Modelling a humanoid virtual leader assistant for a remote and isolated caregivers

Aryana Zaleh Collins Jackson

DOCTORAT BRETAGNE LOIRE / MATHSTIC

ÉCOLE NATIONALE D'INGÉNIEURS DE BREST

THÈSE DE DOCTORAT DE

L'ÉCOLE NATIONALE
D'INGÉNIEURS DE BREST

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *human-computer interaction*

Par

Aryana COLLINS JACKSON

Modelling a Humanoid Virtual Leader Assistant for a Remote and Isolated Caregiver

Thèse présentée et soutenue à Plouzané, le 19/9/2022
Unité de recherche : Lab-STICC, UMR 6285, CNRS
Thèse N° : 14

Rapporteurs avant soutenance :

Catherine PELACHAUD — Directrice de Recherche CNRS, UPMC
Alexandre PAUCHET — Directeur de Recherche, INSA Rouen

Composition du Jury :

Président :	Catherine PELACHAUD	Directrice de Recherche CNRS, UPMC
Examinateurs :	Brian RAVENET	Maître de Conférences, IUT d'Orsay
	Domitile LOURDEAUX	Directrice de Recherche, UTC Compiègne
Dir. de thèse :	Ronan QUERREC	Professeur des Universités, ENIB
Co-dir. de thèse :	Pierre DE LOOR	Professeur des Universités, ENIB
Encadrante :	Elisabetta BEVACQUA	Maîtresse de Conférences, ENIB

Acknowledgement

I would like to thank my supervisors, Elisabetta Bevacqua, Pierre De Loor, and Ronan Querrec for their guidance and support throughout this thesis; Yann Glémarec and Dr. Eimear Wall for their advice and expertise; my CSI committee, Domitile Lourdeaux and Brian Ravenet, for their guidance during this these; the annotators of the speech dataset (Stuart Exshaw, Dean Exshaw, and Jacob Mahon; and finally, Stuart for his tremendous support in various ways throughout the entire work.

Table of Contents

Part I

Context & Literature

Chapter 1

Context for the thesis

Contents

This chapter presents the background to this work, in particular the context for this work, the research questions we address, and the approaches we have taken to answer them. The fields of embodied conversational agents, pedagogy, and medicine are briefly introduced (sections 1.1 and 1.2), and the motivation for this work is presented (section 1.3). Finally, the solution that this work proposes is discussed (section 1.4), and the chapter concludes with a summary of the contributions of this work, (section 1.5), the list of publications resulting from the work (section 1.6, and the structure of the manuscript (section 1.7).

1.1 Introduction

With the development of computers, humans gained the ability to delegate tasks that were otherwise cumbersome (Ensmenger 2012). Computers became our assistants, performing tasks that made our lives easier and allowed us to advance our technological abilities. Sectors like finance, commerce, education, medicine among others were furthered by the presence of computers since data could now be housed, accessed, and used for many

different purposes by many different kinds of people. Computers themselves became more and more advanced, capable of tasks in design, calculation, prediction, and more. As computers' capabilities grew, so did the humans', with us being able to expand what we could do as individuals, as groups, and as a society.

Often, computers were used to give us information (Ensmenger 2012). Employees would access databases to extract information relevant to their jobs and to their companies' businesses. Websites and search engines were designed that allowed consumers to search for services they needed. Computer programs could use data to output predictions of sales, voting habits, and much more. In cases such as these, humans have used computers to find the information they need without communicating with them directly. The development of virtual agents changed that.

Virtual agents are software applications that use natural language to converse with human beings, often for the purpose of providing answers (Burgoon, Bonito, et al. 2000). They are part of what is called human-computer interaction (HCI), a field of design and use of the interfaces between humans and computers. Virtual agents have often taken the form of chatbots which companies implement in order to field the majority of customer service queries (Følstad and Skjuve 2019). These chatbots use natural language processing (NLP) techniques to understand users' natural language queries and provide answers in natural language as well (Følstad and Skjuve 2019). Users can have restricted conversations with these agents.

Another type of virtual agent are virtual assistants, which are often used to provide information to users as well (Chung and S. Lee 2018). As opposed to the chatbots mentioned above, these agents are designed to converse with humans because they are designed to provide a wider range of information. However, often restrictions are still placed on these assistants because otherwise, the human conversation partners might go outside the assistants' range of knowledge (Bickmore, Trinh, et al. 2018). The important thing to note is that virtual assistants are specifically purposed for providing information to human users.

When conversational agents take visual form with a face and/or body, these agents are referred to as embodied conversational agents (ECAs) (Chetty and M. White 2019; Yalçin 2020). They often take the form of a human being but they can also take the form of animals or characters. Because these agents are designed to be looked at while they are being talked to, the human user engages fully in the interaction. ECAs are designed to be fully autonomous, and while there may be restrictions placed on the kinds of conversations that can be had with the agent, the idea is that the human user is able

(a) BEAT used with the PUNK1 agent (Cassell et al. 2001).

(b) Greta used in AVLaughterCycle (Urbain et al. 2010).

(c) SmartBody in the form of a doctor (J. Lee and S. Marsella 2006).

Figure 1.1 – Three previous ECA systems, from left to right: BEAT, Greta, and SmartBody.

to immerse themselves in the interaction. Within the field of HCI, there is an emphasis on socio-emotional relationships between the agent and the human user.

In the following section, we detailed the context in which this work is situated.

1.2 Context

As this work involves the development of a virtual agent, it is important to first give context regarding founding work in agents. More detailed information about the work specifically related to this thesis can be found in chapter 2, but this section establishes the context in which this thesis is situated.

1.2.1 Embodied conversational agents

Previous research on ECAs has aimed to create a real-time adaptive system which receives input that dictates output in the form of speech and nonverbal behavior (Bevacqua et al. 2009; Cafaro, Bruijnes, et al. 2017; Cafaro, H. H. Vilhjálmsson, et al. 2014). The study on embodied conversational agents (ECAs) gained momentum with two key pieces of research in the early 2000s: BEAT (Cassell et al. 2001) and Greta (Pasquariello and Pelachaud 2002) (shown in Figures 6.4a and 6.4b). Both pieces of work studied and quantified human behavior in order to define a set of rules - hence the term rule-based models - under which a virtual agent can operate.

Early work on Greta focuses only on facial expressions. The use of wrinkles in particular expand beyond the BEAT framework which has very limited use of facial movement to convey emotions and attitudes. In 2002, Greta was the state of the art for facial expression in virtual agents. Since then, however, the research into virtual agents has expanded: not only has bodily expression evolved, but the methods for defining behavioral rules which trigger those expressions have been expanded upon and refined. The work on Greta has involved the intelligent choosing of speech and nonverbal behaviors (Grimaldi and Pelachaud 2021) and the extraction of meaning from text in order for the agent to perform the most appropriate nonverbal behaviors (Ravenet, Pelachaud, et al. 2018) among many other projects.

In 2006, a nonverbal behavior generator was created called SmartBody, in which input text is analysed for its syntactic and semantic structure (J. Lee and S. Marsella 2006) (see Figure 6.4c). The affect of the virtual agent is taken into account to develop a seamless appearance when synchronising text and nonverbal behavior. Since then, SmartBody has demonstrated how different motion blending techniques work for an agent's nonverbal behavior (Yazhou Huang et al. 2012). SmartBody has been used over and over again because it gives agents the ability to use active and passive behavior during an interaction with many different kinds of behavior (Krämer, Lucas, et al. 2018; Yalçin 2020).

Regarding behavior selection, previous work has involved linking agent intentions to behavior (Ravenet, Cafaro, et al. 2015), hard encoding agents with characteristics such as "warmth" and "competence" (Nguyen et al. 2015), building agent-user interaction on trustworthiness and competence (Kulms, Mattar, et al. 2015), and using machine learning techniques to choose appropriate attitudes and behaviors (Chollet et al. 2014). Real-time adaptation allows an agent to display believable and socially-appropriate behavior. An agent's personalized content and conversations have been found to improve user engagement, improve the quality of speech, provide timely feedback during the interaction, provide adaptive training, and allow for self-reflection (Kocaballi et al. 2019).

In terms of producing speech, many of these use finite-state systems that do not have the same flexibility that agents in agent-based states do (Laranjo et al. 2018). Speech in finite-states must adhere to a set protocol, and therefore rule-based methods work best. More advanced methods such as neural networks allow for more open an unrestrained communication between an agent and a user, and these methods are better kept to frame-based or agent-based systems (Laranjo et al. 2018).

Agent speech (and nonverbal behavior) is often handled by a taxonomy that specifies

various dialogue acts that an agent can utter (Bunt 2009). Dialogue acts are communicative functions that express a change in information state that the speaker wishes upon the listener. These dialogue acts can be used in conjunction with an agent's communicative intentions so that the agent is communicating appropriately. Communicative intentions are discussed in more depth in sections 2.6 and 5.1.

ECAs have been used in a number of domains, but two are of particular importance in this thesis: medicine and education. Agents in medicine are relevant as our agent will eventually lead users through a medical procedure, and pedagogical agents are relevant because when agents have been positioned as leaders in the past, it is because they are teaching users something. The project details are discussed further in section 1.3, but for now, keep in mind that both medicine and pedagogy are important in this work.

Within the medical domain, ECAs have been used in various capacities (Laranjo et al. 2018). There are a variety of examples of virtual agents created for use specifically in medical situations, such as agents which act as a liaison between patients and physicians using pre-scripted speech (Bickmore, Asadi, et al. 2015) and agents which are meant to connect emotionally with patients for mental health benefits (Kearns et al. 2020; Montenegro et al. 2019; Yang and Fu 2016).

ECAs have also been used in pedagogical scenarios in which they teach or guide a human user. For example, ECAs have been used for mindfulness and meditation coaching (Hudlicka 2013), social skills training (Tanaka et al. 2017) in which they utilize information states as well information about the human user to tailor their instruction, and fire-fighting training in a virtual immersive environment (Querrec, Buche, et al. 2004).

In order for interactions between humans and agents to work well, the human needs to engage with the agent. Previous research has studied the indicators of engagement from a user, which can include things like facial expressions, head movements, and turn taking (Dermouche and Pelachaud 2019). Analyzing user behaviors like these allows us to (1) understand how well a user trusts the agent and (2) allows us to adapt agents to users when appropriate.

What has not been studied yet is agents as leaders during a medical situation. The project that drives this thesis is discussed in more detail in section 1.3, but keep in mind that there is no existing work which studies an agent as a leader inside the medical domain or otherwise.

1.2.2 Leadership

Because there has been no work on agents as medical leaders, we discuss humans as medical leaders to provide context for leadership and relationships in a medical situation.

There are a number of leadership models that describe how a leader (the individual in charge of a situation) can best manage the task at hand and their relationship to the followers (the individuals following the leader's direction and guidance). Distributed or collaborative leadership involves multiple leaders, each leaders of their own domain (Cuban 1998; Spillane 2005). Instructional leadership asserts that a leader whose purpose is to educate followers is already an effective leader (Bush 2003; Chase and Guba 1955). Transactional leadership describes a relationship where follower's labor is always rewarded (Gümüş et al. 2018). Transformational leadership inspires, motivates, and intellectually stimulates followers (Judge and Piccolo 2004; Shamir and Howell 1999). Servant leadership focuses on the idea that leaders should aspire to serve their followers' needs (Greenleaf 2015). Relation leadership asserts that leadership can come from anywhere rather than from a designated leader (Murell 1997; Uhl-Bien 2003).

Finally, Situational Leadership® (referred to as SL® throughout this thesis) asserts that there is no one best leadership model, and that the best leadership style depends on the specific situation at hand (Hersey et al. 1988). The situation includes what is called a follower's readiness level, which describes where they are on an ability scale (novice to expert) and where they are on a willingness scale (unconfident to confident). The situation also includes other environmental factors. Because generally the research agrees that there is no one best leadership style for everyone and because SL® takes that into account, this is the leadership model chosen in this thesis.

In terms of relationships during a medical situation, a huge part of organizing a successful procedure is maintaining a positive interaction with the caregiver while maintaining the health of the patient (Araszewski et al. 2014; Rhona Flin et al. 2010; Henrickson et al. 2013; Hjortdahl et al. 2009; Yule, Rhona Flin, et al. 2008). Leadership in the medical emergency room manifests itself in the coordinator or the surgeon: the individual who facilitates all coordination between team members and procedural tasks (Forster et al. 2005; Moher et al. 1992). Previous studies have demonstrated that a successful medical leader is one who interacts in a respectful and helpful way with the members of the team and also directs the team towards the best outcome for the task (Hjortdahl et al. 2009; Moss et al. 2002). The team must have trust in the leader to make the right decisions and ensure that all processes are completed efficiently and correctly (Kulms and Kopp 2016). In

order for a leader to gain the trust of the followers, they must display competence (Rhona Flin et al. 2010; Morineau, Chapelain, and Quinio 2016), confidence (Hjortdahl et al. 2009; Morineau, Chapelain, Le Courtois, et al. 2017), and delegate tasks efficiently and appropriately (Henrickson et al. 2013; Hjortdahl et al. 2009; Moher et al. 1992; Morineau, Chapelain, Le Courtois, et al. 2017; Morineau, Chapelain, and Quinio 2016; Yule, Rhoda Flin, et al. 2006).

In order to effectively communicate and run the medical procedure properly, taxonomies of medical leader non-technical skills help a leader choose the right way to communicate and the right way to behave in all different kinds of situations that may arise. Existing taxonomies include the Non-Technical Skills for Surgeons (NOTSS) (Yule, Rhoda Flin, et al. 2006), the Surgeons' Leadership Inventory (SLI) (Henrickson et al. 2013), and the Anesthesiologists' Non-Technical Skills (ANTS) (Rhona Flin et al. 2010).

A number of papers have been published in recent years and earlier on non-technical skills in a medical context that confirms the ideology behind NOTSS, ANTS, and SLI. In one study, a coordinator's primary role was found to be handling communication and coordination (Moss et al. 2002). In a second study, the skill *task Management* was found to positively and/or negatively affect the outcome of a medical scenario. Each task has prerequisites, corequisites, and postrequisites in terms of tasks or communication, and the coordination of those tasks can determine procedure success (Morineau, Chapelain, and Quinio 2016).

While these taxonomies have been useful for human medical professionals, they have not been used in a virtual agent system for an agent acting as a medical leader. This thesis addresses this gap in the research.

Now that some context has been provided, the thesis project can be explained.

1.3 Project overview

The work in this thesis is part of VR-Mars [1], a project funded by the National Research Agency of France (ANR). VR-Mars surrounds a hypothetical scenario in which a team of astronauts on a mission to Mars encounter a medical emergency. There is no medical doctor among them, and the nearest medical expert is on Earth, which would only be reachable after a minimum of a twenty-minute delay as there is a 10-minute delay in communication between Earth and Mars. During such a situation, the guiding principles

1. Virtual Reality, Medical Assistance and Rescue for Spationauts, `www.enib.fr/vrmars`

of VR-Mars assert that virtual reality and ECAs would improve the coordination of care and better awareness of the situation.

There are three main difficulties presented in this project:

1. Physiological constraints: change in gravity can result in changes to cellular metabolism, the cardiovascular system, the immune system, vision, and sensorimotor coordination (Nicogossian 2016); additionally, extreme conditions, confinement, fatigue, and social isolation can lead to psychiatric symptoms and cognitive errors, especially when dealing with unforeseen critical situations (Fiore et al. 2015);
2. Limited material and human resources: this scenario takes place on Mars where there is a finite number of resources and no possibility of acquiring new resources, and so there is less ability to manage the situation (Descartin et al. 2015);
3. Time constraints: the distance between Earth and Mars can result in a delay of twenty minutes or more, and so communication with the ground control center on Earth is difficult.

The main objective of VR-Mars involves the design of an experimental prototype representing a medical assistance system accessible by a caregiver in an isolated environment. This system will be equipped with a knowledge base comprising a collection of illnesses and medical procedures intended for the astronauts (the caregivers in the situation). The system will also link the caregiver to the experts with a remote supervision center taking into account the latency between the two sites. The system will send information such as the patient's condition and the procedures started and/or completed to the control center on Earth and likewise transmit instructions from the control center to the human being(s) on Mars.

In order to facilitate this communication between the control center on Earth and the astronauts on Mars, a virtual agent is proposed. The agent could be useful for a number of different reasons:

— To combat the latency in communication between the control center on Earth and the astronauts on Mars - an autonomous agent could interact independently with the caregivers if it has all the information about various illnesses and procedures within the system;
— To aid in communication with Earth - the agent could correctly relay instructions from the medical experts on Earth and could correctly relay information about the procedure to the experts;

— To manage the emotional states of the astronauts - the astronauts could be very stressed and could need a constant reminder of what they should do while taking care of their colleague.

With the VR-Mars project in mind, we present the research questions driving this thesis. The overarching question is this: How can an agent system effectively lead a follower through a medical procedure? This question can be split into further research questions:

1. How can an agent system effectively identify a follower's correct readiness level during a procedure?
2. How can an agent system identify the agent's most appropriate leadership style during a procedure?
3. How can an agent perform different styles of leadership through multimodal behavior?
4. Under what leadership style do caregivers of each type perform better?

In the next section, we explain the work that this thesis proposes with relation to the research questions above and the ultimate contributions to the field of ECAs that this thesis provides.

1.4 Proposition

This thesis concerns the management of the relationship between the agent and the caregivers as well as the agent framework which allows the agent to monitor the procedure and guide different kinds of caregivers through each task. This thesis is theoretical in nature, and so the majority of the work is an exploration into how the application should work rather than the application itself. The end result is work that is applicable to not just the VR-Mars project but a wide range of work on virtual agents, HCI, and human-human relationships as well.

The proposed ECA is equipped with knowledge of its human users' capabilities and the medical procedure at hand (including all steps, user roles, and possible consequences of each action). The SAIBA-compliant agent framework involves text-to-speech speech, without an emphasis on intonation. The agent can display nonverbal behavior including gaze behavior, facial expressions, and gestures. The agent system is able to monitor caregiver behavior and adapt the agent's behavior according to the caregiver.

While the chosen medical procedure that the caregivers must perform is the priority, also of great importance is how the agent interacts with the human caregivers. It is vital that the agent is able to lead the caregivers toward the best outcome: an efficient procedure in which all users work effectively together and maintain the health of the patient (Manser 2009). To create behavior that accomplishes this task, the agent is regarded as a leader and the caregivers its followers. SL® is employed as a mechanism through which the agent can communicate with, guide, and assess the followers while accepting their behavior as well as situational changes as input.

By monitoring the environment, including the astronauts, the agent will be able to choose and then perform the best leadership style with multimodal behavior. An agent that leads a person through a medical procedure also acts as a tutor since it is guiding a person through steps that they may not be familiar with if the person is a novice caregiver or is unfamiliar with the medical procedure at hand.

When guiding human beings during a stressful situation, communication is key. The agent system must choose its words wisely in order to guide the human through each step of the procedure. The human caregiver must be motivated by the agent and given an appropriate amount of instruction.

Note that in this theoretical work, only interaction between the agent and a single caregiver is examined.

1.5 Contributions

The scientific contributions of this thesis, in order of importance, are as follows:

1. The creation and implementation of an algorithm that uses follower behavior and outputs readiness level: first, an algorithm is developed that uses several parameters with respect to each follower behavior and calculates readiness level in real time. Second, a thorough study of previous work is used to identify what kinds of follower behaviors present during a medical emergency are relevant to the VR-Mars project. Second, this algorithm is validated over multiple simulated medical scenarios, which validates the use of the algorithm. This part of the thesis addresses research question 1.

2. The creation of an algorithm that determines the most appropriate leadership style: first, environmental factors that should be used to determine the most appropriate leadership style at any given time are identified, such as changes within the

procedure and the patient's state of health. Second, an algorithm is developed that uses several parameters with respect to each of those environmental factors and calculates readiness level in real time. This algorithm is also simulated over multiple different medical scenarios, which validates the use of the algorithm. This part of the thesis addresses research question 2.

3. The design and construction of an agent framework that uses SL® to allow an agent to lead a follower through a medical procedure: the system allows for the collection of data from several environmental factors as well as multiple types of data from the followers. The system is flexible and is applicable to human-agent interactions beyond those in the medical domain and also allows for different kinds of user behavior to be monitored. Additionally, the system allows for communicative intentions to be created using the results from contributions 2 and 3. This part of the thesis addresses research questions 1-3.

4. The identification of the most appropriate leadership style for each type of follower: with analysis of the perception of speech experiment, we have found the leadership style that each type of follower performs best under and therefore we validate the SL® model as it applies to a human-agent interaction. This part of the thesis addresses research question 4.

5. The identification of medical leader speech so that the agent can use appropriate speech in each leadership style: a dataset of medical leader speech was compiled, annotated with leadership style by four annotators, and then used in a number of analyses to extract the linguistic and semantic properties of each utterance. The final properties were then validated by participants in an experiment. Work on the linguistics that should belong to each leadership style has not been done before, and so this part is applicable to not only human-computer relationships but to human-human relationships as well. This part of the thesis addresses research question 3.

6. The development of a taxonomy for an agent leading a medical procedure: a taxonomy is designed that combines non-technical skills that a medical leader needs during a procedure with communicative intentions as well as speech acts from existing agent taxonomies. Even though this taxonomy was developed to help manage the human-agent interaction in our system, it can also be used for human-human interaction in the medical domain, especially when more structure beyond what currently exists is needed.

This part of the thesis addresses research question 3.

7. The identification of appropriate nonverbal behavior for a medical leader in each leadership style: through the study of existing work on both human and agent nonverbal behavior, a compilation of behaviors that could be used by an agent emulating different leadership styles. Again, this compilation is applicable to not only to human-agent interaction but to human-human interaction as well. This part of the thesis addresses research question 3.

8. The design of an experiment that validates the computational model of SL®: An experiment was designed to validate the computational model described in contributions 5 and 6. This experiment aims to validate the method of choosing readiness level, and the method of choosing leadership style by analysing whether the correct readiness level has been chosen and whether the most appropriate leadership style has been chosen. The experiment evaluates the values within the model, compares the computed readiness level with what the participant's readiness level should be, and also evaluates how participants perceive the agent's leadership style and examines what leadership style leads to the best outcome in the medical procedure. Unfortunately, due to Covid-19, this experimentation was not able to be carried out.

Thus the research questions have been addressed: an agent system is proposed which allows an agent to effectively lead a caregiver through a medical procedure autonomously regardless of whether the caregiver is a novice or an expert. Regarding research question 1, an agent can identify a follower's correct readiness level with the behaviors and algorithm designed. Regarding research question 2, an agent can identify the most appropriate leadership style by using environmental factors chosen and the second algorithm designed. Regarding research question 3, an agent can perform different styles of leadership through appropriate nonverbal behaviors chosen through careful research and through different styles of speech that are proposed. Finally, regarding research question 4, a user evaluation that was conducted reveals what leadership style each type of caregiver performs best under and therefore validates SL®.

The work in this thesis presents several novel works that are applicable to the domain of virtual agents but also to areas outside of virtual agents. The computational model developed in this thesis can be applied to any agent that acts as a user's leader or guide in any domain. The work on agent speech is the first of its kind to classify speech according

to leadership style, and thus it is also applicable to human-human relationships. This new model has been implemented in a flexible agent framework, which has not been constructed before.

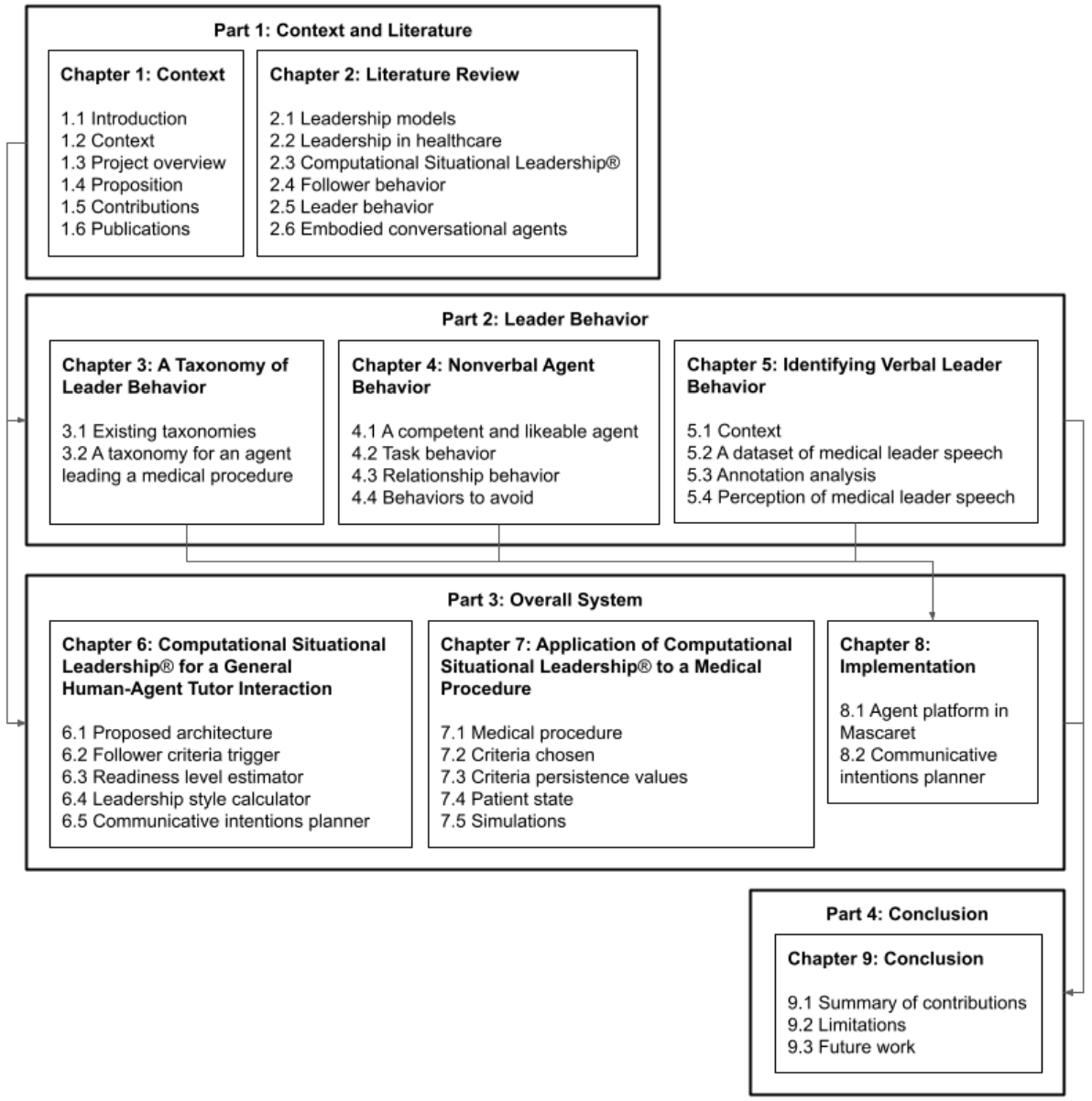
1.6 Publications

The work in this thesis has resulted in the following international publications:

1. Collins Jackson, A., Bevacqua, E., De Loor, P., and Querrec, R., "Modelling an embodied conversational agent for remote and isolated caregivers on leadership styles", in *Proceedings of the 19th International Conference on Intelligent Virtual Agents*, IVA, Paris, France: ACM, July 2-5, 2019. DOI: 10.1145/3308532.3329411;
2. Collins Jackson, A., Bevacqua, E., DeLoor, P., and Querrec, R., "A taxonomy of behavior for a medical coordinator by utilizing leadership styles", in *Proceedings of the International Conference on Human behavior and Scientific Analysis*, ICHBSA, Miami, Florida: WASET, Mar. 2020. DOI: 5913527.YfQCpFjMLOQ;
3. Collins Jackson, A., Bevacqua, E., DeLoor, P., and Querrec, R., "Designing speech with computational linguistics for a virtual medical assistant that uses situational leadership", in *Proceedings of the International Junior Researcher Conference on Human Perspectives on Spoken Human-Machine Interaction*, SpoHuMa, Online: FRIAS, Nov. 15-17, 2021. DOI: 10.6094/UNIFR/223815;
4. Collins Jackson, A., Bevacqua, E., De Loor, P., and Querrec, R., "A computational interaction model for a virtual medical assistant using situational leadership", in *Proceedings of the International Joint Conference on Web Intelligence and Intelligent Agent Technology*, WI-IAT, Essendon, VIC, Australia: ACM, Dec. 14-17, 2021. DOI: 10.1145/3498851.3499019;
5. Collins Jackson, A., Gilles, M., Wall, E., Bevacqua, E., De Loor, P., and Querrec, R., "Simulations of a computational model for a virtual medical assistant", in *Proceedings of the 14th International Conference on Agents and Artificial Intelligence*, ICAART, Online: SCITEPRESS, Feb. 3-5, 2022. DOI: 10.5220/0010910700003116;
6. Collins Jackson, A., Glemarec, Y., Bevacqua, E., De Loor, P., and Querrec, R., "Speech perception and implementation in a virtual medical assistant", in *Proceedings of the 14th International Conference on Agents and Artificial Intelligence*, ICAART, Online: SCITEPRESS, Feb. 3-5, 2022. DOI: 10.5220/0010914500003116.

1.7 Thesis outline

The organization of this thesis can be described by the graphic below. The document is organized into four parts: (1) context and literature, in which a foundation is provided for all the work completed, (2) leader behavior, in which study, analysis, and experimentation are done to explore behaviors indicative of each leadership style, (3) overall system, in which the model and implementation of the model are explained, and finally (4) the conclusions and future work.



Chapter 2

Literature Review

Contents

Because this thesis involves aspects of leadership, medicine, and virtual agents, previous work from each of those multiple domains is relevant. The work completed in this thesis that is discussed in the rest of this document relies on all the information provided here as a foundation. In this chapter, the previous related work is discussed in detail. Section 2.1 details the most relevant work on leadership models in human-human interactions. Section 2.2 delves into leadership models used in healthcare environments. Section 2.3 goes over existing work on computational Situational Leadership ® (referred

to as SL®). Section 2.4 details examples of follower behavior that can influence their readiness level. Section 2.5 details examples of leadership behavior with regard to SL ®. Finally, section 2.6 provides relevant work on virtual agents. Conclusions from this chapter are found in section 2.7.

2.1 Leadership models

In this section, the theoretical basis for SL® this system is discussed in more detail as well as other important and relevant leadership models. As mentioned in the introduction, the agent must be able to lead the human caregiver through the medical procedure. In order to do that successfully, the agent must adopt a style of leadership that will lead to a successful procedure.

2.1.1 Situational Leadership®

As mentioned in chapter 1, this thesis focuses on utilising SL® to employ an agent who can lead a medical procedure. The concept of SL® was first introduced by Hersey et al. in 1969 and revised in 1988 (Hersey et al. 1988) and expanded upon in 1993 (K. H. Blanchard et al. 1993). Born out of a series of studies that appeared to have conflicting results in terms of what kind of leadership yields loyal and competent followers, SL® addresses those discrepancies by proposing a leadership model that suggests the most appropriate leadership style and leader behavior depending on the situation. The goal and present circumstances inform the behavior of the leader. The followers' levels of competence, commitment, and confidence affect the leader's socio-emotional (referred to here as relationship) behavior and directive (referred to here as task) behavior.

Note that SL® is a model not a theory, meaning that the focus is on the practical nature of leader-follower relationships. While SL® was developed from observational studies and previous literature on theories of leadership (Hersey et al. 1988), this model focuses on implementation.

There are a number of factors that influence leader effectiveness, including the leader themselves, the followers, the job demands, and the decision time among other things. However, the main influences are the leader and the followers, and therefore those two will be focused on here.

The follower is assessed by examining their readiness level, a description of compe-

tence, regarding a particular task. Readiness level is composed of two factors: ability and willingness. Ability is the knowledge, experience, and skill that a follower brings to a particular task and is measured with three elements:

1. Past job experience;
2. Job knowledge;
3. Understanding of job requirements.

Willingness is the extent to which the follower has confidence, commitment, and motivation to accomplish a particular task, and is measured by the following elements:

1. Willingness to take responsibility;
2. Achievement motivation;
3. Commitment.

Although ability and willingness are different dimensions, they interact with each other to form the four readiness levels, each denoted with an R:

R1: The follower is unable and lacks commitment, motivation, and/or confidence;

R2: The follower is unable but is motivated and making a genuine effort OR the follower is unable but is confident as long as the leader is there for guidance;

R3: The follower is able but is not willing to use their ability OR the follower is able but is not confident or apprehensive about doing the task on their own;

R4: The follower is able and committed and/or confident.

Followers in levels R1 and R2 are leader-led while those in styles R3 and R4 are self-led. The transition from leader-led tasks in level R2 to self-led tasks in R3 can lead to some apprehension, hence the lower willingness in level R3. No individual is ever completely ready for everything, so the evaluation of readiness level must be task specific in order to evaluate the follower's abilities with respect to specific tasks. For example, someone who is a professional doctor might have a readiness level of R1 with regard to painting.

Similarly, leadership styles are composed of both task behavior and relationship behavior. Task behavior is the extent to which a leader presents the duties and responsibilities of an individual. When task behavior is high, the leader is not concerned about the emotions of the follower but instead is intent on the follower completing the task. When a leader performs task behavior, they:

1. Specify the goals people are to accomplish;

2. Organize the work situation;
3. Set timelines;
4. Provide specific directions;
5. Specify and require regular reporting on progress.

Relationship behavior is the extent to which the leader engages in two-way communication with the follower, thus performing listening, facilitating, and supporting behaviors. When a leader performs relationship behavior, they:

1. Provide support and encouragement;
2. Involve people in give-and-take discussions about work activities;
3. Facilitate people's interactions with others;
4. Seek out and listens to people's opinions and concerns;
5. Provide feedback on people's accomplishments.

The four leadership styles in detail are:

S1: (Directing) Provide specific instructions and closely supervise performance;

S2: (Coaching) Explain decisions and provide opportunity for clarification;

S3: (Supporting) Share ideas and facilitate in making decisions;

S4: (Delegating) Turn over responsibility for decisions and implementation.

Note that leadership style 1, directing, can be viewed as "crisis leadership" because it is most appropriate during times of crisis as it involves instructions only.

Figure 2.1 displays the task and relationship axis graph with leadership styles as well as the corresponding readiness levels.

In order to aid leaders in performing the right behaviors according to the leadership style chosen, twelve descriptors are used, shown in Table 2.1. Leaders can use these descriptors to guide their own behavior and interactions with their followers.

Leaders using high task behavior should communicate precisely, and so directing speech is expected to take the form of direct orders in the imperative mood (Hersey et al. 1988). Speech for high-relationship leaders (coaching and supporting) should create a sense of autonomy for the follower by engaging in two-way communication. This often ends up being speech that is not a direct order, as orders do not allow for the follower to make a choice about whether or not they will comply with a request. An example given is a sentence that begins with "I'd appreciate it if you..." (Hersey et al. 1988). This sort

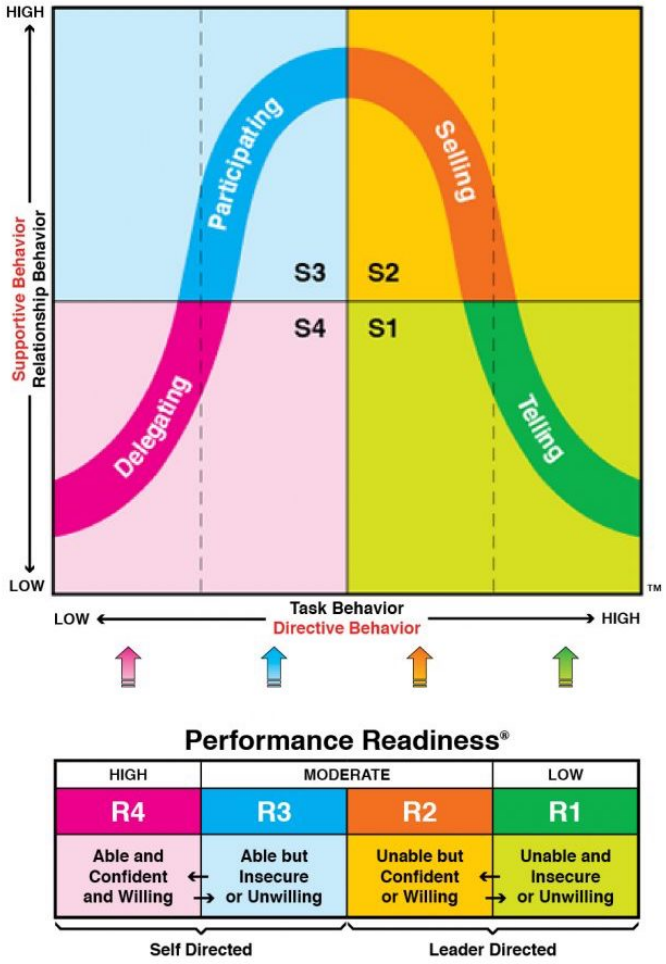


Figure 2.1 – The original task and behavior axes along with how they relate to the the four leadership styles as well as the four readiness levels that correspond (Hersey et al. 1988). Note that in the original work, *supporting* leadership was referred to as *participating*. Due to the fact that the virtual agent does not participate in the medical procedure and following the lead from the original work which encourages renaming leadership styles when necessary (Hersey et al. 1988), we have called the third leadership style *supporting* instead.

of structure provides instruction but also grants the listener a great deal of autonomy by not directly telling them what to do. Delegating leaders interact with followers less than any other type of leader and instead only observe with the intention of stepping in when needed.

More on specific leadership behaviors are discussed further in section 2.5.

SL® aims to help followers progress in their skills and confidence and ultimately grow.

Directing	**Coaching**	**Supporting**	**Delegating**
telling	selling	participating	delegating
guiding	explaining	encouraging	observing
directing	clarifying	collaborating	monitoring
establishing	persuading	committing	fulfilling

Table 2.1 – The original twelve descriptors that guide behavior in each of the four leadership styles (Hersey et al. 1988).

There is no leadership style that is better than the others because each one is appropriate for different followers. Ultimately, it is up to the leader of the situation to determine what leadership style to use. Paying close attention to the followers and what areas of their tasks are most important will lead to the most appropriate style of leadership (K. H. Blanchard et al. 1993; Hersey et al. 1988).

When determining whether a follower's readiness level, and therefore the corresponding leadership style, should advance, a follower's performance relative to the task at hand should be the main determining factor (Hersey et al. 1988). More on specific behaviors that are used to measure readiness level are discussed in section 2.4. Ultimately, the followers hold the power in leader-follower relationships because it is their behavior that determines the leader's behavior.

SL® was tested in an educational setting in which teachers were trained and told to implement the styles to their classes and on an individual basis (K. H. Blanchard 1967). Students were found to have more enthusiasm for the material and statistically significantly performed better. Additionally, their readiness level as a group progressed as long as the teacher implemented SL®. In another study, sixty-five managers in a corporate environment implemented SL® with their direct reports (Gumpert and Hambleton 1979). Those who used SL® rated their subordinates as higher performing *and* they themselves were rated as more effective managers than those who did not use SL®.

There are many more examples of SL® being implemented, although many of these studies use surveys as their sole method of analysis. This can be problematic when follower readiness level is self-assessed by the followers themselves (Thompson and Glasø 2018). This study found that leader assessment of follower readiness level is much more accurate, and that follower performance is highest when their leader-assessed readiness levels are matched with the corresponding leadership style. Employees were also found to perform better when matched with the corresponding leadership style in Lee-Kelley's work (Lee-Kelley 2002).

SL® has also been implemented in medical contexts to train medical professionals with varying success. In a study involving surveys filled out by surgeons, it was found that empowerment of followers (resident doctors) is incredibly important in a training/medical situation as they need to be recognized for the experience that they have even though they are not experts (Sims et al. 2009). When there is a crisis situation and the patient's state is at risk, leadership style should be directing. If the resident's experience level is low for a particular procedure, then leadership style should also be directing. Otherwise, the surgeon in charge should allow the resident more autonomy by leading with supporting or delegating leadership.

However, during another study involving interviews with supervisors and supervisees in a clinical supervision setting for therapists, SL® was less successful (Papworth et al. 2009). The authors note that SL® lacks empirical support. However, this may be due to the fact that although the trainee therapists' readiness levels were correctly identified, they were not matched with the most appropriate leadership styles.

According to the original work, SL® should not be employed to stick to hard rules (Hersey et al. 1988) and instead leaders should monitor followers to see if their leadership is having the desired/best effect. Followers should constantly be monitored.

2.1.2 Situational Leadership II®

This thesis is based on the original work on SL® that was developed in 1969 by Paul Hersey and Kenneth Blanchard (Hersey et al. 1988). However, Blanchard went on to create his own model, a new take on SL®, called Situational Leadership II® (SLII®) (K. Blanchard et al. 1999).

Based on SL®, SLII® presented a different perspective for the evolution of follower readiness level. In SLII®, leadership style 3 is called supporting rather than participating. Ability is referred to as competence, and willingness is referred to as commitment. Followers begin with low competence but high commitment and progress to a state of low competence and low commitment. This is inverse to SL® in which followers progress from low willingness to high willingness. SLII® is displayed in Figure 2.2.

SLII® asserts that followers begin tasks enthusiastically, but as they progress, they realize how much they do not know and therefore become uncommitted. SL® argues that willingness has less to do with enthusiasm and more to do with apprehension about one's own skills, and thus followers begin as insecure (or unwilling as the original work says). This thesis works from SL® because novice caregivers better match that definition of

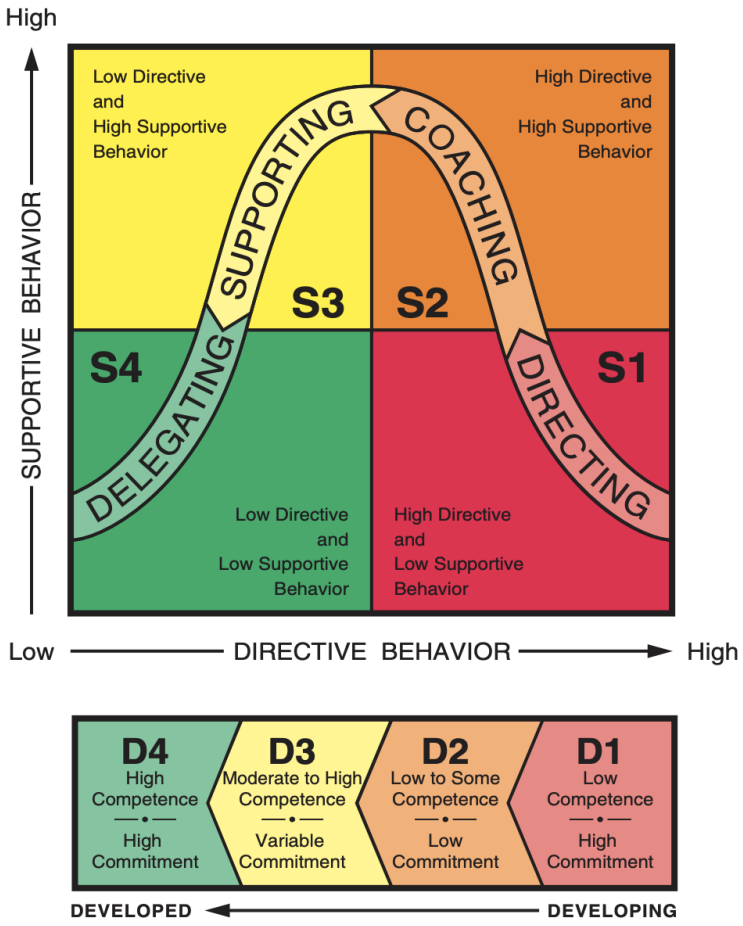


Figure 2.2 – The task and behavior axes along with how they relate to the the four leadership styles as well as the four readiness levels that correspond in SLII® (Northouse 2007).

readiness level R1 and the expected progression better matches the follower evolution in SL® (Hersey et al. 1988).

Another difference between the two models is in task-related development: SL® does not view a follower's progression in terms of their ability as necessarily linear, evolutionary or predictable. Although followers are expected to progress through one readiness level to the next, it is understood that followers may progress and regress in their ability throughout a task. SLII®, on the other hand, expects that followers make a continuous progression without regressing. Again, SL® is chosen as during a medical procedure with varying tasks completed by potential novice caregivers, caregiver ability may rise and fall repeatedly (Hersey et al. 1988; Jewell 2013).

There are several other differences between the two models, but the two mentioned above are the most important.

Like SL®, SLII® has been tested and implemented by various researchers. In 2017, researchers studied SLII® in human resources development through surveys given to employees in various organizations worldwide (Zigarmi and Peyton 2017). They found that followers most often perceive leadership styles S2-S4, with few examples of S1. However, they found that followers often did not get matched with the leadership style they needed. This study was based on what followers felt they wanted. The research behind SL® states that followers self-reported readiness level is often inaccurate. This possibility, combined with the fact that this study did not actually measure the effectiveness of leadership styles matched with their corresponding readiness levels, makes this particular work difficult to interpret.

2.1.3 Other leadership models

Of course, SL® and SLII® are not the only theories of leadership. It is worth mentioning other popular theories of leadership and their strengths and weaknesses. To provide a bit more context here, we must define what a theory or model of leadership actually is: a guide that suggests specific leader behaviors based on situational factors (Sims et al. 2009). Leadership means *influence*, as in the ability to influence others (Sims et al. 2009).

In 2018, a review of various leadership models in the education domain was published (Gümüş et al. 2018). The most commonly discussed leadership models between 1980 and 2014 include distributed/collaborative, instructional, transactional, and transformational leadership. Although these leadership models are specifically assessed in relation to their use in educational settings, all leadership models are broad enough to be applied to a variety of situations. The scenario that this thesis centers on, a virtual agent leading a potentially novice caregiver through a procedure, involves a pedagogical scenario, and so leadership styles that have been used specifically in education are relevant.

Distributed or collaborative leadership is based on the idea that one individual leader is not enough because it is impossible for one individual to have all the knowledge necessary to bring about successful results to the team (Cuban 1998). In distributed leadership, all individuals work together as leaders of their own domains (Spillane 2005).

Instructional leadership asserts that an individual can be an effective leader when their purpose is solely to instruct others (Chase and Guba 1955). In educational settings, instructional leadership is mostly carried about my the school principal, the person visibly

at the head of a school (Bush 2003).

Transactional leadership involves leaders who make requests with the understanding that the labor their followers complete will be rewarded for something and often involves hierarchical relationships (Gümüş et al. 2018). Transactional leadership has much in common with task behavior from SL® in terms of what it looks like and what it hopes to achieve in follower behavior.

On the other hand, transformational leadership inspires, motivates, and intellectually stimulates followers. Transformational leadership is also called charismatic leadership because leaders should embody some charismatic characteristics in order to influence followers. Transformational versus transactional leadership has been studied extensively in the past (Judge and Piccolo 2004; Shamir and Howell 1999). These studies found that transformational leadership had a significant positive effect on followers in educational settings. Transformational leadership is quite similar to relationship behavior in SL® in terms of what it looks like and what it hopes to achieve in follower behavior.

Servant leadership focuses on the idea that leaders should aspire to serve their followers; they should bonding with the followers, empathize with their needs, take care of them, and empower them (Greenleaf 2015). We can see that there are many similarities between this and relationship behavior from SL® and transformational leadership.

Relation leadership asserts that leadership can come from anywhere rather than from a designated leader (Murell 1997; Uhl-Bien 2003).

Leadership styles are the method through which these leadership theories are implemented (Sacavem et al. 2019). Leadership styles are sets of behavioral guidelines that help leaders implement the theory of their choice. Historically, leadership styles were composed of only task-oriented or interpersonal-oriented styles (Keyton 2018). Another classification of styles included autocratic, democratic, and laissez-faire leadership styles (Lewin and Lippitt 1938, 1939; R. White and Lippitt 1960). Autocratic leaders share many principles with directive leaders from SL® and transactional leadership. Democratic leaders are quite similar to supporting leaders from SL® and transformational leaders. Laissez-faire leaders are much like delegating leaders from Situational Leaderhip®.

More recently, research has suggested that the more leadership styles a leader has to choose from in their leadership, the more effective they will be (Goleman 2000). In 2000, six leadership styles were proposed: (1) coercive style which requires immediate compliance from the followers, (2) authoritative style which mobilizes people toward a vision, (3) affiliative style which builds emotional bonds and harmony with followers, (4) democratic

style which aims to develop consensus through follower participation, (5) pacesetting style which expects greatness and self-direction from followers, and (6) coaching style which develops followers for the future. Similarities can be seen between some of these styles and those from SL®.

While the six leadership styles above are all relevant and important in a number of settings, some of them are irrelevant during a medical procedure. For example, what Goleman 2000 considers coaching is not relevant during a one-time procedure. SL® is a more flexible leadership model because it allows for different leadership styles depending on situational factors. During a medical procedure with caregivers that could be either novices or experts, a number of situations could arise. Thus SL® remains the leadership model that makes the most sense for this thesis.

In the following section, leadership specific to healthcare settings is discussed.

2.2 Leadership in healthcare

When discussing leadership in healthcare, we mean leadership between medical leaders (such as surgeons) and followers (such as medical residents). Our agent acts as a leader in a medical situation, and so studying leadership models specific to healthcare is relevant. There have been varying arguments for one leadership style or another in general, such as servant leadership (Trastek et al. 2014) and SL® (Sims et al. 2009). There have also been a number of reviews of leadership in healthcare settings (Harris and Mayo 2018; Trastek et al. 2014; van Diggele et al. 2020). In this section, specific implementations of leadership models in healthcare settings are examined.

SL® was explored as a method of organizing personnel in the trauma center in 2009 (Sims et al. 2009). The results found that using a situational approach makes sense because a number of positive and negative events can occur during surgery and the leader (the surgeon) needs to be able to appropriately respond to each one of them. The study provides an in-depth step-by-step process in which leaders can choose the best leadership style for the situation. Five, rather than four, leadership styles are defined in terms of the leadership behavior involved in each and the situations and followers that they correspond to:

1. Aversive: a leader who is serious and severe in order to deal with problematic followers;

2. Directive: clear and direct and offering low socio-emotional support (same definition from SL®);
3. Transactional: a leader who makes it clear that desired behavior will be rewarded;
4. Transformational: a leader who aims to motivate and inspire followers (shares similarities with coaching and supporting leadership from SL®);
5. Empowering: a leader who helps followers take on responsibility in order to become leaders themselves (shares similarities with delegating leadership from SL®).

We can see that the five leadership styles above begin with high task behavior and low relationship behavior according to the definitions provided by SL®. As we move further down the list, the amount of task behavior decreases while the amount of relationship behavior increases, exactly like the theory described by SL®.

Importantly, this study asserts that during surgery in the trauma center, if the patient's state is critical, then the leadership style should always be directive (Sims et al. 2009). In the original work on SL®, Hersey et al. noted that other variables will sometimes take precedence: things like situational crises, time crunches, etc. may change the appropriate leadership style (Hersey et al. 1988). Like Hersey et al. stated, the followers' expertise and knowledge is a valuable determinant of the leadership style. However, more importantly is the severity of the circumstance. No matter how competent and committed a follower is, the situation at hand may call for a directive leadership style simply because there is no time that can be allocated to encourage or support (Sims et al. 2009).

2.3 Computational Situational Leadership®

Now that previous work on various leadership models have been discussed and the reasons for choosing SL® have been given, it is time to delve into how a situational model of leadership can be implemented computationally. The works covered in this section do not pertain specifically to healthcare or to virtual agents, but their work provides an instrumental foundation regardless.

There are two previous works that create computational models of SL® (Ben-Asher et al. 2018; Bosse et al. 2017). In 2018, researchers created a model that aimed to provide a leader with the most appropriate feedback a follower in each readiness level should receive in order to help them progress (Ben-Asher et al. 2018). In this model, follower trust and adaptability to tasks are the main factors when determining leadership behavior. This

	R1	R2	R3	R4
1	Defensive behavior	Nodding head	Being hesitant	Sharing creative ideas
2	Complaining behavior	Seeming eager	Being resistant	Being result-oriented
3	Intense frustration	Speaking intense and quickly	Feeling overworked	Being willing to help others
4	Late task completion	Listening carefully	Seeking reinforcement	Keeping boss informed of task progress
5	Performance only to exact request	Accepting tasks	Feeling over-obligated	Shows confidence
6	Argumentative behavior	Acting quickly	Lacking self-esteem	Making efficient use of resources
7	Discomfort in body language	Seeking clarity	Focusing on potential problems	Being responsible
8	Confused unclear behavior	Making yes I know comments		
9	Fear of failure	Answering questions superficially		
10	Concern over possible outcomes			

Table 2.2 – The thirty-three follower behaviors belonging to each readiness level from previous work on computational SL® (Bosse et al. 2017).

model mainly focuses on the followers' progression rather than the followers' efficiency and performance overall. In particular, they investigated how often and what kind of feedback a follower should receive in order to progress. A sliding window is used to determine readiness level from follower behavior over time.

In 2017, a computational model was developed that allows the leader to adapt their leadership style by monitoring a follower's behavior (Bosse et al. 2017). The researchers created this work based on an interaction between a student and a supervisor over the course of a thesis. A persistence parameter is used in their algorithm to calculate readiness level over a long period of time. In this model, the extent to which a follower exhibits thirty-three different behaviors determines readiness level. These thirty-three behaviors are categorized by readiness level, indicating that each behavior is indicative of only one readiness level (shown in Table 2.2).

As shown from Table 2.2, these behaviors include general outward attitudes, nonverbal behaviors, speech, and emotions. Some of them are things that a leader can perceive from observing a follower (e.g., defensive behavior and seeming eager), but some are things that the follower must express themselves (e.g., fear of failure and feeling over-obligated). Some

of them are quite vague as well (e.g., being responsible and confused unclear behavior). Additionally, each of these behaviors is organized by readiness level, meaning that this model considers certain behaviors as only indicative of a particular readiness level.

The computational model itself is discussed in more detail in chapter 6, but it provides a flexible model that methodically calculates a follower's readiness level from their own behavior. Once readiness level has been determined, the corresponding leadership style is chosen for the leader to embody.

In the following section, follower behavior that should influence readiness level is discussed further.

2.4 Follower behavior

In the scenario that drives the work in this thesis, the follower is the human caregiver. In order to determine their readiness level according to SL®, their behavior must be monitored. Therefore, we examine previous work that studied follower behavior to understand what might be the indicators of a person's readiness.

In one pedagogical scenario, student performance was measured by the number of errors they make, the time they take to complete tasks, and the number of times they request help (Nakhal 2017). These parameters share similarities to the behaviors listed in Table 2.2.

In SL®, there are some general rules given for behavior of followers in each readiness level. For example, followers in readiness level R1, which is an unable and unwilling or insecure, may exhibit more hesitation or may ask more questions. Followers in R2, those who are unable but willing or confident, may have less hesitation but may make more mistakes. Able but unwilling or able but insecure followers, those in R3, may have more hesitation but may make fewer mistakes and may ask fewer questions. Followers in R4, those who are able and willing or able and confident, may perform the fastest and the most accurately of all the follower types (Hersey et al. 1988).

Unfortunately, there is not much previous work on what constitutes an R1, etc. follower in the emergency room. It is generally expected that anyone inside an emergency room completes each step correctly and quickly. Additionally, there is not much room in an emergency situation for the follower to lead themselves, as they would while in levels R3 and R4 (Jewell 2013). Regardless, there has been research into the management of tasks by healthcare workers, and these are worth mentioning.

In 2013, a literature review of works on supporting novice nurses was published (Jewell 2013). The findings from these works include that novice nurses often do not ask questions when they should and that asking questions is a sign of growth and progression (Andersson and Edberg 2010; Delaney 2003). Expert nurses are able to communicate effectively in a variety of situations while novices may not be able to (Andersson and Edberg 2010; Delaney 2003). While these behaviors are for nurses in general and not emergency room-specific, they do provide insight into how novice versus expert caregivers may behave.

Additionally, examining behavior from students and tutees is helpful as well. As mentioned in chapter 1, this thesis involves a pedagogical scenario, and so the human caregiver can be considered a student. In 1986, a study was conducted to monitor the behaviors of students during peer tutoring (McKellar 1986). Among the tutee behaviors that were positively correlated with academic performance is asking questions. However, note that because this piece of research surrounds students who are in need of academic help in the first place, questions are a sign of proactiveness.

In 2019, low- and high-performing computer science students were studied (Liao et al. 2019). This research found that while all students asked questions, there were differences in behavior between low- and high-performers: low performers were more likely to give up when confused, try to memorize information, and ask others for answers. High performers were more likely to create their own problems to solve for practice, seek out other resources, and continue upskilling even after the deadline passes.

These resources have all validated SL® which explains that novices (level R1) not only lack ability but lack motivation as well (Hersey et al. 1988). Growing novices (level R2) are motivated and show it by asking questions and staying engaged, as shown by the work covered in this subsection.

2.5 Leadership behavior

In this section, the most appropriate leadership behavior with relation to SL® is explored. In some cases, in-depth information is expanded upon in a later section or chapter, and so those chapter references are provided.

Previous studies outside the medical domain have shown that the best results are attained by a leader embodying a style personalized to a follower (Schyns and Mohr 2004; Sims et al. 2009).

Within the medical domain, there has been a lot of work regarding human-human

relationships. The person leading an emergency medical procedure holds an important role in that they manage both the procedure tasks, the health of the patient, and the interaction with the caregiver(s). Thus having the trust of the caregivers and being competent at their work are two of the most important qualities a leader can have (Hjortdahl et al. 2009; Montenegro et al. 2019; Morineau, Chapelain, and Quinio 2016). Additional qualities that a medical leader should embody include communication, negotiation, autonomy, creativity, and appreciation of the caregivers (Araszewski et al. 2014; Hjortdahl et al. 2009; Moher et al. 1992; Morineau, Chapelain, Le Courtois, et al. 2017; Morineau, Chapelain, and Quinio 2016; Yule, Rhoda Flin, et al. 2006). Many of these qualities are therefore included show up in medical professional behavior taxonomies (Rhona Flin et al. 2010; Henrickson et al. 2013; Yule, Rhona Flin, et al. 2008).

The rest of this section is dedicated to non-technical skills for medical coordinators in the emergency room, leaders' nonverbal behavior, and leader speech.

2.5.1 Leadership behaviors in medicine

In 2002, a study was conducted on communication and coordination in the operating room (Moss et al. 2002). In this study, the leader in the operating room is the charge nurse. The findings include that everyone involved in the operation should receive the same communication, taking the form of a board that everyone can see. The charge nurse is tasked with updating the board with new and changing information. The most common piece of information communicated was the status of surgical cases and their next room assignment. The most important finding here was that the leader, the charge nurse, needs to always be aware and in charge of the communication that all caregivers are receiving.

Later, in 2009, researchers studied the specifics of medical leader communication (Hjortdahl et al. 2009). First, the main determinant of trauma team function was found to be leadership, which although a broad term, indicates that the very presence of a leader who acts as the manager of the team is the most important aspect. Some general characteristics that the leader should have include an interest in emergency medicine, confidence and calmness, good communication skills, good listening skills, good focus, and trust of the medical team. It was also found that in order to be a good leader, the person would also need to be a skilled trauma surgeon. In other words, leaders need to be competent themselves.

Hjortdahl et al. proposed five non-technical skills that are important in the emergency room: (1) awareness of the situation including future states and current changes at all

times, (2) selecting the right decisions quickly, (3) the ability to change the team's direction at any time, (4) building confidence among team members, and (5) ensuring information is communicated properly (Hjortdahl et al. 2009).

In 2016, a study explored the specifics of how nurses and other caregivers should behave in various situations (e.g., they should stand at the headboard when performing CPR on a patient) (Morineau, Chapelain, and Quinio 2016). This work studied the deficiencies in current workflows for a variety of operating room procedures and found consequences for these deficiencies. For example, if the operating room is not cleaned, then the presence of paper and other small garbage can cause nurses to lose track of where the correct instruments are. Ultimately, this study came to the conclusion that the management of the work environment is one of the most important skills a medical leader can have during a procedure. In an emergency situation, non-technical skills often affect the outcome more than technical skills (Hjortdahl et al. 2009; Morineau, Chapelain, and Quinio 2016).

Because behavior in the operating or emergency room has such immediate effects on patient health and caregiver stress, it is important that there are strict guidelines for the leader's behavior. Three taxonomies which establish important non-technical skills and behavioral examples influence this thesis: the Anesthesiologists' Non-Technical Skills (ANTS) created in 2010 (Rhona Flin et al. 2010), the Surgeons' Leadership Inventory (SLI) created in 2013 (Henrickson et al. 2013), and the Non-Technical Skills for Surgeons (NOTSS) created in 2016 (Yule, Rhoda Flin, et al. 2006). The idea is that by following these taxonomies, fewer mistakes can be made.

ANTS Defines a taxonomy for anesthesiologists so that they can strengthen their non-technical skills and successfully manage their own and others' behaviors in the operating room (Rhona Flin et al. 2010). Each non-technical skill includes examples of both good and bad behavior. SLI was created based on interviews with caregivers and video recordings of operations and provides guidelines for surgeons' behavior in the operating room (Henrickson et al. 2013). NOTSS also includes examples of individual behaviors that leaders in the medical room should follow (Yule, Rhoda Flin, et al. 2006). The specifics of these taxonomies is discussed further in chapter 3.

2.5.2 Nonverbal behavior

Nonverbal behavior exists to (a) provide information, (b) regulate the interaction, (c) express intimacy, (d) act as social control, (e) present identities and images, (f) affect management, and (g) facilitate service and task goals (Patterson 1990). Nonverbal behaviors

can then be thought of as belonging to one of these seven functions.

In 2004, researchers studied nonverbal leadership behaviors by isolating the nonverbal behavior indicative of various leadership theories (Schyns and Mohr 2004). They found that while communication and expressiveness are different things, they can often be perceived as the same thing, and that leadership itself is an interactive process. Facial expressions, gestures, and body movement may not do much to further communicate what has already been said with words and inflection, but the addition of them adds meaning to the situation. Schyns and Mohr elaborate on the minute details of how various nonverbal behaviors are perceived, but fails to come to a conclusion regarding which are significant and effective. Perhaps this is due to the fact that leadership styles are not taken into account.

Other research has come to conclusions regarding specific nonverbal behaviors that are effective for leaders such as managers and teachers: For example, engaging in gaze with followers while speaking was found to be a high-powered behavior (Carney, Hall, et al. 2005; Chaudhry and Arif 2012). Forward leans, smiles, and a variety of hand gestures were found to increase follower engagement (Chaudhry and Arif 2012; Greven 2017). Using a variety of tone and speech pace was also associated with follower engagement (Chaudhry and Arif 2012). A palms-upward gesture was found to be associated with leader proactiveness (Greven 2017). The duration of gestures rather than the gestures themselves was found to be significantly associated with follower job satisfaction in a work setting (Ciuffani 2017).

Ultimately, it was found that nonverbal behaviors should match verbal behavior in terms of meaning (Chaudhry and Arif 2012; Schyns and Mohr 2004). There is also a plethora work on cooperative versus dominant nonverbal behaviors, but those are less relevant to a leader-follower relationship because the leader does not necessarily need to be cooperative because they should be in charge, but also the leader should not be dominating the scenario. However, in general, more expressiveness of emotion, especially positive emotion (e.g., smiling) was found to be cooperative (Schug et al. 2010). More on nonverbal behaviors is found later in section 2.6.2 and in chapter 4.

2.5.3 Speech

It is worth briefly mentioning linguistics as a part of this work. Note that the ECA created in this thesis uses text-to-speech, and so elements like prosody and pace are not taken into account. Speech Act Theory (SAT), a theory of linguistics that explores how

words work together to form utterances and intentions directed at listeners (Searle 1979), has been useful for extracting meaning from text, e.g., in opinion mining (Pluwak 2016). Indirect statements are analyzed through semantics and syntax to extract opinions with Wordnet [1], FrameNet [2], and SenticNet [3]. Others have used SAT for meaning extraction in text, primarily for sentiment analysis (Ensink and Sauer 2003; Lakoff 2002). More detailed information is available in section 5.1.2.

Because speech acts involve a speaker's intentions, they are closely related to communicative intentions. A communicative intention is, simply put, what the speaker wants to achieve with a piece of communication. Communicative form is the format that a piece of communication takes. These together change the addressee's information state by adding information or correcting information (Bunt 2009).

Communicative intentions and speech acts are so important in this thesis because identifying how a follower (the caregiver) should behave is vital to the patient's health. By identifying the ultimate goal for follower behavior (the communicative intention), we can choose the best form for that intention to take.

Speech acts are useful not only to help determine the desired intention behind the utterance but also because they can help shape what communication *should* look like in certain situations. Taxonomies for communicative intentions and skills outside of a medical context have been developed and used for various situations long before virtual agents were a topic of interest. For example, in 1976, behaviors of young children were organised into a taxonomy, which laid a baseline of defining desires, needs, and interaction (Allwood 1976). In 1992, a baseline for general overriding dimensions was established: speech management, interaction, and focused messaging (Allwood et al. 1992). In 2001, a taxonomy was defined for educational purposes, focusing on objectives and cognitive processes (Krathwohl 2002). Taxonomies specific to agents are discussed later in section 2.6.2.

As shown from these previous works, there are many different speech acts and there is no agreed-upon and finite list of acts. This is because researchers have studied a wide variety of text and speech, and different speech acts will appear in different contexts. Communicative intentions and speech acts are discussed in more detail later in this chapter in section 2.6 and in chapter 5.

In terms of leadership speech, the twelve descriptors that guide leader behavior in each leadership style (Table 2.1) can be thought of as speech acts that should be used in

1. `https://wordnet.princeton.edu/`
2. `https://framenet.icsi.berkeley.edu/fndrupal/`
3. `https://sentic.net/`

each leadership style (Hersey et al. 1988). In fact, previous work has indeed defined two of these descriptors found in coaching leadership, clarification and explanation, as speech acts (Walton 2007).

Although there is a plethora of work on linguistics, there is a huge gap in the research regarding leader speech. To our knowledge, there is no work that defines what speech should look like in terms of semantics or syntax in different styles of leadership. The work in this thesis addresses that gap in chapter 5.

In the following section, the last section of the literature review, previous work on ECAs are discussed.

2.6 Embodied conversational agents

When virtual agents take visual form with a face and/or body, they become embodied conversational agents (ECAs). ECAs are designed to invite total human engagement during interactions. ECAs are of particular importance for our research because during a medical situation in which an agent is leading a potentially amateur caregiver, it may be necessary for the agent to express instructions nonverbally. Therefore, we explore previous work on ECAs.

The human in the situation needs to trust the agent enough to successfully lead him or her through a series of steps (Kulms and Kopp 2016; Kulms, Mattar, et al. 2015). The concept of developing trust is inherent in all types of human-computer interactions, regardless of domain, as it is a prerequisite to a positive interaction (Hoegen et al. 2019; Kulms and Kopp 2016; S. K. Lee et al. 2021). Additionally, trust leads to greater efficiency when the human is completing tasks (Kulms and Kopp 2016). Therefore user engagement is intertwined with user trust of an agent during an interaction. Measuring engagement is important when evaluating a user's attitude, and can help create a more personalized interaction (Cisneros et al. 2019). Previous research has studied the indicators of engagement from a user, which can include things like facial expressions, head movements, and turn taking which allows us to (1) understand how well a user trusts the agent and (2) allows us to adapt agents to users when appropriate. (Dermouche and Pelachaud 2019).

In this section, we discuss related work pertaining to virtual agents in general, agents within the medical domain, pedagogical agents, and agent systems using virtual reality.

2.6.1 SAIBA

In an effort to unify the multi-modal behavior generation process, the SAIBA [4] framework in conjunction with the Behavior Markup Language (BML) and the Function Markup Language (FML) were established (Cafaro, H. H. Vilhjálmsson, et al. 2014; Kopp et al. 2006; H. Vilhjálmsson et al. 2007). SAIBA is a framework for ECAs that allows for multi-modal interaction between one or more ECAs and one or more humans. An agent system utilizing the SAIBA structure includes an intent planner where agent intentions are formed, a behavior planner where the communicative intentions are translated into verbal and nonverbal signals, and a behavior realizer which translates these signals in animation. FML is the language that encodes communicative intentions into signals, and BML encodes those signals into animation.

In 2007, updates to BML specifically were published (H. Vilhjálmsson et al. 2007). Included in this paper is a list of toolkits that can be used to help ECA frameworks better produce nonverbal behaviors. These include the Expressive Gesture Repository (Zsofia Ruttkay 2001), ECAT: The ECA Toolkit (Zsófia Ruttkay et al. 2006), and the BCBM Behavior Rule Builder (Thórisson 2002).

In 2009, an extension was developed for SAIBA which allowed for better reactive behaviors (Bevacqua et al. 2009). Previously, because all behavior had to originate as intentions in the intent planner, the agent was delayed when reacting to the user in real time. A new module was created that bypasses the intent planner to create low-level behavior so the agent can perform behaviors that synchronize with the human user. Additionally, FML was broken into chunks to further reduce delays in behavior planning.

In 2014, a unified FML was proposed that used contextual information, participant's culture and socio-relational goals in order to translate communicative intentions into verbal and nonverbal signals (Cafaro, H. H. Vilhjálmsson, et al. 2014). The researchers behind this work found that a communicative function can come from a conscious, planned communicative intention that the participant aims to accomplish or unconsciously, as might occur from the participant's mental-emotional state.

Later, in 2017, an FML-template dialogue manager for expressing communicative functions in SAIBA was developed (Cafaro, Bruijnes, et al. 2017). This dialogue manager takes input as text and outputs agent behavior in real time. A series of FML templates for template behavior were designed as well. However, when the input is a sentence, it is not broken up into smaller pieces of information, and thus the dialogue manager cannot

4. SAIBA refers Situation, Agent, Intention, Behavior and Animation

respond to specific parts of the user's words (Nakhal 2017).

2.6.2 Agent behavior

Outside of the SAIBA framework, there has been a lot of work done on producing agent behavior. In this section, prior work on agents' verbal behavior, nonverbal behavior, and the coordination of both verbal and nonverbal behavior are discussed.

Verbal behavior generation

In 2009, a taxonomy of agent communication was developed (Bunt 2009). The Dynamic Interpretation Theory++ (DIT++) taxonomy is a clear and flexible dialgoue-act taxonomy of agent communication that can be applied to various contexts based on ISO standard 24617-2:2012 [5]. DIT++ defines ten dimensions, or classes, of communication that contain general-purpose communicative functions and dimension-specific communicative functions. These functions behave like the speech acts mentioned in section 2.5.3.

The DIT++ also specifies aspects of communication borrowed from SAT, such as direct and indirect communications ("What time is it?" vs "Do you know what time it is?"). Additionally, the DIT++ categorizes certain and uncertain answers from users and how verbal and nonverbal behavior often co-occur. The DIT++ taxonomy is discussed in more detail in section 3.1.2.

Some examples of the communicative functions (speech acts) under the dimension *informing* are *inform*, *agreement*, *disagreement*, and *correction*. Many more dimensions and functions are specified relating to a number of different circumstances that arise during interaction between an agent and a human user.

In 2019, a dialogue-act taxonomy for a virtual coach for improving the lives of the elderly was created based on the DIT++ (Montenegro et al. 2019). The work places a huge emphasis on empathy: how the agent can establish a relationship with the human(s) it works with. Importantly, multi-modal communication is taken into account, meaning both verbal and nonverbal communication. The taxonomy introduces a method of tags that act like functions and user reactions from previous papers (Allwood et al. 1992; Bunt 2009): topics, which help the ECA keep track of conversation changes; intents, like *question*, *inform*, etc.; polarity, meaning positive, negative, or neutral; and entities, such as dates, quantities, etc. These four tags take on a hierarchical structure, with *topics*

5. `https://dit.uvt.nl/`

existing at the top and *entities* at the bottom, which allows semantic information to be gathered at each step of the process.

Another speech act taxonomy was developed in 2019 utilizing a layered approach (Bernard and Arnold 2019). The authors assert that intentions are multi-layered. Intention trees are used to get to the root of a problem or to achieve the ultimate goal after several iterations. Utterances convey multiple things at once, with (1) a physical layer (producing sound), (2) a linguistic layer (making a sentence), (3) a semantic layer (providing meaning), (4) a pragmatic layer (changing the cognitive context), and (5) a cooperation layer (contributing to a common goal).

These taxonomies are used to both generate agent speech and detect the content of a user's speech during an interaction. When detecting a user's speech content, the dialogue acts need to be automatically tagged. This has been done in the past with machine learning techniques (Anikina and Kruijff-Korbayova 2019; Malik et al. 2018).

Agent speech is often implemented with a dialogue engine such as Flipper (van Waterschoot et al. 2018). Dialogue engines can work with a SAIBA-compliant model, incorporate information states, preconditions, and effects in order to produce seamless agent speech in real-time.

Choosing an agent's nonverbal behavior

Agent frameworks like SAIBA can manage agent behavior by utilizing FML and BML to translate intentions into signals and transform those signals into nonverbal behavior. In terms of actualizing that behavior, different engines and methods of implementing nonverbal behavior have been developed. For example, in 2001, a lexicon was developed in which signal-meaning pairs are coded in memory to produce relevant gestures (Poggi 2001). In 2002, a state-of-the-art facial model was developed which simulated the human face in a rapid and believable manner (Pasquariello and Pelachaud 2002). Recently, in 2021, multimodal generation for performing both verbal and nonverbal in BML was demonstrated (Grimaldi and Pelachaud 2021).

In terms of choosing the nonverbal behavior to generate with these systems, a number of different approaches have been taken. Some researchers have chosen to present various nonverbal behaviors to human beings and determine from there what makes appropriate and inappropriate nonverbal behaviors (Krämer, Simons, et al. 2007; Straßmann et al. 2016). Others have taken a more machine learning approach by utilizing large amounts of data and thus determining the best behavior to implement (Chiu et al. 2015; Haag and

Shimodaira 2016; Hasegawa et al. 2018; Yuyun Huang et al. 2016; J. Lee and S. C. Marsella 2010; Mariooryad and Busso 2012; Ravenet, Ochs, et al. 2013; Sadoughi and Busso 2017). Non-learning methods have also been used to coordinate speech and gestures (Bergmann et al. 2013; Cassell et al. 2001; J. Lee and S. Marsella 2006; S. Marsella et al. 2013; Ravenet, Pelachaud, et al. 2018). Each of these approaches are relevant and worth delving into.

In 2007, a perception study was conducted to see how people perceived an agent's nonverbal behaviors (Krämer, Simons, et al. 2007). They found that self-touching gestures performed by the agent led it to be evaluated as more natural, more warmhearted, more agile and more committed. However, frequent self-touching during the interaction led participants to rate the agent as more strained and aggressive. When the agent raised its eyebrows, participants generally had negative feelings toward it.

In 2016, a study was conducted on an agent's nonverbal behavior in relation to their dominance and cooperativity (Straßmann et al. 2016). Overall, they determined that expressivity of nonverbal behavior, more than the behavior itself, determined an agent's cooperativity rather than dominance. However, they did establish that the akimbo posture (Standing straight with hands on the hips), crossing of the arms, and head tilts toward the user conveyed a degree of dominance while open-body gestures, head tilts to the sides, and chin rotations to the sides conveyed submissiveness (Straßmann et al. 2016).

2.6.3 Agents in the medical domain

Now that the specifics of how agents are developed and behavior is generated have been discussed, we delve into some agents that were designed for particular purposes. In this section, we discuss agents that have been used within the medical domain as our agent exists within a medical context.

In 2013, an agent system was developed in order to improve patient diagnoses (Bennett and Hauser 2013). These researchers used Markov Decision Processes (MDPs) to model how belief states change over time and how variables have causal relationships with others. Essentially, a patient's health status is tracked over time and the algorithm re-plans the treatment strategy. The system was able to recommend decisions made throughout the period that the patient was monitored.

In 2016, a "Hospital Buddy" agent was developed, which acted as a liaison between patients and physicians (Bickmore, Asadi, et al. 2015). The virtual agent was designed to allow patients to provide feedback on their care by using pre-scripted speech and a range of synchronized nonverbal behavior. Pre-scripted speech was chosen because in the

medical domain, human-agent interaction should be constrained (Bickmore, Trinh, et al. 2018).

In general, using some conversational agents for medical information is harmful because they are not able to be experts in every area (Bickmore, Trinh, et al. 2018). It is recommended that when dealing with medical information, user input should always be constrained so that the agent's response can be thoroughly validated for every scenario. Every possible interaction scenario needs to be thoroughly validated because providing open-ended assistance could lead to harm to the user (Bickmore, Trinh, et al. 2018).

In 2020, a Wizard-of-Oz (WoZ) system was developed to provide family caregivers with self-management skills and problem-solving skills (Kearns et al. 2020). Using different empathetic personas, the system communicates with caregivers and provides response templates for grounding and intervention when the caregivers need it.

These are certainly not the only examples of medical agents. There have been many more examples of agents in the medical domain, such as those that converse with patients by conducting psychiatric interviews (Philip, Dupuy, et al. 2020) and questionnaires and diagnostics (Lucas et al. 2017; Miner et al. 2016; Philip, Bioulac, et al. 2014; Philip, Franchi, et al. 2017) as well as patient monitoring (Black et al. 2005).

Next, we examine agents developed for pedagogical scenarios.

2.6.4 Pedagogical agents

When we talk about pedagogical agents, we are specifically talking about agents that aim to teach or guide a human user. Our agent will have to guide a human caregiver through a medical procedure that they may not be familiar with, and so examining prior works in which an agent has guided a human is relevant.

One agent interaction model specific to pedagogical scenarios is Mascaret, a UML-based meta-model that permits the modeling of semantic, structural, geometric, and topological properties of the entities in the virtual environment and their behaviors (Nakhal 2017; Querrec, Taoum, et al. 2018; Taoum et al. 2018). Mascaret also defines the notion of a virtual agent by their behaviors, their communications, and their organisation. Essentially, it is a framework in which an embodied virtual human can interact with a user. Mascaret works within a SAIBA-compliant framework, and allows for the creation of high-level intentions in real time. The agent recognizes the user's actions through the interface.

In 2017, a tutoring system was built using Mascaret in which human user performance

was measured by the time of execution, the number of committed errors and the number of requests for assistance (Nakhal 2017). Later, the tutoring system was built to respond to the user's previous experience and level of knowledge. Only a certain amount of complexity of steps is kept in working memory according to the prior knowledge of the user (Querrec, Taoum, et al. 2018; Taoum et al. 2018).

A non-Mascaret system also utilized information states for the agent (Chetty and M. White 2019). Additionally, this system involved different agent personas in order to deal with different scenarios. Users performed better with those personalized personas rather than one static agent for everyone.

Another non-Mascaret system used a framework which trains medical leaders to deal with their teams (Lourdeaux et al. 2019). In this work, leadership and hierarchical relationships were taken into account. However, in this work, the human user is the leader and the ECAs are the followers. The agents' nonverbal behaviors are used to influence the user's behavior. Thus this scenario is the inverse of the one driving this thesis.

Many agents are created and then tested with human users, sometimes to test the perception of nonverbal behavior (A. L. Baylor et al. 2009; Frechette and Moreno 2010). One piece of research studied the effects of various facial expressions and gestures on human students and found that participants involved in attitudinal instruction learned more without deictic gestures (pointing) which those involved in procedural instruction learned more with them (A. L. Baylor et al. 2009). They confirmed that identifying the right nonverbal behavior is tied to the learning outcome. Another study found again that students learned more with an agent using deictic gestures (Frechette and Moreno 2010).

Personalized instruction has been found to be particularly useful. In 2013, a virtual mindfulness and meditation coach was designed and implemented (Hudlicka 2013). This agent was found to be more effective than a self-administered course, demonstrating the value of virtual tutors. Chris, the coach, can display limited nonverbal behaviors and communicate via text-based natural language. Importantly, Chris is able to adapt to the human user's knowledge and motivational state by asking questions and then using keyword and template matching to understand the user's responses as the coaching progresses.

Another example of a successful tutoring system helps individuals with autism spectrum disorders to improve their social skills (Tanaka et al. 2017). The system analyses facial analysis of its users and provides personalized feedback based on that analysis. Adapting to individual human users was also done with different personas, created by

varying agents' speech patterns and gestures, when training users in VR for a number of different domains (Chetty and M. White 2019).

While these examples involve an agent leading a human being through a task or series of steps, tutor agents are also able to perform more social supportive behavior. A study in 2010 with a robotic tutor found that behavior such as role modeling, nonverbal feedback, attention building, empathy, and communicativeness can build a stronger relationship with the human user and also increase the user's learning (Saerbeck et al. 2010).

Users' attitudes towards virtual tutors have been studied as well (Pecune et al. 2010). The agent once again adapted to the learner by responding based on his or her previous experience and level of knowledge. Participants rated agent's social relation, social status, and performance. Participants' perception of high power and high status of the agent was associated with their lowered performance while perception of lower status and likeability were associated with their higher performance during the task.

In the following section, we briefly discuss prior work related to virtual reality.

2.6.5 Virtual agents in virtual and augmented reality

In this final section of the literature review, we discuss a few prior works which utilize virtual reality for their virtual agents. Virtual and augmented reality environments each have their advantages when it comes to ECAs. Interaction with an ECA aims to be engaging for the human. When the human interacts with the agent in virtual reality, they are immersed in the virtual environment and are able to see a carefully curated scene. This could be useful, for example, when the medical experts at the control center on Earth want to see what is happening on the remote site. With virtual reality, they will be able to see exactly what is happening elsewhere.

Virtual reality was implemented in VICTEAMS, which was mentioned earlier in this section (Lourdeaux et al. 2019). The user is immersed in a virtual environment with multiple autonomous agents acting as followers who can interact with each other as well as the user. The environment provides an exercise in which the user must manage the virtual agents and the surrounding environment.

Augmented reality also has its benefits. With augmented reality, an ECA appears in the real-world environment but is able to interact with real-world objects. This could be very useful in our human-agent interaction when the agent needs to point at medical tool or a part of the patient's body. Augmented reality has been implemented in an ECA system that asks human users to find objects in the real-world environment (I. Wang et al.

2019). In another application that studied the positioning of ECAs in space, users had to interact with ECAs in augmented reality (A. Huang et al. 2022). In both of these pieces of research, the appearance and positioning the agent in the augmented reality application led to different user perceptions, but users were able to interact with both the agent and the real world at the same time.

The Mascaret model lends itself very well to both virtual and augmented reality (Buche et al. 2004; Nakhal 2017). In these systems, a user can be fully immersed in a training scenario, such as for learning about blood analysis (Nakhal 2017) or firefighting training (Querrec, Buche, et al. 2004).

Because of the possibilities and advantages of using virtual and augmented reality for the VR-Mars project specifically, the final version of the agent that the work in this thesis pertains to will be applicable to both augmented and virtual reality.

2.7 Conclusions

This chapter has covered a plethora of related works that each provide a foundation for the work completed in this thesis. In section 2.1, we discussed prior work on various leadership models. As explained, there has been a lot of work exploring which leadership styles are best for different situations, which have helped us understand how an agent might be able to lead a follower. In section 2.2, existing leadership models specific to healthcare situations were explored. Situational Leadership® was chosen as the leadership model for our agent system because it describes a method of choosing the most appropriate leadership style depending on the situation.

In section 2.3, we explored prior works that involve mathematically computing a follower's readiness level and the leader's leadership style. One of these works in particular provides a foundation for our work in terms of determining a follower's readiness level. In section 2.4, we examine prior works that discuss behaviors and their meanings that might be performed by followers in various situations. This section expands upon the follower behaviors discussed in section 2.3 to provide a wider range of follower behaviors that may indicate one readiness level or another.

In section 2.5, we explore existing works that discuss behaviors and their meanings that might be performed by leaders in various situations. However, there is no prior work on verbal leadership behavior as it pertains to SL®. Finally, in section 2.6, we discuss existing work on ECAs in a number of domains. Because our work involves an agent in

the medical domain and in a pedagogical scenario, medical and pedagogical agents are of particular importance.

Notably, there is no prior work on (1) virtual agents as medical leaders or (2) the linguistics of SL®, which are the gaps that this thesis addresses.

Key points from Chapter 2

- Situational Leadership® was chosen over other leadership models because it is simple yet flexible and is able to be adapted for any kind of follower;
- There has been one highly relevant work that developed a computational model of Situational Leadership®;
- A variety of human behaviors could influence a follower's readiness level;
- Leadership behavior should vary depending on what kind of follower the leader is interacting with;
- Agent frameworks like SAIBA and interaction models like Mascaret provide an effective interaction model for a pedagogical medical agent.

Part II

Leader Behavior

Chapter 3

A Taxonomy of Leader Behavior

Contents

Part II of this thesis is focused on the identification of leader behavior. The chapters in part II therefore explore behavior that an agent acting as a leader should perform in order to contribute to a positive working relationship with the human caregiver and thus ultimately result in a successful medical procedure.

In this first chapter of part II, we begin to explore how an agent could perform behavior indicative of each leadership style (covered in section 2.1.1). To aid in the creation of agent nonverbal and verbal behavior, a taxonomy is designed to provide a framework within which the agent can operate and interact with the human caregiver.

Taxonomies are road maps in the form of rules and guidelines that ensure a person or agent behaves in a systematic manner in order to elicit behavior from the other people or agents involved in the interaction. This thesis explores how an agent can assume the role of a medical coordinator to successfully lead a caregiver through medical tasks. Before going forward, we must clarify the definition of a medical coordinator. The coordinator is the person who facilitates all coordination between team members and procedural tasks (Forster et al. 2005; Moher et al. 1992). Because the agent cannot participate in the procedure themselves but only serves to guide the caregiver, the agent is considered to be the medical coordinator. The human caregiver is operating on the patient, and because the agent needs to elicit certain behavior from the caregiver in order to provide the best care possible, the agent needs a taxonomy to adhere to.

In this thesis, emphasis is placed on non-technical skills as the agent itself does not perform the procedure. Existing taxonomies for both human medical leaders and virtual agents work to ensure that the human trusts them by presenting them as competent and confident. This part of the thesis aims to identify the fundamental non-technical skills that a medical coordinator can enact during a medical emergency and the communicative intentions that are applicable for the agent.

In this chapter, previous taxonomies for medical leaders and virtual agents are discussed in section 3.1 and a taxonomy of medical coordinator non-technical skills, communicative intentions, and specific individual behaviors is proposed to enable an agent to act as a medical coordinator in an emergency situation is detailed in section 3.2. Section 3.3 contains the concluding remarks and main takeaways from this chapter.

3.1 Existing taxonomies

Existing taxonomies that present behaviors for medical leaders and ECAs in an organized fashion in order to strategically choose the best behavior to perform were first mentioned in section 2.5.1. In this section, we go more into depth about how these taxonomies work and what they do.

3.1.1 Taxonomies for medical leaders

Before diving into the taxonomies themselves, here is a word about where they come from: the Joint Aviation Requirements: Translation and Elaboration of Legislation (JAR-TEL [1]) established recommendations for the management of teams during aviation. The idea was that the safety of airline operations is heavily connected to human factors. Cultural differences were found to influence pilot behavior and attitudes, and therefore guidelines for the management of interpersonal relationships during flight were established.

Criteria from the JAR-TEL project informed the first taxonomy that we discuss here: the Non-Technical Skills for Surgeons (NOTSS) system (Yule, Rhoda Flin, et al. 2006). Designed in 2006, NOTSS aims to address key skills that surgeons need during an operation that enables them to lead a team of medical caregivers. In many medical procedures, a coordinator is not present and instead the surgeon acts as the leader (Henrickson et al. 2013; Hjortdahl et al. 2009; Yule, Rhoda Flin, et al. 2006). After analysis from a series of

1. https://trimis.ec.europa.eu/project/joint-aviation-requirements-translation-and-elaboration-legislation

videos of procedures, leadership was found to be the primary determinant of procedure success. Some of the sub-skills under leadership deemed important include elements like competence, interest in medicine, calmness, and alertness (Hjortdahl et al. 2009; Yule, Rhoda Flin, et al. 2006).

NOTSS has been tried and tested in various medical environments (Yule, Rhoda Flin, et al. 2006). Five non-technical skills are identified: *Situation Awareness*, *Decision Making*, *Task Management*, *Leadership*, and *Communication and Teamwork*. Each skill is divided further into elements that describe concrete tasks that a surgeon performs during a procedure (Hjortdahl et al. 2009; Yule, Rhoda Flin, et al. 2006). For example, under the non-technical skill *Gathering information*, there is an element which says *Ensures that all relevant investigations (e.g. imaging) have been reviewed and are available*. Under the non-technical skill *Projecting and anticipating future state*, there is the element *Verbalises what may be required later in operation*.

Unlike NOTSS, the Anesthesiologists Non-Technical Skills (ANTS) taxonomy was made primarily for anesthesiologists, not surgeons or leaders of a medical procedure, and contains the following four non-technical skills: *Task Management*, *Team Working*, *Situation Awareness*, and *Decision Making* (Rhona Flin et al. 2010). These categories are broken down into various elements, similarly to NOTSS, and examples of good and bad behaviors. For example, under the non-technical skill *Task management*, and then under the element *coordinating team activities*, the example of good behavior is *confirms roles and responsibilities*. Many of the skills and elements are shared between NOTSS and ANTS. Unfortunately, there has not been as much reliability testing done with ANTS since its creation in 2010.

In 2013, the Surgeons' Leadership Inventory (SLI) taxonomy was developed from a series of videos of surgical operations, similarly to NOTSS Henrickson et al. 2013. The ability for a leader to make decisions was labelled as a vital skill - this serves as a reminder that the baseline is making decisions at all, not jumping to making positive or negative decisions. In other words, indecisiveness is the worst attribute for a leader to have during a medical procedure. In the SLI, *Communication and Teamwork* operate under separate categories, with separate elements belonging under each. The study mentions that surgeons' leadership could depend on the type of operation as different tasks are required under different procedures.

Several articles in recent years and earlier on non-technical skills in the medical context have been published, confirming the ideology behind NOTSS, ANTS, and SLI. In one

study, it was found that the main task of the coordinator was to handle all communication and coordination (Moss et al. 2002). The second study found that *Task Management* positively and/or negatively affected outcomes in medical scenarios (Morineau, Chapelain, and Quinio 2016). It was found that each task has prerequisites, corequisites, and postrequisites, and the coordinator needs to manage those for the team in order to guarantee procedure success (Morineau, Chapelain, and Quinio 2016).

3.1.2 Dynamic Interpretation Theory (DIT++)

The DIT++ taxonomy, mentioned first in section 2.6.2, is hugely influential in the field of virtual agents and has been used widely to create agent behavior. The DIT++ taxonomy specifies both dimensions and functions of speech, where functions are essentially communicative intentions and dimensions are classes that these communication functions belong to (Bunt 2009).

The DIT++ consists of both forward- and backward-looking functions, with forward-looking functions being communicative functions, or speech acts, that seek to change the information state without provocation and backward-looking functions being communicative functions that are provoked by the human's speech (Bunt 2009). These categories of communicative functions are divided into dimensions, which specify groups of communicative functions by common purpose (e.g., *statement*, *information request*, and *answer*). Within these dimensions, there are specific dialogue acts. For example, under the *information request* dimension, there are dialogue acts such as *direct question* and *indirect question*.

These dimensions and functions are general-purpose and are meant to be used in a wide variety of virtual agents (Bunt 2009). They help define an interaction with a human from start to finish. There are a multiple dimensions and functions in the DIT++ taxonomy to account for a number of different situations that can occur during interaction, and they all assume that the human will respond to the agent and vice versa.

In the following section, our proposed taxonomy that is specific to an agent leading a medical procedure is thoroughly explained.

3.2 A taxonomy for an agent leading a medical procedure

As established by JAR-TEL, the ability to lead a group of novices and experts alike is imperative and relevant in an age of continued exploration in which medical professionals are not present. For emergencies on remote sites, like ships and space, often the caregivers are not medical professionals (Descartin et al. 2015; Fiore et al. 2015; Nicogossian 2016). Thus a coordinator needs to be able to lead both novices and experts.

In order to effectively manage a caregiver, a virtual agent coordinator must be equipped with the tools necessary to manage both the user and the tasks at hand. Because the context for our virtual agent is a medical emergency, it is vital that the agent's generated behavior adhere to strict guidelines set by both the task and the caregivers.

There are no taxonomies that specify guidelines for agent behavior when that agent leads a medical procedure. Because a medical procedure involves the health of a human patient, it is crucial that the agent follow precise guidelines in order to ensure that there is no unpredictability in terms of agent-human communication during the procedure. Therefore, a taxonomy is proposed that combines both the necessary non-technical skills from the NOTSS, ANTS, and SLI taxonomies as well as the DIT++ taxonomy for agent speech.

The procedure steps and other nominal information is provided to the coordinator beforehand. When the coordinator communicates a task, the task in question comes directly from the nominal procedure or from decisions made by the medical experts in the remote medical center. The agent's role is to communicate effectively with the caregivers so that the procedure goes smoothly in terms of the patient's health and the caregiver's stress level. This is discussed in more detail later in chapter 8.

3.2.1 Taxonomy structure

The taxonomy is structured with high-level information in the form of non-technical skills, low-level information in the form of speech acts, and information in between. Note that we refer to the acts borrowed from the DIT++ as speech acts rather than dialogue acts because the goal of our ECA is to lead rather than to converse.

The only skills and elements that are borrowed from existing work are those that involve interaction with a follower. While a great deal of non-technical skills that do not

involve communication are important to a human medical leader, they are not applicable to a virtual agent. Furthermore, our medical coordinator agent is specific to an emergency procedure in which only certain skills are applicable. Therefore we chose to eliminate all non-communication-based skills entirely.

The taxonomy begins with four non-technical skills that a medical coordinator needs in order to successfully facilitate an emergency procedure by communicating with the caregivers: situation awareness, decision-making, task management, and team management. Situation awareness involves the agent being aware of the patient's and follower's states as well as communications from the medical experts standing by on Earth. Decision-making involves the agent making sense of the situation in order to make an intelligent decision regarding the follower's next desired behavior. Task management involves the agent ensuring the follower is completing tasks correctly. Finally, team management involves the agent managing the relationship with the caregiver.

Each non-technical skill is divided into one or more elements: various sub-tasks that fall under the overarching non-technical skills. These elements provide more specific but still general goals that the agent has under each non-technical skill. Elements share many similarities to the elements in NOTSS, ANTS, and SLI as well as the dimensions specified by the DIT++ taxonomy, and thus they can be thought of as classes of communicative intentions and speech acts.

Elements are further divided into communicative intentions, which were first discussed in section 2.5.3. Communicative intentions what the speaker wants to change about the information state of the listener or what the speaker wants the listener to do. They should lead to an understanding of information or an action. The agent uses communicative intentions to communicate in such a way that information is either provided to the caregiver or information is obtained from the caregiver (H. Vilhjálmsson et al. 2007). These intentions can also show an emotional state or a level of certainty about the current circumstances or about the information the caregiver is imparting. Communicative intentions, as they are defined in this work, do not change with a change in leadership style. Rather, the leadership style only dictates *how* the intention is communicated.

Communicative intentions are not part of the NOTSS, ANTS, or SLI taxonomies, but they are present in other non-agent taxonomies in order to organize communication during human-human interaction (Allwood 1976). Of course, they also organize communication in agent-human interaction (Bunt 2009). By including them in our taxonomy, we establish that each non-technical skill that an agent coordinator needs during a medical procedure

is tied to an intention that the agent has for the human follower.

Lastly, each communicative intention is assigned a speech act. These speech acts were chosen based on SAT (mentioned briefly in section 2.5.3 and discussed in more depth in section 5.1.2), DIT++, and other works which have used the DIT++ for their own specific purposes (Anikina and Kruijff-Korbayova 2019; Bunt 2009; Searle 1979). Each act chosen is one that could be used from a medical leader to a caregiver during a medical procedure:

— Instruct;
— Inform;
— Offer;
— Request information;
— Respond;
— Support.

Respond and *offer* are backward-looking functions, and the rest are forward-looking. A speech act from the list above is assigned to each communicative intention. The speech acts provide a system of classification for agent speech and also a method by which a function of speech can be chosen in order to carry out the communicative function. The goal of this proposed taxonomy is to allow the agent to choose the most relevant non-technical skill, class of communicative intentions, communicative intention itself, and speech act that will ultimately communicate most clearly to the caregiver. These speech acts are discussed in terms of their implementation in chapter 8.

Non-technical skill	**Element**	**Communicative intention**	**Speech act**
Task Management	Planning and preparation	communicates plans	inform
	Flexibility/ responding to change	redirects tasks	instruct
	Prioritising	communicates priority of tasks	inform
	Setting and maintaining standards	states standards and expectations	instruct
	Using authority	gives orders	instruct
		states case and provides justification	inform

Table 3.1 – A subset of the elements, communicative intentions, and speech actions under the non-technical skill *task management*.

Table 3.1 displays a segment of the taxonomy, with the taxonomy in full displayed in Appendix A. As shown, a single non-technical skill (*Task management*) is split into five elements. Each element then contains one or more communicative intentions. Finally, each communicative intention is assigned one of the six speech acts.

This taxonomy is designed to be applicable to a variety of medical procedures and enables the agent to find the appropriate speech and nonverbal behaviors to use in different situations. Note that while the agent developed in this thesis communicates with just one caregiver, there are parts of the taxonomy that refer to multiple caregivers. Communication to multiple caregivers is discussed in section 9.3.

This taxonomy does not involve SL® in any capacity. The agent can communicate nonverbally and verbally using the taxonomy in any of the four leadership styles, as will be explained in chapters 4 and 5. The non-technical skills, elements, communicative intentions, and speech acts do not contain any syntactic rules, allowing for more flexibility of communication in each of the four leadership styles.

3.3 Conclusions

In order to manage interaction between an agent and a human follower in a medical situation, a taxonomy guiding agent behavior was necessary. No existing taxonomy for an agent leading a medical procedure exists, and so we created our own.

The taxonomy for an agent leading a medical procedure (covered in section 3.2) weaves together medical coordinator non-technical skills, virtual agent communicative intentions, and speech acts. It is founded on existing and proven research on taxonomies for human-human and human-agent interaction both in and outside the medical domain (covered in section 3.1). Ultimately, this taxonomy is flexible enough to be used for human-human relationships as well by providing more guidance to medical leaders. In terms of human-agent communication, it allows an agent the high-level and low-level detail needed to guide a medical procedure. This taxonomy guides our work on both nonverbal and verbal agent communication. The taxonomy is used later in our work on both nonverbal and verbal agent behavior.

However, the taxonomy presented here has not been validated with an evaluation. In future work, it would be valuable to conduct an evaluation in which multiple annotators independently annotate a dataset with the non-technical skills, elements, communicative intentions, and speech acts that are included in this taxonomy. Such an evaluation would

be necessary in order to prove that the taxonomy is robust and useful when understanding communication from a medical leader.

Key points from Chapter 3

- Existing taxonomies for medical leaders provide important non-technical skills that are applicable to a virtual agent leading a medical procedure;
- Existing taxonomies for virtual agent verbal and nonverbal communication provide speech acts that are also applicable to a virtual agent leading a medical procedure;
- A new taxonomy containing both non-technical skills and speech acts is proposed;
- Elements and communicative intentions were developed to link non-technical skills and speech acts together, thus creating a simple-to-use and robust taxonomy;
- The proposed taxonomy is applicable to both human-human and human-agent interaction.

Chapter 4

Nonverbal Agent Behavior

Contents

This second chapter of part II focuses on the agent's nonverbal behavior. The taxonomy presented in chapter 3 and available in Appendix A can be used to design an agent's nonverbal behavior when that agent leads a medical procedure. In general, nonverbal behavior exists to (1) provide information, (2) regulate the interaction, (3) express intimacy, (4) act as social control, (5) present identities and images, (6) affect management, and (7) facilitate service and task goals (Schyns and Mohr 2004). Therefore, when designing nonverbal behavior for an agent, it is important to be intentional and ensure that each behavior is adding to the communication.

Previous work on agents' nonverbal behavior was specifically discussed in section 2.5. However, there is a lack of research on nonverbal behavior in the emergency room. Because medical coordinators and leaders are usually involved in the procedure themselves, they typically do not have the capacity to perform certain nonverbal behaviors, like gestures, at all (Forster et al. 2005; Moss et al. 2002).

When designing nonverbal behavior, speech acts and the context of the situation should be influential (Key 1973). This means a few things: (1) the taxonomy presented in chapter 3 should dictate what nonverbal behavior is performed; (2) the context of the situation in terms of SL® should be taken into account so that the behavior is leadership style-specific; and (3) the context of the situation in terms of environmental factors should determine whether the nonverbal behavior is appropriate. For example, smiling can be viewed as a supportive behavior generally, but would not be supportive in a crisis

situation (Darioly and Mast 2014). Intentions behind both nonverbal and verbal communication should match (Chaudhry and Arif 2012).

In section 4.1, behaviors that an agent should embody regardless of leadership style are discussed. In section 4.2, nonverbal behaviors that are indicative of high task behavior (corresponding to leadership styles *directing* and *coaching*) are provided. In section 4.3, nonverbal behaviors that are indicative of high relationship behavior (corresponding to leadership styles *coaching* and *supporting*) are detailed. Section 4.4 contains nonverbal behaviors that our ECA should avoid at all times regardless of situational factors. Our final remarks are found in section 4.5.

Note that this chapter acts as a concentrated literature review with regards to nonverbal behavior that is indicative of different leadership styles. Therefore, the behaviors discussed in this chapter are not evaluated directly. This is discussed further in section 4.5.

4.1 A competent and likable agent

Throughout the literature, several behaviors emerged that appear to always generate a positive user response and caused a perception of competence and likeability. The perception of competence and likeability in a leader is important because it builds trust with a follower and ultimately builds a stronger relationship (Rhona Flin et al. 2010; Montenegro et al. 2019; Morineau, Chapelain, and Quinio 2016). These behaviors discussed in this section are those that an agent, regardless of style, should employ.

Perception of leaders is based largely on composure, competence, as well as warmth and general likeability (Maricchiolo et al. 2009). Good leaders are seen as more open and physically expressive. In terms of body expression, they tend to have more erect posture and more forward lean (Carney, Hall, et al. 2005). Expressivity itself is seen as being of higher power (Carney, Hall, et al. 2005). Regardless of their positivity or negativity, bodily expression itself is seen as being more cooperative than uniform expression (Schug et al. 2010; Straßmann et al. 2016).

Eye contact is generally perceived as a sign of good leadership by maintaining focus on the individuals or object of most importance and also sustaining the relationship aspect of the interaction (Smith 1979). Gaze also implies power; more eye contact towards the followers is perceived as more powerful (Carney, Hall, et al. 2005). In a medical context, the more a leader looks at a follower, the more they are perceived as being proactive (Guillaume et al. 2018). This sign of proactiveness can indicate that followers

view the leader as more effective (Greven 2017).

The use of hands, more than heads and gaze, have been found to increase both humans' and virtual agents' verbal eloquence (Bergmann et al. 2013). In fact, users have been able to interpret more from gestures than speech, tone, facial expressions, and gaze (Maricchiolo et al. 2009). Ideational gestures, ones that directly refer to objects or ideas, such as the drawing of a circle with the fingers to indicate a circular concept, are effective at communicating ideas. They often result in more understanding and thus a perception of higher competence of the leader (Maricchiolo et al. 2009). Steepling of hands translates as confidence and competence (Navarro and Karlins 2008, pp. 133–164), which leads to higher trust of the leader (Hjortdahl et al. 2009; Yule, Rhoda Flin, et al. 2006).

The presence of hand gestures increases the perception of competence (Biancardi et al. 2017). Hand gestures where the hands are open and palms are facing one another or hand gestures where there is a lot of movement led to higher follower satisfaction (Ciuffani 2017).

4.2 Task behavior

Recall that task behavior is behavior a leader performs to present the duties and responsibilities of a follower (Hersey et al. 1988). A leader performing high-task behavior is not concerned about the emotions of the follower but instead is intent on the completion of the task. Thus, in general, the speech act *instruct* is used often in high-task behavior, although remember that the taxonomy is not split up by leadership style and that generally the non-technical skills can be used by a leader in any style. Leaders that use high task behavior are more direct when providing instructions. In general, more "robotic" or stiff behavior can be considered more direct (Saerbeck et al. 2010).

While performing the speech act *instruct*, there are a few nonverbal behaviors that may be helpful when a leader exhibits task behavior. Maintaining eye contact with each follower is hugely important to ensure followers are listening and able to follow along (Carney, Hall, et al. 2005). Ideational gestures are a way of conveying specific ideas like where the follower should act on the patient (e.g., for resuscitation, abdominal palpitations, etc.) (Guillaume et al. 2018). High-task leaders should clearly articulate instructions to followers who may not otherwise know what to do.

Additionally, pointing gestures are seen as being especially proactive, since they are a clear ideational gesture indicating what the user is meant to do (Guillaume et al. 2018).

In several studies, agents that used pointing gestures led to an improvement of students' learning (A. L. Baylor et al. 2009; Frechette and Moreno 2010).

The agent may perform the speech act *respond* when the leader has to correct or disagree with something the follower has said or done. In cases such as these, palms-downward gestures can indicate a wish to stop the current situation: they may want to interrupt the situation because the user does not understand, they disagree with the way things are going, or things are moving too fast (Kendon 2004; Matsumoto and Hwang 2012). A shake of the head can indicate disagreement as well (J. Lee and S. Marsella 2006).

4.3 Relationship behavior

Relationship behavior refers to socio-emotional support that a leader provides to their followers (Hersey et al. 1988). Note that cooperation is mentioned often in this subsection. Cooperation and dominance often act as opposites when it comes to social interaction, where cooperative qualities rank higher in users' perceptions of the individual and dominant qualities rank lower (Guillaume et al. 2018; Sims et al. 2009; Straßmann et al. 2016). Cooperation refers to a person's ability to work well with others, which is hugely important in the context of a medical scenario (Guillaume et al. 2018; Ishikawa et al. 2006; Sims et al. 2009). Users have been found to be more trusting of cooperative leaders rather than dominant ones (Kulms and Kopp 2016; Kulms, Mattar, et al. 2015; Schug et al. 2010; Straßmann et al. 2016) (covered further in section 4.4).

In general, the speech acts that a leader using high-relationship behavior might perform more often would be *offer* and *support*. When an agent is relatable, people often find them more supportive, and so human-like behavior is encouraged (Saerbeck et al. 2010). For example, self-touching behavior can be perceived as both warmth and commitment (Krämer, Simons, et al. 2007). Smiles are seen as expressions of joy or friendliness (Chollet et al. 2014; Samter 2006).

When the agent is performing the speech act *offer*, a nonverbal behavior that might accompany it is a palms-upward gesture: the agent may be showing or giving something (even if the "something" may be intangible) (Kendon 2004).

When the agent performs the speech act *instruct* with high relationship behavior, instead of resorting to pointing, the agent might do a head tilt instead. Head tilts in general are perceived as cooperative (Straßmann et al. 2016). Lateral head movements

can refer to specific objects in the space, directing users to look towards something (J. Lee and S. Marsella 2006).

A nod or shake of the head should accompany the speech act *respond* because it is a reactive behavior (Saerbeck et al. 2010; Straßmann et al. 2016). A happy face in the case of a correct answer or a sad face in response to a mistake is also highly responsive and therefore indicative of high relationship behavior (Saerbeck et al. 2010).

Other cooperative behavior involves forward leans (Carney, Hall, et al. 2005) (Navarro and Karlins 2008, pp. 85–108), head tilts (Straßmann et al. 2016), and overall more expressive behavior (Schug et al. 2010; Straßmann et al. 2016). Wide eyes can indicate praise or support of the followers (Navarro and Karlins 2008, pp. 165–204).

When hands are together and palms are turned upward, the leader may be indicating that they are withdrawing from the current task or do not want to intervene in the current situation, which happens in follower-led leadership styles supporting and delegating (Greven 2017; Matsumoto and Hwang 2012).

4.4 Behaviors to avoid

There are behaviors that all leaders regardless of style should avoid. Dominance, referring to a person's ability to control a situation and other people, can lead to animosity or a feeling of being threatened (Burgoon and Dunbar 2006). Additionally, submissive behaviors should be avoided as a leader should be recognised as the authority figure throughout the procedure (Guillaume et al. 2018).

Expansive gestures in which arms and hands are open and away from the torso are perceived as dominant and should be avoided (Biancardi et al. 2017; Burgoon and Dunbar 2006; Carney, Hall, et al. 2005). Other dominant behaviors to avoid include the akimbo posture [1] (Straßmann et al. 2016), crossing of arms (Straßmann et al. 2016), hands clasped together (Chollet et al. 2014). The turning of the head towards a fellow follower is seen as being less cooperative (Straßmann et al. 2016). Both upward head-tilts and downward head-tilts can indicate dominant assertion and condescension (Lance and S. C. Marsella 2007; Mignault and Chaudhuri 2003). Gazing less at followers, particularly when they are speaking, is seen as dominant behavior (Carney, Hall, et al. 2005). Finally, the raising of eyebrows can be interpreted as negativity (Krämer, Simons, et al. 2007).

In terms of submissive behaviors, constrictive gestures (where limbs are held tight to

1. standing with straight posture and both hands on the hips

the body; the opposite of expansive body posture) are regarded as being powerless (Carney, Cuddy, et al. 2010). Additionally, although self-touching gestures may increase perception of warmth and friendliness (Krämer, Simons, et al. 2007; Straßmann et al. 2016), they have also been found to indicate low power and insecurity (Carney, Hall, et al. 2005; Harrigan et al. 1991; Krämer, Simons, et al. 2007).

Tense lips, lips that are pursed together and unsmiling, can be indications of thinking or of disagreement. They can be perceived negatively and so are best to avoid (Navarro and Karlins 2008, pp. 165–204). Finally, too many hand gestures could be perceived as a lack of confidence (Greven 2017).

4.5 Conclusions

In this chapter, we provide a number of recommendations for nonverbal behavior for an agent acting as a leader of a medical procedure. A list of all behaviors discussed in this section can be found in Appendix B.

As mentioned at the beginning of this chapter, these behaviors are not evaluated. However, each nonverbal behavior discussed can be thought of as a hypothesis to a future experiment where its validity as a task- or relationship-related behavior is tested as well as its relation to the speech acts in our taxonomy. This future experimentation is also important because of slight contradictions in the literature. For example, some research suggests that self-touching gestures increase perception of warmth and friendliness (Krämer, Simons, et al. 2007; Straßmann et al. 2016), while other suggest that they they indicate insecurity (Carney, Hall, et al. 2005; Harrigan et al. 1991; Krämer, Simons, et al. 2007). A participant evaluation would resolve these inconsistencies.

Such an experimentation could take one of several forms depending on the desired result of the agent system. For example, nonverbal behavior could be evaluated without simultaneous speech in a between-subjects experiment. Alternatively, different nonverbal behaviors could be evaluated with the same speech in a between-subjects experiment to determine the effects of different behaviors. These experiments could be done using a survey-style approach or they could be done during the procedure itself, with the participants acting as the caregivers. Ultimately, the goal would be to understand if nonverbal behaviors affect participants' ability and willingness and perception of the agent leader and if so, which behaviors lead to greater ability and willingness and the most trust of the agent.

The final nonverbal behaviors will be stored in a gestuary in the agent system to be used in conjunction with other elements of the taxonomy (this is discussed in more detail in section 8.2.2).

Key points from Chapter 4

— Previous works on human-human and human-agent interactions were analyzed for their work on nonverbal behavior;
— Nonverbal behaviors were recommended to accompany certain speech acts in certain leadership styles.

Chapter 5

Identifying Verbal Leader Behavior

Contents

In this final chapter of part II, we explore how an agent acting as a leader can perform leadership behavior through speech.

Despite various studies on the performance of situational leadership, no prior work has been completed to discover, concretely, the linguistic elements of each leadership style. In this final chapter of part II, we embark on several projects to define linguistic rules for speech in each leadership style. The work presented here provides novel contributions to the fields of human behavior, healthcare, and intelligent virtual agents.

Before we delve into agent speech, note that communication requires that (1) two or more individuals are willing and capable of communicating; (2) the listener is willing and capable of perceiving the behavioral, verbal, or other means whereby the speaker is signalling information; and (3) the listener is willing and capable of understanding the content that the sender is displaying or signalling (Allwood et al. 1992). Therefore, these assumptions are present throughout this chapter and the rest of this thesis.

In order for situational leadership to work, the followers must be able to perceive leadership style correctly. Therefore, this work explores speech whose leadership style is perceived the same way by multiple people. The agreement on leadership style by multiple people indicates that the characteristics of that utterance are almost universal and can be perceived largely the same by a group of caregivers, no matter who they are.

The agent developed in this thesis utilizes text-to-speech, and so inflection and prosody are not the subject of this thesis. Speech for an agent in the medical domain should also have restrictions placed on what they can say so that errors overall are minimized and so that followers do not get off-track (Bickmore, Trinh, et al. 2018), and so the speech developed in this chapter is quite restricted by design.

In this chapter, context is provided for work on linguistics in section 5.1, the creation and annotation of a dataset of medical leader speech is explained in section 5.2, analysis of the annotated dataset resulting in a set of linguistic rules is covered in section 5.3, and our experimentation to study the perception of those linguistic rules is detailed in section 5.4. Our final remarks and conclusions are found in section 5.5.

5.1 Context

We provided some context for leadership speech and agent speech in sections 2.5.3 and 2.6 respectively. In this section, we go into more depth regarding existing works that

provide necessary context in order to understand the process of our work for identifying appropriate agent speech.

5.1.1 Grammatical moods

Before delving into existing work on linguistics, it is necessary to define what a grammatical mood is. A grammatical mood is a verbal form that allows speakers to express their attitude toward the subject of their speech. In English, there are five main moods (*The Five Grammatical Moods* 2022), with the verbal component in bold:

— Indicative mood, used to express a fact ("She **likes** the gift.");
— Imperative mood, used to express a command or a request ("**Clean** your room.");
— Interrogative mood, used to express a sense of uncertainty by asking a question ("**Are** you **coming** to the summer camp?");
— Conditional mood, used to express a condition statement ("If you want to visit your friends, you **should study** now.");
— Subjunctive mood, used to express a wish, doubt, demand, or a hypothetical situation ("If I **were** in her situation, I would never drive.").

The main moods that could appear during an interaction between an agent and a caregiver are imperative, interrogative, and indicatives. Therefore, those three moods are discussed throughout this chapter.

5.1.2 Speech Act Theory

As mentioned in section 2.5.3, a foundational piece to this research is Speech Act Theory (SAT) (Searle 1979). In this research, the use of language in terms of intention and form are studied. An illocutionary act is defined as an attempt to communicate which includes both the illocutionary force (related to the communicative intention) and the form that the communication action takes.

An utterance can be direct or indirect. Direct utterances have no alternate possible meanings, whereas indirect utterances do have some other possible meanings other than the most obvious. Another definition is that in indirect utterances, the form and proposition are not literal (Mann 1980; Searle 1979). e.g., the difference between "Pass me the salt" and "Can you pass me the salt?". In daily life, we will most often accept that the two sentences have the same meaning (a request for the salt). However, the second utterance could also be asking about literal ability. Therefore the first example is direct and

the second is indirect (Searle 1979). Although the two sentences have the same meaning, they have different grammatical moods (imperative and interrogative respectively - these are discussed in more depth in section 5.1.3) because those moods can facilitate different intentions.

There are five different illocutionary forces:

— Assertives: statements of fact or opinion with an active role taken by the speaker;
— Directives: statements to elicit an action from someone else;
— Commissives: statements displaying a commitment to a task by the speaker;
— Expressives: statements with an expression of emotion or psychological state felt by the speaker;
— Declaratives: statements of fact or opinion without an active role by the speaker.

As a side note, our taxonomy (based on the DIT++) includes speech acts that would be considered directives or commissives in SAT: *instruct* and *offer*. Other illocutionary forces other than directives and commissives are present in our taxonomy, e.g., *inform* behaves often like an assertive or declarative, and *support* behaves like an expressive.

The five categories of illocutionary forces specified in SAT are not exclusive or exhaustive. e.g., a commissive-directive act exists in which a leader requests and action that he or she also takes part in. Therefore the illocutionary forces in SAT are not exclusive or rigid (Hussein et al. 2012). Also of interest to note is that SAT can be understood to be interrelated to SL®. e.g., the act of *explaining* is tied to coaching leadership, and explanations can be expressed with declarative illocutionary force.

Also note that these illocutionary forces can be present at varying degrees. For example, assertives can be statements of fact or of hypothetical situations (Searle 1979).

The five grammatical moods are not indicators of illocutionary force. That is to say, the communicative intention behind speech does not necessarily dictate whether the utterance is an imperative, interrogative, or indicative. There are no set rules about which structures belong to certain forces or acts as specified by SAT (Wilson and Sperber 2012). Previous research in SAT in conjunction with natural language processing (NLP) has proven this lack of direct connection between force and structure as well (Mann 1980). For example, examine the sentence "You are standing on my hand." This sentence is indicative in form. However, the utterance is actually a directive because the speaker utters the declarative for the purpose of getting the addressee to step off his or her hand. This is an indirect speech act. The example above would suggest that extracting illocutionary force using syntax or word sequencing is not a valid method, especially for indirect statements.

SAT also talks about the relationships between words and parts of sentences, also called dependencies. Dependencies within a sentence can be graphically represented by a dependency tree, such as the one in Figure 5.1. Dependency trees are discussed in more detail in section 5.3.4, but essentially, this tree represents the different clauses in the sentence.

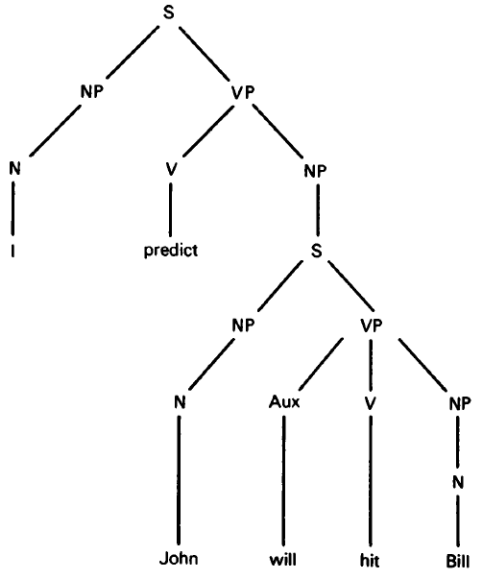


Figure 5.1 – The structure of the assertive-declarative indicative sentence "I predict John will hit Bill" (Searle 1979, pp. 20–27).

There are some interesting semantic points in SAT, such as those associated with the word "please": when "please" is added to a declarative directive sentence, the sentence then becomes clearly directive. Thus it is essentially an indirect sentence by using the indicative mood to convey an order or request (Searle 1979, p. 40).

In the following section, we discuss previous work on linguistics in more detail than was covered in section 2.5.3.

5.1.3 Previous work

In this section, we examine works on linguistics which study how language can convey intentions in different ways. The agent has to convey sometimes the same intentions with both low and high task and relationship behavior, which correspond to different intentions.

A speech act is the combination of communicative intention and the proposition (the content of the speech) (Allwood et al. 1992; Mann 1980). A communicative action is then

the form that a speech act takes. Therefore, some authors consider a speech act to be a mood (Allwood et al. 1992). When this assumption is made, intention does not have to match the speech act (making the utterance indirect) (Allwood et al. 1992; Searle 1979). It was found that the task of identifying illocutionary force from semantic and syntactic features of utterances was much more difficult when working with indirect utterances. For example, the declarative "You are standing on my hand" is actually a directive "Get off of my hand." Regardless, Mann found that it was possible to detect illocutionary force from text. Other research has established that indirect statements are not indicative of good leadership (Pluwak 2016).

However, there is still a strong link between intentions, speech acts, and semantics. In 1987, a dictionary of words was established that indicated the presence of various speech acts, confirming that there are many cases in which semantic content is directly linked with communicative form (Wierzbicka 1987).

Research has also been conducted on automatically identifying speech acts from speech and text (Felice and Deane 2012; Vosoughi and Roy 2016). In 2012, researchers built a computational model using a maximum entropy classifier for speech act identification that achieved an accuracy of 79.28%. This model was built using both native and non-native English speakers' writings. The speech acts that were searched for were *requests*, *orders*, or *commitments* (Felice and Deane 2012).

Later, in 2016, another speech act classification algorithm was built to identify speech acts from Twitter (Vosoughi and Roy 2016). The authors aimed to detect assertion, recommendation expression, question, request, and miscellaneous speech acts. The model uses both semantic features, such as opinion words, vulgar words, and n-grams. An n-gram is a common sequence in a string made up of n words. For example, in the string "You are standing on my hand", the 3-grams are "You are standing", "are standing on", "standing on my", and "on my hand." They also studied syntactic features such as dependency sub-trees and punctuation, and their findings include things like the phrase "I think" indicating an expression and the phrase "could you please" indicating a request. Vosoughi and Roy found that their logistic regression model was 70% accurate.

In 2018, a dialogue act classification algorithm was developed using a variety of machine learning techniques on both manually annotated data and machine-annotated data with an aim to improve the classification of dialogue acts in machine-annotated data. Using a random forest model on n-grams between 2 and 7, an accuracy of 66.87% was achieved (Malik et al. 2018), solidifying n-grams as a viable way of detecting dialogue or

speech acts.

A difficult part of extracting speech acts from text is when the text is indirect: when the intention does not match the form. This was tried in opinion mining where indirect statements are analyzed through semantics and syntax to extract opinions with Wordnet [1], FrameNet [2], and SenticNet [3] (Pluwak 2016). Others have used speech act detection for sentiment analysis as well (Ensink and Sauer 2003; Lakoff 2002).

Lastly, the original work on SL® contains guidelines as well for speech in each leadership style. High task behavior should contain precise speech and imperative utterances (Hersey et al. 1988). High relationship speech should create a sense of autonomy for the follower, which means not using direct orders as orders do not allow for the follower to make a choice about whether or not they will comply with a request. An example given is a sentence that begins with "I'd appreciate it if you..." (Hersey et al. 1988). The leadership descriptors in Table 2.1 can be thought of as speech acts themselves, with words like "telling" depicting how the speaker behaves and what the speaker intends for the listener to do or understand (Hersey et al. 1988).

Now that some context has been provided for how linguistics can convey intentions and meaning in utterances and text, we explain our methods for finding linguistic rules for an agent leading a medical procedure. The first step is to gather examples of medical leader speech.

5.2 A dataset of medical leader speech

The first step of identifying linguistic rules for agent speech in each leadership style involves creating and annotating a speech dataset, which we detail in this section. In section 5.2.1, the creation of the dataset is detailed, and in section 5.2.2, the process of annotating the dataset is explained.

5.2.1 Creation of the dataset

While there are existing corpora that involve team communication (Anikina and Kruijff-Korbayova 2019; Litman et al. 2016), none contain speech specific to the emergency room. Therefore, an emergency room-based dataset was necessary.

1. https://wordnet.princeton.edu/
2. https://framenet.icsi.berkeley.edu/fndrupal/
3. https://sentic.net/

The dataset created contains coordinator speech from a variety of emergency room simulation and training videos as well as some previous literature (a list of which is available in Appendix C). The videos chosen were made by university hospitals and made to be examples for medical professionals-in-training in emergency and non-emergency medicine. By choosing only videos from university hospitals, we ensured that the behaviors of the medical staff in the videos was verified as being desirable behaviors. Each video lasted anywhere from two to ten minutes long. Parts that closely resembled parts in other videos that were already added to the dataset were skipped in order to maintain variety in the dataset.

Each video includes a leader, either a lead nurse (a medical coordinator) or a doctor, and has one to three different followers. Videos were watched, and speech from the medical leader was manually copied into text format using video subtitles.

Previous work has discussed "splitting points" when working with corpus text as a pre-processing measure (Kruijff-Korbayova et al. 2015; Weisser 2018). Splitting points are points at which it makes sense to split an utterance or a string so that the individual part can be examined separately. Commonly, corpora text is split into segments logically based on the information that needs to be gathered. In the case of our research, mood and communicative intentions can be apparent in whole sentences and also sentence segments, and so state changes and commas which already separate an utterance into two complete sentences are used as splitting points.

The speech spoken in the videos was split up initially by complete utterance. Complete utterances were separated by long pauses, others speaking, and changes of situation state (e.g., (1) before a patient receiving fluids and (2) afterward). These utterances were then separated by complete sentence and then further split into individual segments if the sentence contained multiple subject-verb pairs.

For example, one utterance in the dataset is "Okay, great. Let's continue with normal saline at 250 mls an hour. Let's do a stat chest x-ray, ECG, and CCU bloodwork. We will arrange a bed in ICU, and great work, everyone. Good job." Below are the complete sentences that make up this utterance:

1. Okay, great.
2. Let's continue with normal saline at 250 mls an hour.
3. Let's do a stat chest x-ray, ECG, and CCU bloodwork.
4. We will arrange a bed in ICU, and great work, everyone. [sic]

5. Good job.

The segments that make up this utterance include:

1. Okay, great.
2. Let's continue with normal saline at 250 mls an hour.
3. Let's do a stat chest x-ray, ECG, and CCU bloodwork.

4a. We will arrange a bed in ICU,

4b. and great work, everyone.

5. Good job.

As shown, sentence 4 was split up into two segments because the sentence comprises two different subject-verb pairs, and consequently, two different non-technical skills. Segment 4a contains a complete sentence with the subject-verb pair "we, will arrange". Segment 4b, on the other hand, is a sentence fragment and is understood to be a shortened form of "You did great work". Therefore, the subject-verb pair is an implied "you, did".

All utterances were assigned the following according to the taxonomy presented in chapter 3, using the context in the videos rather than from the utterances alone.

— Non-technical skill;
— Element;
— Communicative intention;
— Speech act.

Because the non-technical skills, elements, etc. are not ambiguous or open to interpretation given the context surrounding each utterance and our taxonomy is based on others with high interrater agreements (Bunt 2009; Montenegro et al. 2019), these labels were completed by the PhD student.

Each utterance was also assigned a mood. In our dataset, there are not many examples of the subjunctive or conditional mood. Additionally, the same verb forms can be used between those three moods. For example, examine the phrase "I get sunburned" in these sentences: "I get sunburned in hot weather" (indicative) and "If it's hot, I get sunburned" (conditional). The phrase "I get sunburned" does not change. Therefore, for the purposes of our research, the conditional, subjunctive, and indicative moods are all grouped together under *indicative*. Finally, each utterance was assigned a label indicating whether it was direct or indirect, as explained in section 5.1.2.

The dataset contains 294 complete utterances that are expanded to 375 total segments, whole sentences, and complete pieces of speech due to the splitting points mentioned earlier in this section. The size of the dataset is in line with select other research in which very specific data is examined: In Song et al.'s work, two datasets of 377 utterances and 500 utterances, each with five labels, was used to train a classifier for speech recognition (Song et al. 2014). In Mahabal et al.'s work, only 320 training examples were used for text classification (Mahabal et al. 2020). Furthermore, in Malandrakis et al.'s work, text generation is used to build a dataset of virtual agent speech; this dataset includes as few as 8 samples per category (Malandrakis et al. 2020).

Table 5.1 displays the distribution of non-technical skills and speech acts within the dataset. There are some expected patterns. e.g., the non-technical skill *situation awareness* contains many examples of *inform* and *request information* while the skill *task management* contains many examples of *instruct*.

	instruct	inform	offer	request information	respond	support
Situation Awareness	0	77	0	42	15	0
Decision-making	48	0	0	0	0	0
Task Management	168	6	0	0	0	0
Team Management	0	7	1	0	0	11

Table 5.1 – Distributions of the speech acts and non-technical skills within the dataset.

The taxonomy presented in chapter 3 contains communicative intentions and speech acts that are not specific to an emergency, such as *offer* (although they are specific to medical procedures), and so it is not unusual that some do not appear often in the dataset.

In the next section, we discuss how the dataset was annotated with leadership style.

5.2.2 Annotating the dataset

As mentioned at the beginning of section 5.2, the dataset that we created was also annotated. In order for situational leadership to lead to a more successful procedure, the followers must be able to perceive leadership style in the sense that they perceive different levels of task and relationship behavior. Therefore, having multiple people assign leadership style to each utterance was necessary. An agreement among annotators indicates

that the leadership style of that utterance is largely perceivable, whereas a disagreement indicates the opposite.

Four people were chosen as annotators, one woman and three men, all between the ages of 21 and 29. All four were native English speakers, from the US and from Ireland, and had a minimum education level of some college experience. None of them had any experience in the medical field. This medical inexperience was intentional so as to ensure that the leader speech rules decided by the analysis in this paper would be applicable to novice caregivers. The three annotators between the ages of 27 and 29 all had significant work experience working under a boss while the annotator who was 21 did not have much work experience.

The annotators were given the following information about situational leadership: (1) the definitions of task and relationship behavior, (2) the definitions of each leadership style, and (3) a list of the original descriptors for each leadership style, available in Table 2.1 (Hersey et al. 1988). Annotators were not given any contextual information regarding each utterance. They were asked to assign a leadership style to each utterance of the dataset. The order of utterances was randomized for each annotator to ensure that it did not affect the results as much as possible. An example of the annotators' ratings is in Table 5.2. Recall that S1 refers to directing leadership (high task and low relationship behavior), S2 refers to coaching leadership (high task and high relationship behavior), S3 refers to supporting leadership (low task and high relationship behavior), and S4 refers to delegating leadership (low task and low relationship behavior).

	Utterance	**Ann. 1**	**Ann. 2**	**Ann. 3**	**Ann. 4**
1	Okay, let's resume compressions and administer epinephrine 1 milligram push.	S2	S1	S1	S1
2	Hold your hands in position.	S1	S1	S1	S1
3	Tubes going in.	S4	S4	S1	S1
4	I'm your primary nurse.	S3	S4	S4	S4
5	Inline stabilization: who's going to... ?	S3	S2	S2	S4

Table 5.2 – An example of 5 utterances from the dataset and their leadership labels as given by each of the four annotators.

Note that each annotator annotated the dataset independently of the other annotators. There was no annotation comparison phase carried out in which annotators discussed their decisions amongst themselves to either come to a consensus or discuss reasons for why they

differ. At the time of this analysis, we felt it was most important to find the utterances that annotators agreed-upon based on their own opinions without discussing with others. Because annotation comparison was not carried out, we do not have information regarding why annotators disagreed when they did.

Once annotation on the dataset was complete, analysis could be conducted to see where annotators agreed and disagreed with regard to leadership style. The following section details the analysis conducted on the annotated dataset.

5.3 Annotation analysis

In order to define linguistic rules of speech belonging to each leadership style, we need to find linguistic patterns among the utterances that were labeled with each style. Finding patterns in a set of strings is one of the principal aspects of natural language processing (NLP) and involves examining characteristics of those strings. These characteristics include individual words, parts-of-speech (POS) tags, and sentence structure (Searle 1979). Breaking up a sentence into these various components leads to different understanding of the sentence: its semantic meaning and its syntax. Each kind of analysis leads to different valuable insights.

The majority of work on pattern discovery in utterances and in text is semantic in nature, often with syntax analysis used as a tool for semantic classification with clustering and similarity measures (Brody 2005; Oliva et al. 2011; Özateş et al. 2016; Stevenson and Greenwood 2009; W. Wang and Pan 2019) or by examining semantic meaning directly from the meaning of individual words and phrases with word and sentence embeddings, clustering, similarity measures, and bag-of-words models (Colace et al. 2014; Khoury 2012; Popat et al. 2017). That is to say that many works define similarity of sentences as the similarity of the meanings.

While semantics can be very useful for understanding some intentions (Searle 1979; Wierzbicka 1987), we also want to focus on syntactic elements to find out whether sentence structure impacts intention. As discussed in section 5.1, the communicative intention can influence form, so it is this relationship that we wish to explore. Additionally, the videos that the dataset uses contain a variety of different procedures in order to avoid defining linguistic rules that might be specific to one procedure or another. Thus the vocabulary in the dataset is very broad and analyzing those utterances semantically could lead to results based on procedure-specific vocabulary, which we want to avoid.

In this section, we detail the methods we used to discover structural patterns from the dataset using linguistic labels and components from our taxonomy. We discuss our statistical analysis methods in sections 5.3.1 and 5.3.2, discuss clustering methods in section 5.3.3, discuss dependency parsing methods in section 5.3.4, discuss chunking methods in section 5.3.5, compile a list of rules resulting from all the annotation analysis in section 5.3.6, and finally examine individual annotators' responses in section 5.3.7.

5.3.1 Agreement analysis

In order to evaluate how the annotators assigned leadership style to each utterance in the dataset, the Fleiss kappa statistic is used. Fleiss kappa measures interrater reliability when there are more than two raters and results in a value that shows how much more those raters agree as opposed to a random selection of values (Landis and Koch 1977). Therefore the Fleiss kappa statistic is a more nuanced value than simple agreement percentage that takes into account the fact that a random selection of values from each rater would result in some agreement. The original interpretations of Fleiss kappa statistics are: (1) 0.21-0.40: fair agreement, (2) 0.4-0.60: moderate agreement, (3) 0.61-0.80: substantial agreement, and (4) 0.81-1.00: almost perfect agreement. The Fleiss kappa statistic for all annotators over the whole dataset was 0.415, indicating moderate agreement.

The dataset was then analyzed according to mood, length, directness, and speech acts. In this section, the analysis was done on the entire annotated dataset and included the Fleiss kappa statistic of agreement for utterances based on these characteristics (mood, length, directness, and speech acts).

Mood

Table 5.3 displays the Fleiss kappa statistics for the annotated dataset split by both leadership style and by mood. Note that an imperative containing "let's" ("let us") is often interpreted differently in English than imperatives with other verbs (e.g., "Let's go home" versus "Go home"; the first implies that the speaker is involved whereas the second does not imply involvement by the speaker (Goddard 2002; Searle 1979)). Therefore imperatives are split into two groups: one containing "let's" as the imperative verb and one not containing "let's" as the imperative verb.

The utterances included in this analysis are only those which have one mood, not more than one. Out of the 375 utterances in the dataset, only 328 contain just one mood

with 47 containing more than one grammatical mood. This is why the Fleiss kappa differs slightly between these 328 utterances (kappa of 0.404) to the kappa statistic for the entire dataset (kappa of 0.415).

	Imperatives			**Interrogatives**	**Indicatives**
	all	*with "let's"*	*without "let's"*		
Directing	0.187*	0.000	0.274*	-0.008	0.264*
Coaching	0.217*	-0.008	0.326*	0.126*	0.374*
Supporting	-0.043	-0.036	0.256*	0.075	0.310*
Delegating	0.111	0.094	0.010	0.097	0.547*

**p-value < 0.05*

Table 5.3 – The Fleiss kappa statistics of utterances that only included one sentence structure each, for a total of 328 utterances (73 imperatives, 44 without "let's"; 76 interrogatives; 178 indicatives). The overall Fleiss kappa is 0.404.

The annotation results in Table 5.3 indicate imperatives with "let's" are more ambiguous than those without - there is more agreement among annotators in terms of imperative utterances without "let's" than those with "let's". Annotators generally agreed upon the leadership style of interrogative utterances the least and generally agreed upon the leadership style of indicative utterances the most.

There are only 22 interrogative utterances that were agreed upon, all of which are part of coaching style. Two examples are below:

1. Okay, can someone get a hold of the family and inform the attending physician as well for me, please?
2. Okay, can someone tell me what happened here, please?

Utterances 1 and 2 are indirect orders: the communicative form of the utterances is an interrogative which seeks information, yet the communicative intention is an order. The context of the videos that they appeared in confirms this but so does the fact that they both contain the word "please" which indicates a directive. All of the agreed-upon interrogative utterances have speech act *instruct*, yet all of them are in question format, which invites a dialogue between the speaker and the hearer. This invitation to respond establishes the hearer's autonomy, and establishment of autonomy of another person can be considered high relationship behavior (Goddard 2002; Searle 1979), which is confirmed by the annotators' agreement that utterances such as these contain high relationship behavior according to situational leadership.

Looking at Table 5.3, we can see that annotators most agreed upon indicative utterances that they labeled with delegating leadership. This could indicate that utterances with the indicative mood are likely to be understood as being spoken with low-task and low-relationship behavior.

Additionally, utterances that contain more than one mood are almost entirely concentrated in coaching style. The individual sentences and segments that make up these longer utterances are labeled as imperatives and indicatives, and individually, they are labeled with all four leadership styles. This indicates that an utterance containing a mix of moods and leadership styles will make the overall utterance coaching. In order to explore this, utterances are examined by length.

Length

Table 5.4 displays the Fleiss kappa statistic for all utterances in the dataset, separated by segments only, whole sentences, and the utterances longer than one sentence.

	Total dataset	**Segments**	**Whole sentences**	**Utterances longer than one sentence**
Directing	0.356*	0.297*	0.386*	0.263*
Coaching	0.428*	0.321*	0.438*	0.388*
Supporting	0.297*	0.252*	0.312*	0.299*
Delegating	0.533*	0.576*	0.499*	0.634*

**p-value < 0.05*

Table 5.4 – The Fleiss kappa statistics for the four annotators on the entire dataset (375 utterances in total, Fleiss kappa of 0.415), split into individual segments only (of which there are 79 in total), whole sentences (251 in total), and utterances that are longer than one whole sentence (45 in total)

The median and mean character lengths of all utterances in the dataset are 42.0 and 63.2693 respectively. The utterances were then split up based on whether they were agreed upon (the agreement group) or not (the disagreement group). In the agreement group, the median is 44.0 characters long and the mean is 77.8740 characters long. In the disagreement group, the median is 41.0 characters and the mean is 55.7903, resulting in a difference of about 22 characters between utterances that were agreed upon and those that were not. In order to explore this difference, a Fisher's f-test and a Welch two-sampled t-test is completed between the number of characters in the agreement and disagreement

groups.

The Fisher's f-test returns a significant value of 0.1863 (p-value < 0.05). The Welch two-sampled t-test returns a significant t-statistic of -2.0165 (p-value < 0.05), indicating that the means are significantly different, and so we conclude that longer utterance length is easier for people to interpret and agree on.

Additionally, utterance length was found to depend on leadership style. A one-way ANOVA was also completed between the utterances in the agreement group (p-value < 0.05) which shows that the means of the character lengths between each leadership style are not the same. According to the results in Table 5.4, directing, coaching, and supporting leadership should be most expressed as single, whole sentences. However, delegating utterances containing more than one sentence are more agreed upon, and so utterances coming from a delegating leader could be better if they are more than one sentence long.

Directness

Recall that directness refers to whether an utterance's literal meaning is different than its contextual meaning; an indirect utterance is one where the communicative intention and communicative form do not match. Table 5.5 displays the Fleiss kappa statistics among the annotators with regard to directness of the utterance and each leadership style. Among the 375 utterances in the dataset, there are 24 that contain both a direct and an indirect segment, and so these utterances are excluded from the analysis.

	Direct	**Indirect**
Directing	0.406*	0.155*
Coaching	0.253*	0.265*
Supporting	0.294*	-0.022
Delegating	0.499*	0.072

**p-value < 0.05*

Table 5.5 – The Fleiss kappa statistics for each leadership style across direct and indirect utterances. There are 281 direct statements (kappa of 0.377) and 70 indirect statements (kappa statistic of 0.193).

As shown in Table 5.5, direct utterances are generally more agreed upon than indirect sentences. However, when the leadership style chosen was coaching, indirect utterances

are more agreed upon, though slightly, than direct utterances. As mentioned earlier when discussing the moods of the utterances in the dataset, there are many interrogatives which are classified as indirect with the speech act instruct. This indicates that interrogatives that direct a follower portray coaching leadership. The following section explores how other speech acts affect the assignment of leadership style.

Speech acts

Table 5.6 displays the Fleiss kappa statistics among the annotators with regard to speech acts of each utterance and each leadership style.

	instruct	**inform**	**offer**	**request information**	**respond**	**support**
Directing	0.427*	0.294*		-0.081		-0.031
Coaching	0.531*	0.396*	-0.500	-0.207*		0.593*
Supporting	-0.016	0.099*	-0.500	-0.088	-0.204	0.455*
Delegating	0.180*	0.555*		-0.029	-0.204	-0.138
Totals	*168*	*138*	*1*	*42*	*15*	*11*

*p-value < 0.05

Table 5.6 – The Fleiss kappa statistics for each leadership style and speech act. The Fleiss kappa statistic for the entire dataset is 0.415.

As shown in Table 5.6, not all kappa statistics are significant, and some are low just because the speech acts are not all distributed evenly throughout the dataset. Speech acts *offer*, *support*, *request information*, and *respond* do not show up often within the dataset, which indicates that they would not present themselves often during an emergency procedure, although there is a possibility that this is due to the size of the dataset.

The vast majority of the dataset consists of utterances encompassing speech acts *instruct* and *inform*. Because the dataset was compiled from emergency room simulation videos that did not contain many examples of *offer*, *support*, *request information*, and *respond*, a leader of a medical emergency should use fewer utterances with those speech acts.

It is also important to note the differences between speech acts in each leadership style in Table 5.6. As expected, annotators had moderate agreement on utterances that contained the speech act *instruct* and agreed that these utterances often belonged to directing and coaching utterances, which both contain high task behavior. There are

other patterns that confirm SL®, too, such as utterances with speech act *support* often being agreed on as belonging to coaching or supporting leadership, which both contain high relationship behavior. By understanding how the speech acts then relate to each leadership style, we can better design speech for those leadership styles.

Agreement analysis results

The results of all the statistical analysis performed thus far in this chapter on the annotated dataset are available in Table 5.7.

	Directing	**Coaching**	**Supporting**	**Delegating**
Mood	Imperatives without "let", Indicatives	Interrogatives, Indicatives	Indicatives	Indicatives
Length	whole sentences	whole sentences, 2+ sentences	whole sentences	whole sentences, 2+ sentences
Directness	Direct	Direct, Indirect	Direct	Direct
Speech acts	instruct	instruct, inform, support	support	inform

Table 5.7 – A list of rules generated with statistical analysis on the annotated dataset.

5.3.2 Analysis of task and relationship behavior

Because SL® is composed of both task and relationship behavior, we also wanted to see whether annotators agreed more based on one or the other. This would help us understand whether annotators primarily considered either task or relationship behavior to be the determining factor of leadership style. Again, the entire annotated dataset is examined in this section.

To examine the differences in annotation of both types of behavior, the leadership styles that annotators assigned each utterance in the dataset were grouped twice, once based on task behavior and once based on relationship behavior. For example, examine the five utterances from the dataset in Table 5.2. Only utterance 2 is in agreement among all annotators. However, when grouped by relationship behavior (low-relationship leadership styles directing and delegating are combined and high-relationship leadership styles

coaching and supporting are combined), we can see that utterance 3 is also in agreement since all annotators marked that utterance as having low relationship behavior (see Table 5.8). When grouped by low and high relationship behavior in this way, the kappa dropps to 0.362 (p-value < 0.05) from 0.415 for the whole dataset when leadership styles are not grouped by relationship behavior.

	Utterance	**Ann. 1**	**Ann. 2**	**Ann. 3**	**Ann. 4**
1	Okay, let's resume compressions and administer epinephrine 1 milligram push.	High	Low	Low	Low
2	Hold your hands in position.	Low	Low	Low	Low
3	Tubes going in.	Low	Low	Low	Low
4	I'm your primary nurse.	High	Low	Low	Low
5	Inline stabilization: who's going to... ?	High	High	High	Low

Table 5.8 – Five utterances in the dataset annotated by four annotators. Their assignments of leadership style are grouped by either low or high relationship behavior.

When grouped by task behavior as shown in Table 5.9 (high-task leadership styles directing and coaching combined and low-task leadership styles supporting and delegating combines), we can see that in addition to utterance 2, utterance 1 is agreed upon since all annotators marked that utterance as having high task behavior, and also utterance 4 is agreed upon since all annotators marked it as having low task behavior. Utterances 3 and 5 do not have any agreement when grouped because annotators did not agree on high or low relationship or task behavior for those utterances. When grouped by low and high task behavior in this way, the kappa jumps to 0.570 (p-value < 0.05) from 0.415 for the whole dataset when leadership styles are not grouped by task behavior.

	Utterance	**Ann. 1**	**Ann. 2**	**Ann. 3**	**Ann. 4**
1	Okay, let's resume compressions and administer epinephrine 1 milligram push.	High	High	High	High
2	Hold your hands in position.	High	High	High	High
3	Tubes going in.	Low	Low	High	Low
4	I'm your primary nurse.	Low	Low	Low	Low
5	Inline stabilization: who's going to... ?	Low	High	High	Low

Table 5.9 – Five utterances in the dataset annotated by four annotators. Their assignments of leadership style are grouped by either low or high task behavior.

These results indicate that annotators agree more on indicators of task behavior than those of relationship behavior and imply that indicators of relationship behavior may be more unique to individual followers. Individual annotators' responses are examined further in section 5.3.7. The difference in kappa statistics between the datasets grouped by task and relationship behavior may also indicate that annotators assign leadership style by first identifying the level of task behavior in the utterance and then secondly identifying the level of relationship behavior.

Finally, we explore whether annotator demographics had any effect on the Fleiss kappa.

Effects of annotator demographics

We first explored whether there were any patterns in how annotators assigned leadership style in terms of age/work experience. The Fleiss kappa statistic for just the male annotators was 0.433 (p-val < 0.001), which is not much higher than the overall kappa statistic of 0.415 and still indicates moderate agreement. The kappa for the three annotators aged 27-29 with significant work experience was 0.387 (p-val < 0.001), indicating fair agreement only rather than moderate agreement. This may indicate that age or significant work experience does not lead people to agree more on what leadership speech looks like.

When the male annotators' ratings were grouped by task behavior, the agreement among them was 0.536 (p-value < 0.001), down from 0.570 for all four annotators' responses were grouped by task behavior. When grouped by relationship behavior, the kappa statistic was 0.397 (p-value < 0.001), higher than 0.362 when all four annotators' responses were grouped by relationship behavior. This might suggest that indicators of relationship behavior change depending on gender since men agreed more on what relationship behavior looks like in medical leader speech. However, the agreement is still rather low, which again points to relationship behavior being very individual.

When the responses from the older annotators with more work experience were grouped by task behavior, the kappa is 0.56 (p-value < 0.001), very close to the statistic 0.570 for all four annotators' responses were grouped by task behavior. When grouped by relationship behavior, the kappa is 0.312 (p-value < 0.001), lower than 0.362 when all four annotators' responses were grouped by relationship behavior. This might suggest that age and/or work experience has little effect on what annotators consider to be task and relationship behavior.

Some of these individual differences are examined in section 5.3.7. However, more research is needed to understand how individuals perceive relationship behavior and how

varying levels of task and relationship behavior influence a follower's performance during a task.

While we gathered some valuable insights by examining the annotated dataset statistically, we move onto other methods to discover further patterns between each leadership style.

5.3.3 Clustering to find common sequences

To further identify linguistic rules in the form of patterns that exist within each leadership style, we use k-means clustering (Berber Sardinha and Veirano Pinto 2021; Colace et al. 2014; Khoury 2012). In the rest of this section, each utterance in the annotated dataset is referred to as a string. The goal is to identify patterns among the agreed-upon strings in the annotated dataset and then check whether those patterns are indicative of one leadership style. First, we examine some prior works which use clustering for a similar purpose.

In this section, only the agreed-upon strings in the dataset, resulting in 127 total strings, are examined. The strings that were not agreed upon in terms of leadership style are not used.

In previous research, multi-dimensional clustering was performed in which multiple variables within the set of strings are used for identifying sameness (Berber Sardinha and Veirano Pinto 2021). In this work, TV program scripts were examined and communicative functions were used as one parameter. The authors argue that communicative functions require multivariate analysis because "of the complex co-occurrence of many different linguistic characteristics across the texts". However, these texts involve a much broader spectrum of different registers (genres), whereas the text in this research comes from a single register (the medical emergency room). Therefore, multi-dimensional analysis is not as necessary.

Common sequences within a string, or n-grams, have been used in natural language processing for many different purposes. For example, n-grams have been used to automatically classify dialogue acts with machine learning techniques (Malik et al. 2018) and have been used to associate semantics with meaning (Vosoughi and Roy 2016).

Clustering has been shown to be particularly effective when using the common occurrence of n-grams to find latent characteristics within a set of strings (Berber Sardinha and Veirano Pinto 2021; Colace et al. 2014). In order to cluster, a similarity measure is needed, which compares two strings to determine how similar they are to each other. Two

similarity measures are often used: edit distance (also called Levenshtein distance) and cosine similarity. Edit distance is the minimum number of edits (insertions, deletions or substitutions) required to change one string into another (Levenshtein 1966). In a string similarity context, edit distance takes into account the order of words, and has been used for both semantic and syntactic similarity (Babur and Cleophas 2017; Ferreira et al. 2016; Kondrak 2005; R. Wang and Neumann 2007).

Cosine similarity determines the cosine between two vectors. Common words or n-grams are identified within a group, and then each string is turned into a binary numeric vector that is the length of the set of words or n-grams and is composed of 0s and 1s based on the presence of those common words or phrases. For example, 2-grams that show up at least three times throughout the whole set of strings might be used (Colace et al. 2014; Oliva et al. 2011; Özateş et al. 2016; R. Wang and Neumann 2007). This method is similar to a bag-of-words model: word order does not matter. Just like edit distance, cosine similarity has been used for semantic and syntactic purposes (Babur and Cleophas 2017; Colace et al. 2014; Khoury 2012; Lin and Wu 2009; Oliva et al. 2011; Özateş et al. 2016; Popat et al. 2017).

We perform clustering on both raw strings and part-of-speech (POS) tags. A raw string is defined as a string in almost its original form with contractions expanded and punctuation intact. Looking at a raw utterance lends itself often to semantic classification using the meanings of individual words to understand the meaning of the entire utterance (Colace et al. 2014; Mahabal et al. 2020).

POS tagging can be done with a number of NLP packages, but we chose Stanford CoreNLP which uses the Penn Treebank [4]. Stanford CoreNLP comes pre-trained, although there is an option to train it on specific data. Each token, or word, in the string is given a POS tag that represents the part of speech that the word assumes in that particular string, as a word might assume a different POS tag depending on the context of the sentence. In order to cluster on POS-tagged utterances effectively, the original words are removed, leaving only the tags and punctuation. A string can be analyzed in a syntactic sense as well by examining what types of words appear in certain orders (Ferreira et al. 2016; Mann 1980; Oliva et al. 2011; Pluwak 2016).

The sum of squared differences (SSD) is used to determine the number of optimal clusters in the dataset, with each cluster containing one or more unique or prevalent patterns that only the strings within that cluster share. The SSD is a mathematical way

4. `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`

to find the clusters with the least amount of variance within them to ensure they contain similar strings. The strings are then clustered with k-means into the optimal number of clusters based on the presence of individual terms or n-grams within each string as explained above. The resulting clusters that contain a single (or nearly a single) leadership style are examined, and the n-grams that define each cluster then define the linguistic rules for each leadership style.

For the purpose of brevity, only the successful similarity measures using both cosine similarity and edit distance are discussed in detail, although some methods that were tried without success are mentioned. Only sequences that do not include medically-specific language are included, as we are examining how structure relates to leadership style, not how utterances specific to one particular procedure are indicative of leadership style.

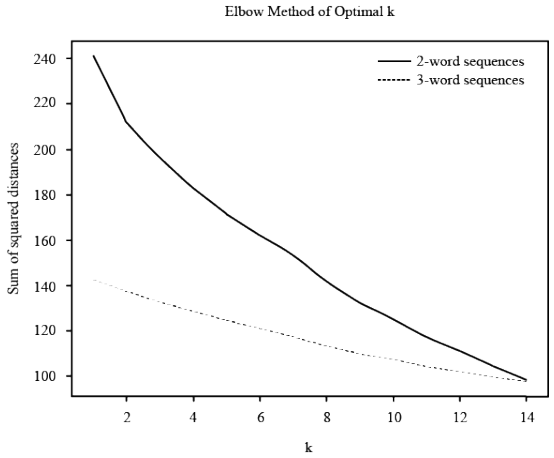
Supervised machine learning, such as Siamese networks, one-shot learning, and random forests, are also valid methods of learning the properties within suspected groups in a dataset (Malik et al. 2018; Song et al. 2014; Vosoughi and Roy 2016; W. Wang and Pan 2019). Siamese networks and one-shot learning were originally used for learning the properties within each leadership style, but the low number of examples coupled with the variation in vocabulary prevented meaningful results from being obtained.

Using the 127 agreed-upon strings from the dataset, latent patterns are discovered with clustering. Figures 5.2a- 5.2c display the optimal k-means values for the methods that lead to successful results.

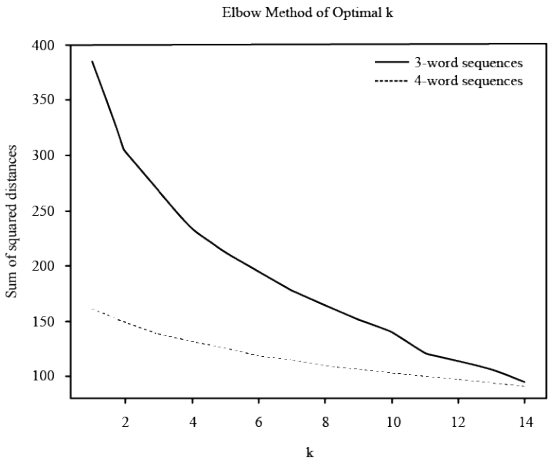
Cosine similarity is used exclusively for clustering on raw utterances, as edit distance was found to be ineffective and led to high SSD values and clusters that did not make human sense. This is likely due to the specialized vocabulary present throughout the dataset.

Figure 5.2a displays the k values when common 2-grams and 3-grams are examined from raw utterances. For the first trial, this means that all 2-grams present in the entire dataset were found, and then only those 2-grams that appeared a minimum of three times throughout the dataset are added to a list, resulting in 161 total 2-grams. Each string in the dataset was then represented by a vector made up of 0s and 1s that is 161 numbers long, with 0 indicating that the specific sequence was not present in the string and a 1 indicating that the specific sequence was present in the string.

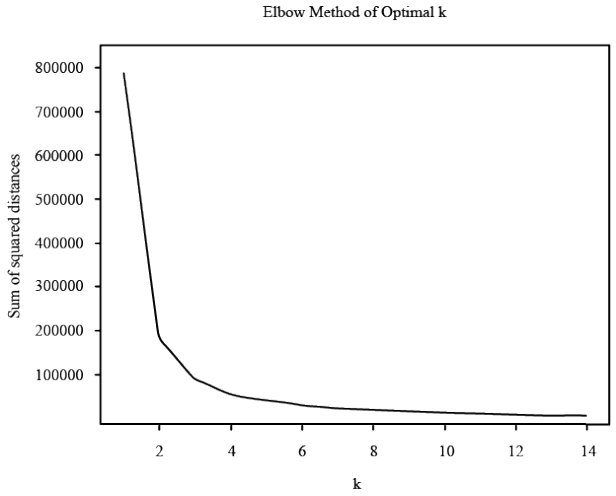
For the 2-grams trial, the strings in the dataset were organized into four and fifteen separate clusters separately. When the strings in the agreement group are organized into four clusters, only two of them appear to follow the leadership styles: Cluster 0 includes



(a) The optimal k values when the entire agreement group is analyzed with cosine similarity based on 2-grams and 3-grams that appear at least 3 times with procedure-specific sequences removed from the list of 3-grams. This resulted in 161 (at k=4, the SSD=171.2239, and at k=15, the SSD is 98.1565) and 79 total sequences (at k=8, the SSD=62.9443) respectively.



(b) The optimal k values when the entire agreement group, POS tags only, is analyzed with cosine similarity based on 3-grams that appear at least 7 times (for a total of 25 sequences; at k=4, the SSD=212.1615) and 4-grams that appear at least 3 times (for a total of 79 sequences; at k=8, the SSD=77.0983). Procedure-specific sequences were removed.



(c) The optimal k values when the entire agreement group, POS tags only, is analyzed with edit distance. At k=3, the SSD=75715.4074.

Figure 5.2 – The SSD of different k values when clustering on the raw utterances of the dataset as well as the utterances' POS tags only. All of these clustering efforts resulted in clusters that provided n-grams present in a specific leadership style.

strings that are all categorized as delegating and include sequence "I see", and cluster 3 only contains interrogative coaching utterances. Additionally, cluster 1 includes coaching and delegating leadership. While this seems to be an odd mix given the combination of task and relationship behavior, the utterances labeled as coaching are long and include segments that were labeled by some or all annotators as delegating within them. Cluster 2 contains a random assortment of strings. No new patterns that were not procedure-specific were discovered from clustering into fifteen clusters.

The process was completed again with 3-grams; the SSD values are shown in Figure 5.2a. Procedure-specific sequences were removed when clustering on 3-grams, resulting in 79 total sequences. The patterns that were discovered were extensions of the sequences discovered from 2-grams clustering. For example, the sequence "can someone" was discovered to be present in utterances only labeled as coaching by the annotators when clustering on 2-grams. When clustering on 3-grams, the sequence "okay, can someone" was found to be indicative of coaching leadership.

Clustering was then completed on POS tags. To account for the abstraction of information that comes with POS-tagging, longer n-grams were gathered than were used for raw utterances (see Figure 5.2b). The eight resulting clusters only affirm the rules that were determined by the clustering done on 2-grams; no new patterns were revealed.

While edit distance was not found to be effective on raw utterances, it was more effective when applied to POS tags. Figure 5.2c displays the SSD values when sentences are clustered according to edit distance. Even though the SSD values are much higher than those obtained when using cosine similarity as the similarity measure, there were some insights when examining the three clusters produced: cluster 1 contains long sentences classified as coaching leadership, with many explanations and various pieces of information. This confirms what was said in section 5.3.1 that longer sentences with a combination of orders and information are often part of coaching leadership.

Table 5.10 displays the patterns for each leadership style found by k-means clustering. The POS tag sequence MD PRP VB refers to a modal verb, a personal pronoun, and a verb (e.g., "can you go"). The sequence PRP MD VB refers to a personal pronoun, a modal verb, and a verb (e.g., "you can go"). The sequence VBZ IN PRP$ refers to a verb in the third person singular present, a preposition, and a possessive pronoun (e.g., "looks like his").

In the following section, we search for linguistic rules for each leadership style by analyzing the utterances' syntactic structures.

	Directing	**Coaching**	**Supporting**	**Delegating**
Sequences of words	"we need to", "I want you", "carry on with", "we will"	"please", "okay, can someone", "for me, please", "as well, please", "please, can we", "can you please", "you can"	"okay, thank you"	"I see that", "it looks like"
Sequences of POS tags		MD PRP VB, PRP MD VB		VBZ IN PRP$

Table 5.10 – A list of rules generated by clustering on the agreed-upon utterances as well as their POS tags. When a sequence of POS tags tended to be a set of specific words, only those specific words were included.

5.3.4 Dependency parsing

Dependency parsing is a method of identifying important structural relationships that extend beyond parts-of-speech, like subjects, predicates, objects, noun phrases, verb phrases, independent and dependent clauses, etc., and studying how they work together within the string. Dependencies within a sentence were first discussed in section 5.1.2 (Searle 1979). While dependency parsing is often used as a first step to semantic understanding (Brody 2005; Jurafsky and Martin 2008; Oliva et al. 2011; Popat et al. 2017; Stevenson and Greenwood 2009; Vosoughi and Roy 2016) and has also been used in conjunction with word embeddings in order to better extract semantic meaning from the sentence (Oliva et al. 2011; Popat et al. 2017; Stevenson and Greenwood 2009; R. Wang and Neumann 2007), the revealed syntactic structures are what we are looking for.

As was done in section 5.3.3, only the 127 agreed-upon strings in the dataset are examined, with the strings that were not agreed upon discarded during analysis.

Stanford CoreNLP has a dependency parsing software that traverses a POS-tagged string and results in a tree which displays the dependencies between each word. As an example, the dependency tree for the string "Hold your hands in position" is shown in Figure 5.3.

As shown in Figure 5.3, the verb "hold" has two leaves, one of which is a direct object ("hands") and one of which is a nominal modifier ("position"). "Hands" contains a possessive nominal modifier "your", and "position" contains a case "in". The numbers refer to the order of the words in the string. The important thing to note here is that in this sentence, there are two dependent parts.

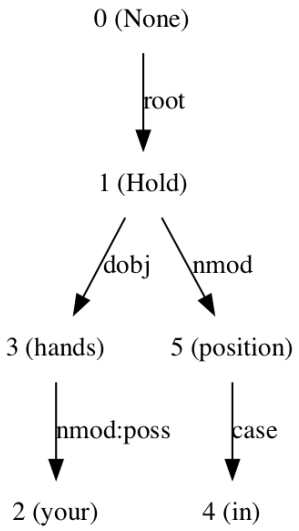


Figure 5.3 – The dependency parse tree for the string "Hold your hands in position".

Dependency parsing can be used to identify clauses and relationships between clauses in text for the purpose of clustering the strings based on common clauses and relationships (Brody 2005). The authors used dependency parse trees in the same way we have used raw utterances and POS tags: finding n-grams and calculating edit distance. We attempted to perform the same methods that we performed in the previous sections; however, the clusters did not make human sense. This is likely due to the use of word embeddings in conjunction with dependency trees in order to extract semantic meaning. However, word embeddings did not make sense for our work because we are only interested in syntax.

Despite the difficulties of clustering on dependency trees, another method of discovering similarity and therefore clustering using the dependency parses of sentences has been established using semantic triples, also known as Resource Description Framework (RDF) triples. RDF triples are a set of three terms in a clause: the subject, predicate, and object. The subject is the part of the sentence that shows who or what is doing the action or who or what is in a state of being. The predicate is what is being said about the subject; it will always include a verb and can also include a subject complement. The object of the sentence receives the action or has the action performed on it (Grammar Wiz 2021). The RDF triples of a sentence go further and describe all relationships between the words of a sentence.

Triples themselves can be generated from dependency parses and tested for similarity

and clustered (Ferreira et al. 2016; Hamoudi 2016; Popat et al. 2017). While they also are used for semantic similarity, they are a great tool for analyzing syntactic structure because they pick up on patterns and characteristics of a string that do not involve meaning or determinate sequences. In other words, RDF triples can display information about dominant or subordinate clauses, active or passive voice, and relationships between two words that can be anywhere in the string instead of needing to be next to each other. For our example sentence "Hold your hands in position", the RDF triples are:

1. ('Hold', 'VB'), 'dobj', ('hands', 'NNS')
2. ('hands', 'NNS'), 'nmod:poss', ('your', 'PRP$')
3. ('Hold', 'VB'), 'nmod', ('position', 'NN')
4. ('position', 'NN'), 'case', ('in', 'IN')

The RDF triples above hold intrinsic information regarding the way in which the sentence is structured. When we compare these to the dependency tree in Figure 5.3, we can see that the triples describe the relationships present in the graph.

In order to utilize this method effectively, the words are removed leaving only the POS tags so that the specialized vocabulary that is unique to many utterances is not taken into account.

A limitation of dependency parsing is that it only examines segments or whole sentences, therefore utterances longer than one sentence are not examined, leaving only 108 utterances that were agreed upon by the four annotators with regard to leadership style.

The ten most common triples using POS tags in the dataset are clustered with cosine similarity, as shown in Figure 5.4. Unfortunately, the three clusters produced by 21 common triples were completely random in terms of leadership style.

On the other hand, when clustering was performed on the most common triples that appear five times throughout the data (see Figure 5.4, two of the clusters produced by the 56 common triples did contain some patterns: cluster 0 contained only directing and coaching utterances while cluster 1 contained almost all supporting and delegating. This indicates that there were some patterns of task behavior that were being detected. In order to examine this further, the set was clustered with 9 clusters instead. Some of the patterns detected are below:

1. *VB-nsubj-PRP*: mostly coaching; indicates a verb performed by a pronoun which acts as the utterance subject (e.g., "we prepare" in "Can we prepare to change compressors, please?"), cluster 1;

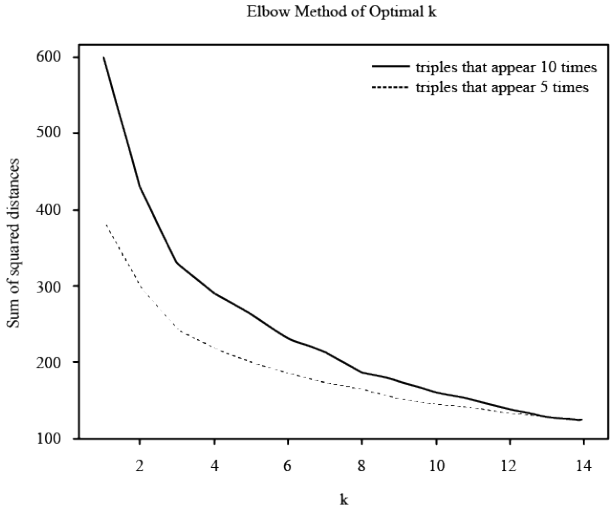


Figure 5.4 – The optimal k values when utterances 1 sentence or shorter from the agreement group are analyzed with cosine similarity based on common triples that appear at least 10 times and 5 times, resulting in 21 triples (at k=3, the SSD=290.4179) and 56 triples (at k=3, the SSD=218.8032, and at k=9, the SSD=146.5075) respectively. *Hold your hands in position.*

2. *VB-aux-MD*: coaching; indicates a modal verb used with a regular verb (e.g., "can prepare" in "if we can prepare for the administration of 300 milligrams of amiodarone IV push, please."), cluster 3;
3. *NN-case-IN*: delegating, indicates a noun that acts as a dependent of a preposition or subordinating conjunction (e.g., "like pressure" in "Looks like his blood pressure is quite high."), cluster 8.

Clustering based on the presence of common triples has confirmed some of the patterns previously discovered with clustering on raw utterances and POS tags only. There were some patterns picked up by clustering that were removed from the list above due to procedure-specific vocabulary. Pattern 1 from the list above indicates that pronouns are often found in coaching utterances; pattern 2 confirms that coaching utterances often contain modal verbs like "can"; and pattern 3 confirms that structures such as "looks like" and "sounds like" followed by a noun are often found in delegating utterances.

A list of rules determined by analysis on RDF triples is shown in Table 5.11. The details regarding each of these triples is available in Appendix D. For more detailed information about what the patterns here indicate, please see Universal Dependencies [5].

5. `https://universaldependencies.org`

	Directing	**Coaching**	**Supporting**	**Delegating**
All Strings		`VB-nsubj-PRP`, `VB-aux-MD`		`NN-case-IN`
Imperatives	`VB-dobj-NN`, `VB-discourse-UH`	`MD-mark-IN`, `VB-ccomp-VB`		
Interrogatives		`VB-dobj-NN`		
Indicatives		`VB-mark-IN`	`VB-dobj-PRP`, `VBG-aux-VBP`, `JJ-advmod-RB`	`NN-cop-VBZ`, `VBN-nsubjpass-PRP`, `VBN-auxpass-VBZ`

Table 5.11 – A list of rules generated by clustering on utterances' RDF triples. These were compiled by examining the most common triples in each group that also made human sense.

5.3.5 Chunking

Another method of examining a string's structure is by chunking, which involves manually defining phrases within a set of strings. The NLTK library in Python provides a method of defining rules for phrase classification [6]. Chunking allows for rules to be set for phrases that may not be clearly supplied or understood from sequences of words or POS tags or from RDF triples.

As was done in sections 5.3.3 and 5.3.4, only the 127 agreed-upon strings in the dataset are examined.

Chunking rules in NLTK are set by a combination of POS tags in a defined order with regular expressions. A chunk may require a certain POS tag, one or more specific POS tags, the absence of a certain POS tag, or it may leave the presence of a specific POS tag optional. Once the chunk is written, the chunking function traverses the POS-tagged string to determine whether that chunk is present and will then label the string with those chunks.

After examining the strings in the dataset described so far in this paper with words, POS tagging, and dependency parsing, it became clear that some information was lost, especially for imperative strings. Many utterances involve imperatives in which there is no explicit subject and a verb begins the string or begins a verb phrase within the string. Additionally, phrases that begin with *if* and *while* were often present in segments or incomplete sentences (e.g., "Denise, if you can let me know when two minutes have passed, please" is an incomplete sentence, with a phrase beginning with *if* acting as

6. `https://www.nltk.org/howto/chunk.html`

an instruction). Therefore it was worth exploring whether these phrases were related to communicative intention and leadership style.

Sixteen chunking rules, available in Appendix E, were written in Python to detect five phrases: verb phrases, noun phrases, and phrases beginning with "when", "while", and "if". We began by defining key phrases that do not necessarily show up when examining sequences or RDF triples. Verb phrases and noun phrases were defined. A verb phrase is the part of a sentence that contains a verb and its dependents, such as the object of the verb (e.g., "to intubate him" in the sentence "I'm going to try to intubate him real quick"). A noun phrase is the part of the sentence is the noun and any of the words that describe it (e.g., "bed in ICU" in the sentence "We will arrange a bed in ICU").

Additionally, we defined phrases beginning with "when", "while", and "if". The descriptor *explain* belongs to coaching leadership (Hersey et al. 1988). Therefore we felt it important to define a phrase that often includes an explanation or further context in English, such as phrases beginning with "when" or "while" and followed by a subject-verb phrase (e.g., "so when he arrives, what I'd like you to do is put in a second IV"). Like "when" and "while" phrases, clauses beginning with "if" seemed to contain information regarding either further context or a polite order (e.g., "Dexter, if you could get my labs for me.").

The chunks resulting from our example utterance "Hold your hands in position" are displayed below.

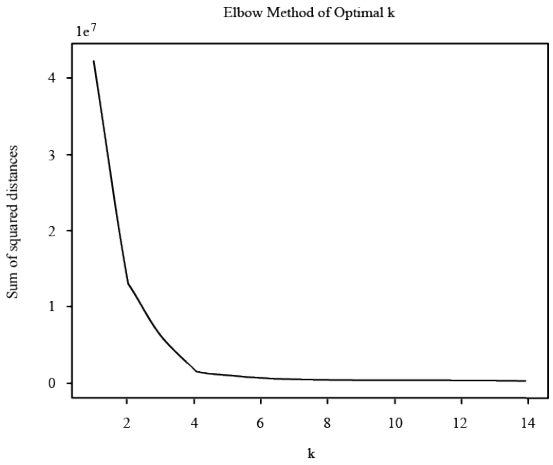
```
(S
  (VB-Phrase Hold/VB
    your/PRP$
    hands/NNS)
  in/IN
  position/NN
  ./.)
```

As shown in the chunks above, a verb phrase containing a verb, a possessive pronoun, and a plural noun is found in addition to a preposition ("in") and a singular noun ("position"). The fact that a verb phrase begins this utterance indicates that it contains the imperative mood.

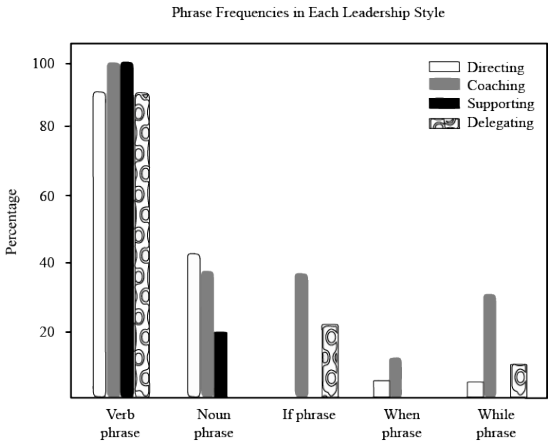
Once chunked, the utterances are clustered using both cosine similarity and edit distance. Unfortunately, after numerous attempts with varying sequence length, the clusters formed by cosine similarity appeared to be completely random in nature. However, clus-

tering using edit distance was more successful, the SSD values of which are shown in Figure 5.5a. Clusters 0 and 1 contained long coaching utterances that involve instructions; this was already confirmed by previous statistical methods. Clusters 2 and 3 were random.

Next, the different phrases present in each leadership style in the agreed-upon utterances from the dataset were analyzed, as shown in Figure 5.5b. As shown, verb phrases exist in most of the utterances regardless of leadership style. Noun phrases are less common across all leadership styles. Phrases beginning with "if", "when", and "while" are mostly limited to coaching and delegating leadership, which confirms that utterances with explanations and context are most commonly perceived as coaching or delegating.



(a) The optimal k number of clusters when utterances are chunked by phrase and are analyzed with edit distance. At k=4, the SSD=996910.8366.



(b) The frequencies at which each phrase appears in each leadership style.

Figure 5.5 – Results from k-means clustering on chunked sentences and frequency distributions of the phrases found.

Note that the chunk rules do not follow any patterns in terms of mood. A summary of the rules is displayed in Table 5.12. Verb phrases and noun phrases are so common that they are not included in the results.

5.3.6 Results of the annotation analysis

The results from all analysis in this section thus far provides rules for various structures within each leadership style. When compiled, there are a plethora of different options for different kinds of sentences, from vocabulary to structure.

Directing	Coaching	Supporting	Delegating
"when" phrases, "while" phrases,	"if" phrases, "when" phrases, "while" phrases		"if" phrases, "while" phrases

Table 5.12 – A list of rules generated by clustering on chunked agreed-upon strings. Each phrase denotes a phrase beginning with the word in quotation marks. There were no chunking patterns found for supporting leadership in particular.

Results from the annotation of the dataset provide a general framework of what to avoid and what to include. NLP analysis has provided more details. Many types of analysis resulted in the same information, which confirms the existence of those rules. Table 5.13 lists the rules that should inform leader speech in different leadership styles in the medical emergency room and which we propose for an agent leading a medical procedure.

Next to each rule in Table 5.13 are two figures: one, the number of total utterances in the agreed-upon portion of the dataset that include that rule and two, the percentage of utterances out of the total agreed-upon utterances that include that rule. For example, the first rule under *Mood* for directing utterances is *imperatives without "let"*. Out of the dataset utterances that were agreed upon as directing leadership, there were 19 utterances that had the imperative mood. There were 23 total utterances that all four annotators agreed belonged to directing leadership, and so $19/23 = 82.609\%$, meaning that 82.609% of the utterances labeled with directing leadership had imperative mood. Note that when the utterance contained more than one mood, the utterance is also counted more than once.

The results in Table 5.13 are demonstrative of the analysis that has been covered in this chapter so far. Because of the size of the dataset (375 total strings with 120 strings that were agreed-upon by all four annotators in terms of leadership style), some rules appear more often than others. Depending on the type of analysis, the entire dataset or the subset that was agreed upon was used. Future work could include creation and analysis of a larger dataset in order to ensure that various patterns appear more times throughout the entire dataset and throughout the subset that is agreed-upon by annotators.

5.3.7 Individual Annotator Analysis

Analyzing the agreed-upon strings is useful for finding characteristics of speech that might be universally recognized, but we also must account for differences between the annotators. Using latent Dirichlet allocation (LDA), we briefly explore each annotator's

	Directing (23)	**Coaching** (52)	**Supporting** (10)	**Delegating** (42)
Mood	Imperatives without "let" (19, 82.609%), Indicatives (4, 17.391%)	Interrogatives (37, 71.154%), Indicatives (25, 48.077%)	Indicatives (10, 100.0%)	Indicatives (42, 100.0%)
Length	whole sentences (18, 78.261%)	whole sentences (31, 59.615%), 2+ sentences (14, 26.923%)	whole sentences (8, 80.0%)	whole sentences (30, 71.429%), 2+ sentences (2, 4.762%)
Directness	Direct (22, 95.652%)	Direct (20, 38.462%), Indirect (42, 80.769%)	Direct (10, 100.0%)	Direct (42, 100.0%)
Speech acts	instruct (22, 95.652%)	instruct (47, 90.385%), inform (5, 9.615%)	support (5, 50.0%)	inform (41, 97.619%)
Keywords	"We need to (1, 4.348%), "Carry on with" (3, 13.043%)	"please" (28, 53.846%), "Okay, can someone" (2, 3.846%), "for me, please" (7, 13.462%), "as well, please" (12, 23.077%), "Please, can we" (9, 17.308%), "Can you please" (22, 42.308%), "You can" (3, 5.769%)	"Okay, thank you" (4, 40.0%)	"I see that" (3, 7.143%), "It looks like" (3, 7.143%)
POS tags		MD PRP VB (19, 36.538%), PRP MD VB (12, 23.077%)		VBZ IN PRP$ (1, 2.381%)
Phrases	"when" phrases (1, 4.348%), "while" phrases (1, 4.348%)	"if" phrases (11, 21.154%), "when" phrases (6, 11.538%), "while" phrases (7, 13.462%)		"if" phrases (2, 4.762%), "while" phrases (2, 4.762%)
Structure	`VB-dobj-NN` (17, 73.913%), `VB-discourse-UH` (4, 17.391%)	`VB-nsubj-PRP` (49, 94.231%), `VB-aux-MD` (39, 75.0%), `MD-mark-IN` (38, 73.077%), `VB-ccomp-VB` (34, 65.385%), `VB-dobj-NN` (47, 90.385%), `VB-mark-IN` (45, 86.538%)	`VB-dobj-PRP` (6, 60.0%), `VBG-aux-VBP` (1, 10.0%), `JJ-advmod-RB` (1, 10.0%)	`NN-case-IN` (31, 73.810%), `NN-cop-VBZ` (22, 52.381%), `VBN-nsubjpass-PRP` (8, 19.048%), `VBN-auxpass-VBZ` (7, 16.667%)

Table 5.13 – A list of rules generated by statistical analysis, performing clustering on the agreed-upon utterances and their POS tags, performing clustering on RDF triples, and analyzing the presence of specific phrases.

assignment of leadership style (Jelodar et al. 2019). LDA is a generative statistical model that allows unobserved groups to explain sets of observations, explaining why certain parts of the data are similar, in a method similar to clustering.

Sequences of words in the dataset did not yield meaningful results, so 3-grams of POS tags were used to find important and distinct groups. An initial assessment using LDA on the agreed-upon strings resulted in many of the same sequences that were produced by clustering. Only some of our results are discussed here.

The first annotator (referred to as Annotator 1 in this section) that we examine is female, age 27, with significant work experience. The most represented POS sequence for strings she labeled with directing and coaching leadership was VB DT NN (e.g., "check the pulse"). Strings containing this sequence were labeled with high-task behavior (directing or coaching leadership) by all annotators, indicating agreement on task behavior when that sequence is used.

Annotator 1 assigned directing leadership to sequence VB JJ PRP (e.g., "make sure you"). Strings containing the former were also labeled with high-task leadership (styles directing or coaching) by all annotators except for the male annotator aged 21 with less work experience, who labeled them as having delegating leadership.

Annotator 1 also assigned coaching leadership to strings with the sequence VB PRP VB, which entirely corresponded to "let's" + a verb. Other annotators assigned these strings styles directing, coaching, or supporting leadership, which confirms the lack of agreement when "let's" is used. If we were tailoring our virtual agent's speech to this annotator in particular, we would use the word "let" to begin utterances with high task and high relationship behavior.

The male annotator aged 21 with limited work experience, Annotator 2, seemed to disagree on leadership style with the other annotators the most. The most representative sequence of strings he assigned with supporting leadership was PRP VBP DT (e.g., "we have a" and "I am a"). There did not seem to be a pattern among the other annotators in terms of what leadership style they assigned when that sequence was present. This indicates that Annotator 2 identifies an introductory statement as well as the use of "we" as being an indicator of high relationship behavior, which is not true for the other annotators. If the speech was based on Annotator 2's assignments only, then utterances with that structure would be reserved for supporting leadership.

Findings such as these demonstrate how even further personalization of a leader's (and therefore agent's) communication might be necessary to correspond to an individual's own

definition of task and relationship behavior.

Through creation of a dataset of medical leader speech, annotation of that dataset, and analysis using multiple methods of the annotated dataset, we have been able to define several linguistic rules for medical leader speech in each leadership style. We have also demonstrated that speech does indeed differ between leadership styles in terms of word choice and structure. However, these linguistic rules have not yet been validated by other people. In the following section, we test the perception of speech created with these rules by conducting an experiment with participants.

5.4 Perception of medical leader speech

Despite the results of the annotation analysis covered in section 5.3, it was clear that there were a lot of things that annotators did not agree on. Furthermore, we did not learn how follower behavior was affected by the perception of a leader's task and relationship behavior. Therefore, we designed and conducted an experiment to both validate everything in Table 5.13 and explore how these rules influence follower behavior.

Our research questions include:

1. Does readiness level influence the perception of task and/or relationship behavior?
2. Does readiness level influence a follower's ability and/or willingness?
3. Is there any correlation between the perception of a leader's task and/or relationship behavior and the follower's ability and/or willingness?
4. Do various characteristics of a sentence influence a follower's ability and/or willingness?
5. Does a follower's performance with regard to ability and/or willingness improve when the follower is matched with the appropriate leadership style?

In this section, we explain relevant prior work before designing the experimentation, the methods used for the experimentation, and our findings.

5.4.1 Prior work

We wanted to explore how speech in text form created according to the rules in Table 5.13 is perceived and whether it influences followers' behavior. There has been no prior work on speech in each leadership style, and so works in other areas are examined.

Similar experiments evaluating the perception of different kinds of text might involve the attitudes of that text. One study invited participants to evaluate the charisma of speech in text format by answering 26 questions about the attitudes present in a number of sentences (Rosenberg and Hirschberg 2009). The survey was conducted online, and each question was answered on a Likert scale. The questions did not ask about charisma directly but instead asked participants to rate their agreement with statements like "The speaker is angry" which the authors used to evaluate the level of charisma present.

Another study was conducted to evaluate the attitudes present in emails that varied in politeness and directness (Hansen et al. 2015). In this study, participants did not rate the emails they read but instead wrote email responses to them, which led to an understanding of how characteristics of text influenced participant behavior.

A third study also varied email text based on politeness and directness and evaluated participants' perceptions of it (Economidou-Kogetsidis 2016). Participants answered ten Likert-scale questions about the content of the emails as well as the personality of the sender of the emails.

One last relevant work is one completed on the perception of a chatbot's personality (Ruane et al. 2020). The within-subject study invited participants to interact with a chatbot and then fill answer questions rating the chatbot's knowledge, quality of conversation, and attitude and personality.

5.4.2 Sentence creation

The first step of designing the experiment was choosing the sentences that participants would eventually evaluate. One sentence's content was chosen so that the content itself did not affect responses. We also wanted to choose something that participants generally would feel capable of doing so that their responses did not depend on their own capabilities. Therefore, we started with a situation in which the agent needed to motivate the caregiver to disinfect the patient's abdomen. The base sentence is a detailed imperative sentence: "Take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left."

Using the rules listed in Table 5.13, we created thirty-three sentences that do not appear in the original dataset covered in section 5.2. Because this thesis involves an agent leading a human being through a procedure by giving orders to both novices and experts, the communicative intentions are mostly limited to motivating the human to perform an action with speech acts *instruct* and *inform*.

Of the thirty-three sentences, thirteen used the guidelines for directing speech, twelve used the guidelines for coaching, two used the guidelines for supporting, and six used the guidelines for delegating. These sentences are described by various attributes:

— Leadership style: which leadership style's guidelines from Table 5.13 were used to create the sentence;
— Mood: whether the sentence was imperative, imperative-let (in which an imperative begins with the word "let's"), interrogative, or indicative;
— Keywords: any relevant keywords that are present in the sentence ("can", "could", "would", "please", "I need", "I want", "I'd like", "we", "help", "I see", and "it looks like");
— Detail level: the level of detail of the instruction (low, moderate, and high);
— Context given: whether an explanation for why the task must be done is present (either *yes* or *no*).

Some examples of the sentences we used, along with their leadership styles, include:

— I need you to prepare the patient by disinfecting the abdomen (directing);
— Could you prepare the patient by disinfecting the abdomen, please? (coaching);
— Do you need any help in preparing the patient? (supporting);
— The patient needs to be prepared before the procedure begins (delegating).

As evident in Table 5.13, there are many different rules that define speech in each leadership style. Some of them go hand-in-hand, such as the speech act *instruct* and the imperative mood, but some do not. It was not possible to test each and every one of these rules individually, so only the rules that lent themselves well to the base sentence we defined earlier were used. That means that some rules, like the rules regarding phrases, were not used in this experiment.

The complete list of these sentences is available in Appendix F.

5.4.3 Experiment Design

The experiment was conducted as an online survey using Google forms. Each participant was assigned a random readiness level among levels R1-R4 and asked to imagine that they are on a remote site with another human being who has suddenly fallen ill and a person who is their boss. Participants were told that the boss was also a human being so as to limit the effect that speech from a virtual agent would have and to demonstrate the effect of leadership speech outside of the context of this thesis. The participant and their boss must work together to save the patient. The boss is experienced in medicine and has

chosen a medical procedure to perform. In future work, we test whether the findings from this study hold up during a medical procedure led by a virtual agent.

The participants filled out some demographic information first: age, gender (male, female, nonbinary, or prefer not to answer), native language, and English level. Each participant then answered four questions for each of the thirty-three sentences to be evaluated. The questions were in a random order for each participant, and the sentences were randomized for each question to minimize any effects that sentence and question order might have. The four questions to be answered on a five-point Likert scale are listed below. In parentheses are the questions we were really trying to answer but did not show the participants.

1. Indicate to what extent (from strongly disagree to strongly agree) you agree that your boss is pushing you to do the job *(What is the participant's perception of the leader's relationship behavior?)*;
2. Indicate to what extent (from strongly disagree to strongly agree) you agree that your boss trusts you to do the job *(What is the participant's perception of the leader's task behavior?)*;
3. Indicate to what extent (from very incapable to very capable) you believe you are capable of completing this task *(What is the participant's perception of their own ability?)*;
4. Indicate to what extent (from very uncommitted to very committed) you are committed to completing this task *(What is the participant's perception of their own willingness?)*

5.4.4 Participants

Participants were recruited through the École Nationale d'Ingènieurs de Brest as well as social media and were each entered into a drawing for five 10-euro prizes.

Eighty-eight people total responded to the survey between October 13th and 26th, 2021. However, one participant responded to every question and every sentence with the same response, so their answers were removed, leaving eighty-seven participants.

Participants ranged from seventeen to sixty-three years old (mean = 32.41, SD = 13.30), forty-three of which were women, forty-four of which were men, and one of which preferred not to report their gender. Over 55% (forty-eight individuals) of the participants spoke English as a native language, with French (twenty-four participants), Arabic

(five participants), German (three participants), Dutch (two participants), Spanish (two participants), Italian (one participant), Polish (one participant), and Ukrainian (one participant) making up the rest. Fifty participants responded that they spoke English as a native language (the discrepancy between the number of participants who selected English as their native language and the number who reported that they speak English at a native level may be explained by participants who had sufficient English exposure and therefore determined their English level to be native as well). Regarding the English level of the rest of the participants, fourteen self-reported that they were fluent, twenty reported that they were high-conversational, and three reported that they were low-conversational.

Readiness levels were randomly assigned to participants: twenty-two were assigned to R1, twenty to R2, twenty-five to R3, and twenty to R4.

5.4.5 Analysis results

In this section, we explain what analysis methods we used to address each of the research questions listed at the beginning of section 5.4. We chose not to standardize participant responses. Only sixteen participants did not use the full range of responses: four participants' responses ranged from neutral to strongly agree, very capable, or very committed, and twelve participants' responses ranged from somewhat disagree, somewhat capable, or somewhat capable to strongly agree, very capable, or very committed. Given that there were only four participants who did not select any response indicating disagreement, incapability, or lack of commitment, we decided not to standardize the responses of every participant.

Influence of Readiness Level on Perception of Task and Relationship Behavior

To address research question 1, we isolated the data to include only responses to one question at a time. We fit a linear mixed effects model with *response* as the outcome variable, with a fixed factor of *sentence*, and a random factor of *readiness level*.

One-way ANOVAs revealed that there was not a statistically significant difference in response to Q1 of the survey (ANOVA, p-value = 0.31) or Q2 (ANOVA, p-value = 0.11) between readiness levels, indicating that readiness level had no significant effect on participants' perceptions of task and relationship behavior.

Regarding Q2, there was a statistically significant interaction (ANOVA, p-value $<$ 0.001) between readiness level and sentence as well, meaning that the extent to which the

participants felt that the boss trusted them to do the task depended on the combination of the sentence and the readiness level, but readiness level alone did not affect the extent to which participants felt that the boss trusted them.

There were statistically significant effects of gender (p-value = 0.01) and native language (p-value < 0.001) on the perception of relationship behavior. There were also statistically significant effects of gender (p-value < 0.001), age (p-value < 0.001), and native language (p-value < 0.001) on the perception of task behavior. These effects are explored further in subsequent sections, but due to the low number of participants in some categories, we cannot adequately explore these effects with our data.

Influence of Readiness Level on Ability and Willingness

To address research question 2, we performing the same linear mixed-effects model for Q3 and Q4. There is no statistically significant difference in response to Q3 between readiness levels (ANOVA, p-value = 0.06), indicating that readiness level had no significant effect on participants' perception of their own ability levels. However, there was a statistically significant interaction (p-value < 0.001) between readiness level and sentence.

There was also no statistically significant difference in response to Q4 between readiness levels (ANOVA, p-value = 0.891), meaning that readiness level had no significant effect on participants' willingness to complete the task.

There were, however, statistically significant effects of gender (p-value < 0.001), age (p-value < 0.001), native language (p-value < 0.001), and English level (p-value < 0.001) on the perception of relationship behavior. There were also statistically significant effects of gender (ANOVA, p-value < 0.001), age (ANOVA, p-value < 0.001), native language (ANOVA, p-value < 0.001), and English level (ANOVA, p-value = 0.01) on the perception of task behavior. Again, these effects are explored further in subsequent sections to find out how participants of different demographics respond differently.

Correlation Between Perception of Task Behavior and Ability

To address research question 3, we evaluate whether the responses to certain questions are correlated. A Kendall correlation revealed that there is a moderate positive correlation between the responses to Q2 and Q3 (Kendall τ-b = 0.21, p-value < 0.001) (Akoglu 2018). We can interpret this to mean that a participant's perception of task behavior is moderately correlated with their perception of their own ability to do the task. It is

possible that perception of the leader's task behavior can influence someone's perception of their own ability.

Among participants who were assigned readiness level R1, the correlation is slightly higher (Kendall τ-b = 0.21, p-value < 0.001), the correlation for R2 participants is higher again (Kendall τ-b = 0.31, p-value < 0.001), the correlation for R3 participants was moderate (Kendall τ-b = 0.25, p-value < 0.001), and the correlation for R4 participants was very weak (Kendall τ-b = 0.07, p-value = 0.04). This indicates that task behavior only affects followers' ability when they are in readiness levels R1, R2, or R3.

Correlation Between Perception of Relationship Behavior and Willingness

Again addressing research question 3, we investigate the relationship between the responses to Q1 and Q4 using Kendall's correlation once again. There was significant evidence to suggest that there was little association between responses to Q1 and Q4 (Kendall τ-b = 0.07, p-value < 0.001). This suggests that participants' perception of relationship behavior is only weakly correlated with their own willingness to do the task.

Among participants who were assigned readiness level R1, the correlation is insignificant and likely nonexistent (Kendall τ-b = 0.05, p-value = 0.10), the correlation for R2 participants is slightly higher (Kendall τ-b = 0.13, p-value < 0.001), the correlation for R3 participants was insignificant again (Kendall τ-b = 0.02, p-value = 0.45), and the correlation for R4 participants was very weak (Kendall τ-b = 0.09, p-value = 0.01). This indicates that relationship behavior only minimally affects followers' willingness.

Influence of Sentence Characteristics on Ability

To investigate research question 4, we isolated the data to include only responses to Q3 and to each readiness level in order to understand the variables affecting followers' ability in each level. Note that there were singularities between *Context given: Yes* and *Detail level: Moderate*, hence the *NA*s in Table 5.14.

R1 With the data limited to participants assigned readiness level R1, a simple multiple regression model was fitted with the dependent variable *Response* and the independent variables *leadership style*, *mood*, *keywords*, *detail level*, and *context given*. The majority of variables were found to be insignificant, as shown in Table 5.14. The best-performing model (adjusted $R^2 = 0.15$) contained the attributes *context given: no* and *mood: imper-*

	R1 Adjusted R^2: 0.14 $F(12, 713) = 11.08$, p-value < 0.001		**R2** Adjusted R^2: 0.33 $F(13, 646) = 25.71$, p-value < 0.001		**R3** Adjusted R^2: 0.05 $F(13, 811) = 4.25$, p-value < 0.001		**R4** Adjusted R^2: 0.01 $F(13, 646) = 1.64$, p-value 0.07	
	Coef.	p-val.	Coef.	p-val.	Coef.	p-val.	Coef.	p-val.
(Intercept)	1.60	<0.001*	1.49	<0.001*	2.15	<0.001*	2.22	<0.001*
Mood: Imperative-Let	0.35	0.33	0.61	0.06	-0.19	0.56	0.03	0.94
Mood: Indicative	-0.06	0.80	-0.11	0.62	-0.12	0.59	-0.10	0.71
Mood: Interrogative	0.35	0.33	0.32	0.34	-0.12	0.73	-0.17	0.69
Detail level: Moderate	0.32	0.07	0.96	<0.001*	0.55	0.00	0.38	0.06
Detail.level: High	1.27	<0.001*	1.90	<0.001*	0.63	<0.001*	0.087	0.67
Context given: No	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
Context given: Yes	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
Can	-0.57	0.10	-0.16	0.61	0.02	0.95	0.26	0.53
Please	0.10	0.49	0.12	0.35	0.08	0.52	0.07	0.67
Could	-0.60	0.08	-0.16	0.61	0.04	0.90	0.26	0.53
Would	-0.42	0.22	-0.20	0.53	0.14	0.66	0.21	0.61
I need	0.00	1.00	0.22	0.33	0.05	0.81	0.22	0.43
I'd like	-0.02	0.92	0.17	0.45	0.19	0.39	0.25	0.38
It looks like	-0.05	0.88	-0.03	0.92	-0.12	0.67	-0.42	0.23
Help	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
We	0.11	0.49	0.20	0.18	-0.01	0.95	0.01	0.96

** significant p-value*

Table 5.14 – Linear regression results when the Likert response to Q3 (evaluating participants' perceptions of their own ability) is the dependent variable. Data from each readiness level was examined separately. Singularities in the data led to *NA*s.

ative, with sentences with no context given increasing the base response by 1.18 (p-value < 0.001) and imperative sentences increasing the base response by 1.18 (p-value < 0.001).

When demographic information from participants was added to the model, the model's performance marginally increased (adjusted $R^2 = 0.30$). Interestingly, native English speakers rated their ability lower and German native speakers rated their ability higher. Because of the low numbers of participants with certain native languages, the differences of English perception between people from different of different native languages cannot be thoroughly reported on with our data but warrants further exploration.

Gender had the largest impact on participants' perceptions of their own ability. When the non-normal response data was analyzed with a Wilcoxon test, we found that men reported their perceived ability significantly higher than women did (p-value = 0.047), although this may not translate to actual performance, only perception.

R2 Like the R1 data, many of the variables were insignificant (see Table 5.14). The best-performing multiple linear regression model fitted with data limited to participants assigned readiness level 2 (adjusted $R^2 = 0.33$) included the attributes *Detail level: moderate*, which increased the participants' perceptions of their ability by 1.02 ($p < 0.001$),

and *Detail level: high*, which increased the participants' perceptions of their ability by 1.95 ($p < 0.001$). The model improved again when demographic information was added (adjusted $R^2 = 0.40$), but no one variable stood out as having a large effect on its own. Again, because of the small number of participants in our study, it is very difficult to point to a certain native language or age range that leads to significant results.

R3 and R4 The multiple linear regression models fitted with data limited to participants assigned readiness levels 3 and 4 did not perform well, and this is likely due to the fact that R3 and R4 followers already have high ability. The best-performing models had low values for R^2 (0.06 and 0.02 respectively). That said, ability did increase for R3 participants when the detail level was moderate or high, as shown in Table 5.14. The model for R4 participants was the only one that was insignificant itself.

Influence of Sentence Characteristics on Willingness

To continue to investigate research question 4, we perform the same steps we did in section 5.4.5 but analysing the responses to Q4 instead of Q3. Unfortunately the variables had far less impact on willingness than they did on ability. The best-performing models had adjusted R^2 values of less than 0.04. Demographic information failed to have a significant effect.

That said, there were some interesting findings regardless. For R3 participants, interrogative sentences beginning with "can" and "could" marginally increased participants' willingness to complete the task by 0.21 (p-value = 0.049) and 0.24 (p-value = 0.03) respectively. Also, indicative sentences including "it looks like" had a significant negative effect of -0.72 (p-value = 0.01).

Influence of Matching Leadership Style on Ability

To investigate research question 5, we limited the data to only responses to Q3 and fit a linear mixed-effects model again with a *match* variable which was *yes* if the participant's readiness level and the sentence's leadership style matched. We found that there was a significant difference between followers matched with the correct leadership style and those who were not (ANOVA, p-value < 0.05). Using a Wilcoxon test, followers matched with the correct leadership style perceived their ability to be significantly higher than those who were not matched with the correct leadership style (p-value < 0.05).

Influence of Matching Leadership Style on Willingness

Again to investigate research question 5, we performed the same steps that were explained in the paragraph above but including only responses to Q4. There was no significant difference in perceived willingness between followers who were matched with the correct leadership style and those who were not (p-value = 0.52). This mirrors our results from the paragraph above in which we found that there were few attributes that contributed to followers' willingness.

5.4.6 Summary of Analysis Results

The results from our participant evaluation on medical leader speech are listed in Table 5.15. As shown, we were able to validate and invalidate some of the findings from the analysis from the annotated dataset. One reason for the rules from the dataset analysis (Table 5.13) being invalidated is the number of people who annotated the dataset (four individuals) compared to the number of participants in the experiment (87 individuals). A second reason could be that some rules are based on patterns that appear fewer times throughout the dataset and others are based on patterns that appear more times. More annotators and/or a larger dataset of speech could mitigate this issue.

As shown in Table 5.15, there were no results regarding relationship behavior. Our evaluation also does not support the part of the SL® model that says supporting leadership should involve low task behavior.

	Directing	**Coaching**	**Supporting**	**Delegating**
Task behavior	high	high	high	low
Mood	Imperatives without "let", Indicatives, Interrogatives	Interrogatives, Indicatives	Indicatives, Interrogatives	Indicatives
Context given	no	no	yes	yes
Detail given	high	moderate-high	moderate-high	low-moderate
Keywords	"we need to", "I want you to", "carry on with", "I need", "can", "could", "would"	"please", "can", "could", "would"	"can", "could", "would", "please"	"I see that", "it looks like"

Table 5.15 – A list of guidelines for speech in each leadership style from our evaluation.

5.5 Conclusions

Speech was explored heavily in this chapter so that our agent can effectively use speech in each leadership style to increase the follower's ability and willingness and therefore ultimately result in a successful medical procedure.

In this chapter, we embarked on two projects to identify linguistic rules for speech in each leadership style. In section 5.1, we provided a foundation of existing work on linguistics. In sections 5.2 and 5.3, the creation of the dataset of medical leader speech, the annotation of that dataset, and the analysis of the annotated dataset are explained. Section 5.4 describes the experiment conducted to validate those linguistic rules. The experiment resulted in some rules being validated and some not. Table 5.16 displays the final set of guidelines regarding speech in each leadership style.

Note that all work in this chapter involves text only. This means that other modes of verbal communication, such as speech prosody and speed, are not examined. Speed and prosody were not able to be manipulated easily using our agent system. However, nonverbal behavior is something that the agent system can handle and is largely the subject of future work.

Speech has also been known to be perceived differently depending on the accompanying nonverbal behavior (J. Lee and S. Marsella 2006; Straßmann et al. 2016). Therefore the analysis and rules covered in this chapter will have to be reexamined when nonverbal behavior is added. As mentioned in section 4.5, the combination of speech and nonverbal behavior will have to be evaluated in a carefully-designed experiment in order to measure the effects of each.

Our work on agent speech is not limited to human-agent interaction in this thesis. Because no work to our knowledge has been published regarding speech differences between leadership styles, this work is applicable to human-human interaction as well. The ultimate goal of this work is to increase followers' ability and willingness in each readiness level with agent speech.

	Directing	Coaching	Supporting	Delegating
Task behavior	high	high	high	low
Context given	no	no	yes	yes
Detail given	high	moderate-high	moderate-high	low-moderate
Mood	Imperatives without "let", Indicatives	Interrogatives, Indicatives	Indicatives	Indicatives
Length	whole sentences	whole sentences, 2+ sentences	whole sentences	whole sentences, 2+ sentences
Directness	Direct	Direct, Indirect	Direct	Direct
Speech acts	instruct	instruct, inform, support	support	inform
Keywords	"we need to", "I want you to", "carry on with"	"please", "okay, can someone", "for me, please", "as well, please", "please, can we", "can you please", "you can"	"okay, thank you"	"I see that"
POS tags		MD PRP VB, PRP MD VB		VBZ IN PRP$
Phrases	"when" phrases, "while" phrases,	"if" phrases, "when" phrases, "while" phrases		"if" phrases, "while" phrases
Structure	`VB-dobj-NN`, `VB-discourse-UH`	`VB-nsubj-PRP`, `VB-aux-MD`, `MD-mark-IN`, `VB-ccomp-VB`, `VB-dobj-NN`, `VB-mark-IN`	`VB-dobj-PRP`, `VBG-aux-VBP`, `JJ-advmod-RB`	`NN-case-IN`, `NN-cop-VBZ`, `VBN-nsubjpass-PRP`, `VBN-auxpass-VBZ`

Table 5.16 – A list of final rules for speech in each leadership style compiled using annotation and experiment analysis.

Key points from Chapter 5

— A dataset of medical leader speech compiled from medical simulations was created;
— Four annotators assigned leadership style to each utterance in the dataset;
— Analysis was performed on the dataset including statistical analysis, k-means clustering, dependency parsing, chunking, and latent Dirichlet allocation to successfully identify elements of speech unique to each leadership style;
— An evaluation was designed and conducted that validated some of the linguistic rules resulting from the annotation analysis.

Part III

Overall System

Chapter 6

Computational Situational Leadership® for a General Human-Agent Tutor Interaction

Contents

In part III of this thesis, we describe the overall agent system that implements Situational Leadership (SL®). In this first chapter of part III, we describe a general model for human-agent interaction using SL®.

As discussed in chapters 1 and 2, SL® is the leadership model chosen for our agent system. Recall from section 2.1.1 that SL® specifies four readiness levels of varying ability and willingness and four corresponding leadership styles involving varying levels task and relationship behavior. SL® allows for different kinds of followers to be led successfully through any kind of task (Hersey et al. 1988). In order for an agent system to use SL® to lead a human through a task, (1) the system must be able to determine the correct readiness level of the caregiver in real-time, (2) the system must be able to determine the most appropriate leadership style that the agent should perform at each moment, and (3), the agent must perform the chosen leadership style.

Because SL® describes readiness level as a combination of low to high ability and willingness, identifying a follower's levels of ability and willingness will lead to the understanding of their readiness level. In an agent model, the best way to represent a follower's ability and willingness is with a numeric value. Then, using other appropriate factors in conjunction with the follower's readiness level, the system should assign an appropriate leadership style.

In section 6.1, our proposed architecture is detailed and we describe our flexible model that is applicable not only to the medical scenario that drives this thesis but to any scenario in which an agent must lead a human. In section 6.2, we discuss how certain user behaviors are triggered in our model. In section 6.3, our proposed method of determining the readiness level for a follower during a procedure is explained. In section 6.4, we describe our process of determining leadership style. In section 6.5, we briefly discuss how communicative intentions are created in this agent framework. Finally, section 6.6 summarizes the contributions of the work contained in this chapter.

The implementation of everything discussed in this chapter is explained later in chapter 8.

Note that throughout this thesis, the terms "follower", "user", "human", and "caregiver" are synonymous. A follower specifically refers to the individual taking direction from a leader in SL®, a user describes a human interacting with an agent system, a human is the human in a human-agent interaction, and a caregiver is a follower specific to a medical procedure.

6.1 Proposed architecture

As first mentioned in section 1.4, the proposed agent should be pedagogical in nature so it can lead followers who may not have knowledge of the procedure. This means that the agent must have a knowledge base of the procedure, be able to use feedback and interact with the user so that it knows the information state of the user, and be able to adapt to the user's information state. The agent should also be flexible enough to work in a virtual reality and augmented reality settings: virtual reality could be useful both for testing purposes; augmented reality could be useful for co-location of the agent and the user in the same environment in order to show the user how to use their environment and perform certain tasks. The agent is an ECA in order to make communication between the user and the system more human-like. As an ECA, being able to show the users how to

complete certain tasks could be very useful.

Our proposed architecture is shown in Figure 6.1. This architecture is a flexible model that can be implemented in any agent model that requires an agent to guide a user through a series of tasks.

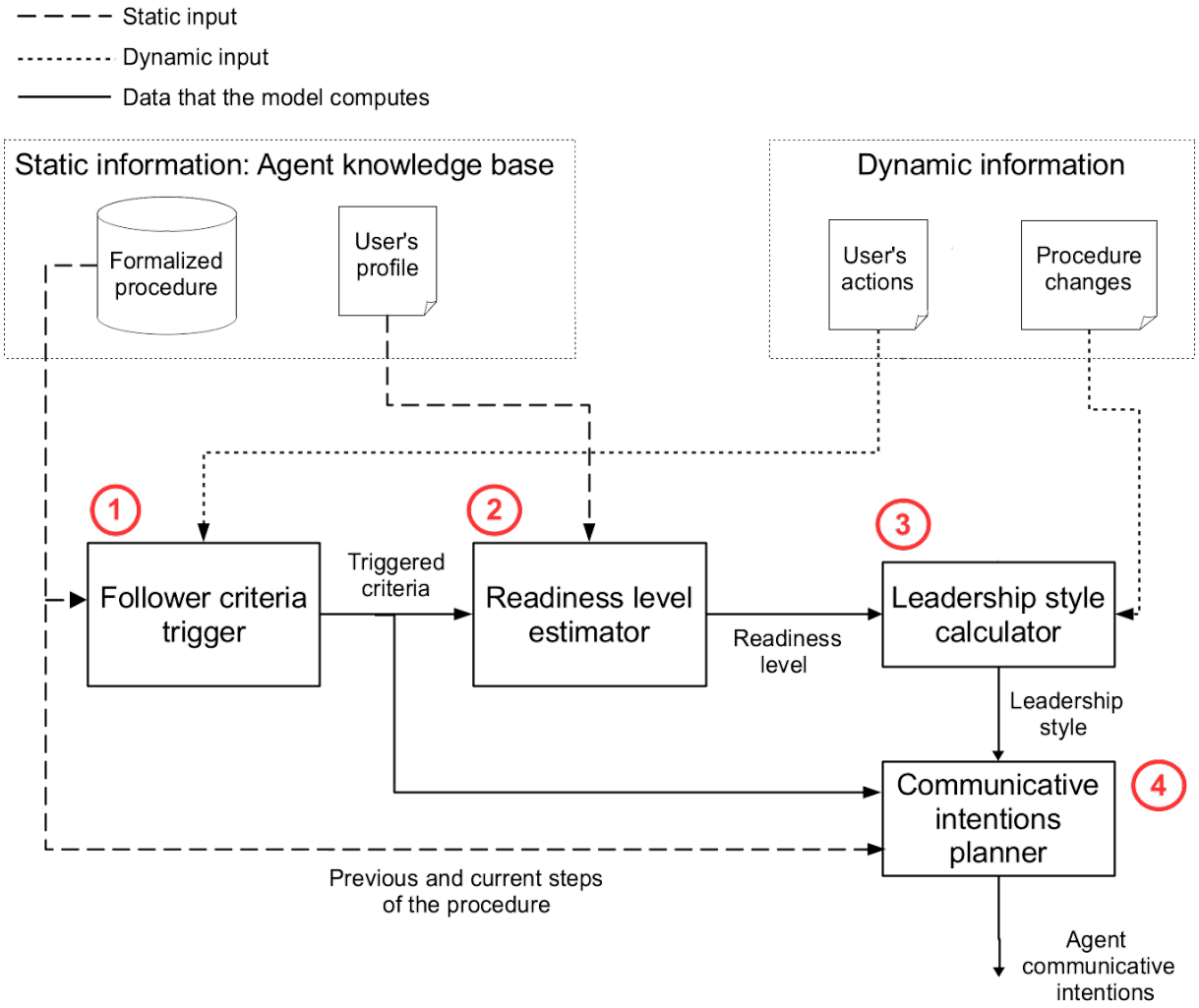


Figure 6.1 – The general architecture and flow of information within the agent framework.

In Figure 6.1, dashed lines refer to static input that is initialized at the beginning of the procedure (although the procedure steps are used throughout the interaction, its tasks do not change on their own), dotted lines refer to dynamic input that may change continually throughout the procedure, and solid lines refer to the data that is computed by the model.

To make everything work, both static and dynamic input is needed. Static input includes the procedure and all the data needed to initialize the the parameters needed by the system, which are discussed in more detail in sections 6.3 and 6.4. Dynamic input includes all the data received during the execution of the application, such as the user's actions.

In our architecture in Figure 6.1, there are four main modules that we propose:

1. Follower criteria trigger;
2. Readiness level estimator;
3. Leadership style calculator;
4. Communicative intentions planner.

When a user's behavior is an indicator of their ability or willingness, we refer to that behavior as a criterion, and so *follower criteria trigger* refers to user behaviors that trigger an indication of their ability or willingness. The *readiness level estimator* is where a user's readiness level according to SL® is computed. *Leadership style calculator* is where the most appropriate leadership style for the agent is determined. Finally, the *communicative intentions planner* determines the agent's communication in terms of both verbal and nonverbal signals. The last two modules can be seen as the intent planner of the SAIBA framework (see section 2.6.1). As we will see in section 8, the agent framework is SAIBA-compliant, so in the rest of this work, we can refer to the communicative intention planner and the leadership style calculator as the agent's intent planner.

The procedure is the series of tasks that the agent leads the user through. The procedure is formalized in the framework and contains detailed information about all the steps and any tools or other aspects of the user's environment that should be used to complete the procedure.

Throughout the rest of this chapter, these four modules along with the other elements in Figure 6.1 are discussed in detail.

6.2 Follower criteria trigger

Module 1 from Figure 6.1 is what we call the follower criteria trigger. As a reminder, the four readiness levels are (Hersey et al. 1988):

— R1: Low ability and low willingness;

— R2: Low to some ability and high willingness;

— R3: Some to high ability and variable willingness;

— R4: High ability and high willingness.

There are a number of different behaviors that a user might exhibit during a procedure. However, only the behaviors that are indicators of a follower's ability or willingness levels are referred to as criteria. The agent model is flexible enough to allow for any number of follower criteria to be added as needed by the type of procedure and/or the type of interaction.

Module one triggers follower criteria according to what the user is doing and what is expected according to the procedure which the user has to follow. When a criterion is triggered, a criteria event is created and sent to the intent planner. The triggered criteria remain in the system as criteria events until the event is resolved. For some criteria, they remain as criteria events until the user performs the correct action. For others, they remain criteria events until the agent communicates (this is explained in more detail in chapter 8). Thus criteria are considered dynamic input throughout the interaction. This implementation is discussed in further detail in chapter 8.

As mentioned in chapter 2, there is one existing work in which an algorithm for calculating readiness level is discussed (Bosse et al. 2017). The behaviors in the existing model by Bosse et al. are a combination of positive and negative behaviors; that is, some of them are behaviors that indicate ability or willingness, while some indicate inability or unwillingness. This brings us to the first difference between their model and ours: our model includes only behaviors that indicate inability or unwillingness. Because the agent should intervene when a criterion is triggered, only negative behaviors should count as criteria. We do not need the agent to do anything necessarily when things are going well.

Additionally, the behaviors included in the Bosse et al. model are categorized by readiness level, with each behavior indicative of only one readiness level. This means that followers are only able to trigger certain criteria when they are in certain readiness levels. This brings us to our second difference between their model and ours: all criteria in our model can be triggered no matter what readiness level the follower is in. Human users can be unpredictable: expert followers are capable of making mistakes and novice followers are capable of completing actions correctly (Andersson and Edberg 2010). Therefore, we group them by whether they are indicators of inability or unwillingness.

This model is quite flexible in that criteria can be added or removed easily as needed. Each criterion has four parameters, two of which are adapted from the existing Bosse et al. model and two of which are our own addition. These parameters of all the criteria

work together to calculate the user's *performance value* which is a value that describes the user's ability and willingness with respect to each criterion. This is explained in detail in section 6.3.1, but note that the term *performance value* is mentioned throughout this chapter. The four parameters for each criterion are displayed in Table 6.1 and are explained below:

	Parameter	**Value**
1	domain	string "ability" | "willingness"
2	error	integer [0,1]
3	persistence	float [0,1]
4	weight	float [0,1]

Table 6.1 – The four parameters for each criterion and their possible values.

1. *Domain*, a parameter that we added, refers to whether the criterion is an indicator of inability or unwillingness; criteria can either exist in the ability domain or the willingness domain, but not both;
2. *Error* refers to whether the follower exhibits the criterion or not and is described by either 0 or 1, where 0 indicates that the follower does not display that criterion, and 1 indicates that the follower does display that criterion. Thus, *error* behaves as a Boolean variable;
3. *Persistence* describes how persistent the follower's previous actions should be in the calculation of the follower's current performance value. In other words, persistence describes how much the presence or lack of a criterion during the previous action influences the current performance. It is described with a float on the interval $[0, 1]$. A value of 0 indicates low persistence while a value of 1 indicates high persistence. The lower the persistence value, the faster the follower's performance value will rise or fall when *error* for that criterion is 0 or 1 respectively;
4. *Weight*, the second parameter that we added, refers to the importance of the criterion within the ability and willingness domains when determining readiness level and is described by a float on the interval $[0, 1]$. All weights for each criterion used must add to 1. The criterion with the highest weight in one domain means that the criterion is more indicative of a follower's readiness than the rest of the criteria. The criterion with the lowest weight in one domain means that that criterion is

less indicative of a follower's readiness than the rest of the criteria. If all criteria have the same importance, then their weight values will be the same.

Only *domain* is a fixed parameter. *Error* is a value that changes over the course of the procedure as the follower progresses through the tasks. In the previous computational model from Bosse et al., a parameter called *extent* is a float on the interval [0,1] because followers can display the same behavior to different extents. This is another major difference between our model and theirs. In our model, we consider that a criterion is either triggered (represented by a value of 1) or not triggered (represented by a value of 0), and therefore we have created a different parameter *error* for which the values must be 0 or 1.

Criteria and their parameters are initialized before the procedure begins by using the user's follower profile, and thus they behave as static input. The follower profile describes the follower's previous experience and knowledge and contains which readiness level the individual has prior to beginning the procedure. We envision that the follower profile is based on previous procedures, external evaluations of the follower, and self-evaluations. The follower profile is discussed in more depth in section 9.3.

6.3 Readiness level estimator

Once criteria have been triggered, readiness level can be determined. Readiness level is then computed using the triggered criteria in module 2 from Figure 6.1. Before the procedure begins, the user's follower profile informs the initialization of the criteria, which in turn determine the follower's readiness level before the procedure has begun. Thus the initialized criteria are considered static input. During the procedure, criteria that are triggered in real time are used to compute readiness level in real time as well, meaning that criteria during the procedure act as dynamic input.

Figure 6.2 displays the concept that we want to achieve. In Figure 6.2, two lines represent a follower's ability and willingness over the course of a procedure. A red line represents the demarcation, a threshold, between low and high ability and willingness. Using this demarcation line, we are able to determine when the follower's ability and willingness values are low or high. Thus, we are able to determine what readiness level the follower belongs to from their ability and willingness values (when the follower's ability and willingness are both below the threshold, they are in readiness level R1; when their ability and willingness are both above the threshold, they are in level R4; etc.).

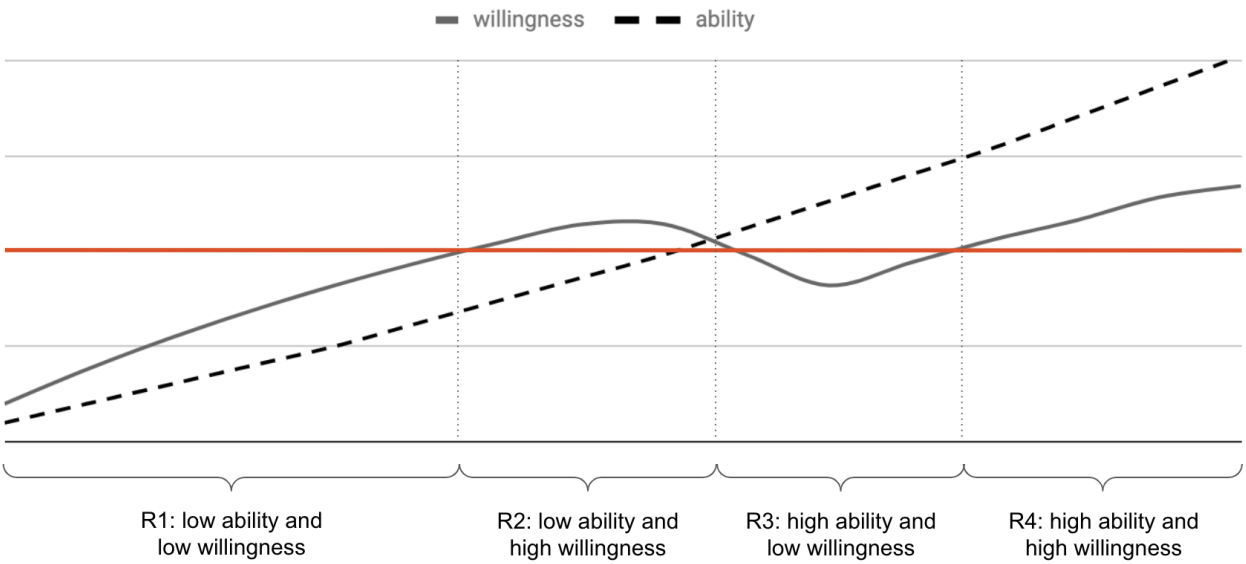


Figure 6.2 – Our wish for the results of a computational model of SL®: two values which represent a follower's performance in terms of ability and willingness, a demarcation between low and high ability and willingness performance, and assignment of readiness level.

Readiness level is defined as "the extent to which a follower has the ability and willingness to accomplish a task" (Hersey et al. 1988). Ability is "the knowledge, experience, and skill that an individual or group brings to a particular task or activity" while willingness is "the extent to which an individual has the confidence, commitment, and motivation to accomplish a specific task" (Hersey et al. 1988). Therefore, willingness can be affected by both general commitment to try to complete a task but also by the follower's perception of their own competence. Because a follower's readiness level can change depending on the nature of the task at hand, computation of the readiness level must occur in real time during a procedure.

In the rest of this section, the model for computing readiness level in real time is thoroughly explained. This model is flexible and can be applied to any scenario in which an agent leads a human user through a series of tasks.

6.3.1 The model explained

The model for calculating a follower's readiness level from criteria involves several equations and methods for ensuring the original definition of SL® is adhered to. In this section, the model is thoroughly detailed before simulations of how it works in practice

are explained in section 7.5. There are three main steps to determining the follower's readiness level, which are detailed throughout this section:

1. Calculating the user's performance value for each criterion;
2. Calculating the overall ability and willingness values;
3. Interpreting the overall ability and willingness values.

As mentioned at the beginning of this chapter, we envisioned that a user's ability and willingness should be represented by a numeric value. In order to satisfy this requirement and address item 1 from the list above, the user's performance of each criterion should also be represented by a numeric value, called the performance value.

Figure 6.3 – The expected progression of the performance for one criterion when the follower begins the procedure in readiness level R1 and the follower makes two errors over the course of the procedure.

Figure 6.3 displays a curve representing how the performance of the user with respect to one criterion should change over the course of a procedure. The follower has low ability at the beginning of the procedure. About halfway through the procedure, the follower makes an error that triggers a criterion, and later, they make the same error. As shown, we expect that both errors and correct actions affect the criterion's performance value.

Step one: calculating the follower's performance value for each criterion

To calculate a user's readiness level, we must calculate a value that describes the user's ability and willingness. However, before we can do this, we must calculate a value that describes the user's performance with regard to each criterion. We call this value a performance value. This value is very important as it will tell us whether the user is performing well or not with regard to each criterion.

The parameters explained in section 6.2 are used in two equations to calculate the performance value described on the interval $[0, 1]$ and represented with the variable v (v is borrowed from Bosse et al.), which defines how high the follower's performance is with respect to each criterion over the course of the procedure. However, our definition of the performance value changes slightly: a low performance value indicates low ability or willingness with respect to that criterion while a high performance value indicates high ability or willingness with respect to that criterion.

Equation 6.1 below is used to calculate the performance value v at time t for criterion c, where p_c represents the persistence for each criterion, and e represents the error, or presence or absence of that criterion. While this equation shares similarities with the equation found in the Bosse et al. model, it has been developed further and changed to suit our proposed agent system.

$$v_{c,t} = (p_c * v_{c,t-1}) + ((1 - p_c) * (1 - e_{c,t})) \tag{6.1}$$

The lack or presence of a criterion (the error e) is essentially a Boolean variable. If $e = 0$, meaning that the follower has not triggered the criterion, the new performance value v is equal to the persistence value multiplied by the most recent performance value, plus one minus the persistence value. The new performance value v will rise either quickly (if p is closer to 0) or slowly (if p is closer to 1).

If $e = 1$, meaning that follower has triggered the criterion, the new performance value v is equal to the persistence value multiplied by the most recent performance value. When p is closer to 0, the new performance value v will drop quickly, and when p is closer to 1, the new v value will drop slowly.

In this way, the new performance value v rises when the follower does not trigger a criterion and drops when the follower does trigger a criterion, resulting in a value that ultimately describes the follower's performance.

The performance value is calculated based on (1) a proportional relationship between

the persistence and the past performance value and (2) an inversely proportional relationship between the persistence and the presence of the error. The persistence value can be thought of as the "importance" variable here: it determines both how important the past performance value is and how important the current presence of an error is.

Equation 6.1 requires the performance value at time $t-1$. At the very beginning of the procedure, default performance values of 0 or 1 are used in place of $v_{c,t-1}$. These default values are established for each criterion which correspond to the follower profile. For a follower profile indicating low or variable ability or willingness, the default performance value is 0, and for a follower profile indicating high ability or willingness, the default performance value is 1. For example, if a follower profile indicates that the individual's readiness level is R3, then they have high ability and variable willingness. The default performance values for all the criteria in the ability domain are 1, and the default performance values for the criteria in the willingness domain are 0. These default values are initialized with the criteria before the procedure begins.

Note that the error values for each criterion remain the same until there is another opportunity to change that value, and performance values and overall values are only updated when error values change. This is demonstrated later in section 7.5.

Step two: calculating the overall ability and willingness values

The second step of computing readiness level is to average all the performance values for each criterion within each domain. We do this because readiness level consists of both ability and willingness. In order to find out what value represents a follower's ability and willingness, we must take into account the performance of the follower with respect to each ability and willingness criterion.

Equations 6.2 and 6.3 are used where C_a and C_w refer to the criteria in the ability and willingness domains, w_c refers to the weight that each criterion holds, and $v_{c,t}$ refers to the performance value at time t of each criterion as calculated by equation 6.1. These equations were developed as part of this research. Equation 6.2 results in a value representing the follower's overall ability A at time t, and equation 6.3 results in a value representing the follower's overall willingness W at time t.

$$A_t = \sum_{c \in C_a} w_c * v_{c,t} \tag{6.2}$$

$$W_t = \sum_{c \in C_w} w_c * v_{c,t} \tag{6.3}$$

Thus a criterion's weight w, along with each criterion's persistence p, acts as an "importance" variable in terms of working to calculate how important each error is when computing a user's readiness level.

Step three: interpreting the overall ability and willingness values

The third step of determining readiness level is to interpret the overall ability and willingness values calculated with equations 6.2 and 6.3. We do this by establishing thresholds that delineate between low and high A and W values.

There are two different thresholds, each for determining what readiness level the follower belongs to. In order to create the first threshold, $th1$, performance cutoffs for each criteria must be identified. This means identifying the highest performance value v that a follower can achieve for a criterion over the course of a procedure before being considered to have high ability or willingness with respect to that criterion only. However, these cutoff values do not represent a threshold on their own. A follower's individual performance for each criterion is only ever used in the context of all the other ability or willingness criteria. The cutoff is only used as an understanding of what the follower should achieve for that criterion, but it has to be taken with the context of all the other criteria as well, which is why the performance cutoff values are only used to create the threshold.

Ultimately, it does not matter if some performance v values are lower than their cutoff and some are higher than their cutoff - the only thing that matters is that the overall ability A and willingness W are above the thresholds. This is why the weight variable is also so important - sometimes, one criteria is not as important as the others.

For example, examine the case of an ability criterion that is not so important in a task. Its weight variable might be set quite low as a result of its lack of importance. If the follower triggers this criterion repeatedly, their v for that criterion may be very low and below its cutoff value. However, because the weight of this criterion is low, A might still be above the $th1_a$ despite the follower's low performance with regard to that single criterion. Thus the algorithm is designed to allow for a high-ability follower's poor performance with respect to a less important criterion. An example of this occurring is available in section 7.5.

Thresholds $th1_a$ and $th1_w$ are created by using each criterion's performance cutoffs

($cutoffV_c$) for $v_{c,t-1}$ in equations 6.2 and 6.3, as shown in equations 6.4 and 6.5. This process is completed for all criteria in both the ability and willingness domains, and the resulting values represent thresholds $th1_a$ and $th1_w$.

$$th1_a = \sum_{c \in C_a} w_c * cutoffV_c \tag{6.4}$$

$$th1_w = \sum_{c \in C_w} w_c * cutoffV_c \tag{6.5}$$

If a follower's overall ability value A at any one time is equal to or less than $th1_a$, they are considered to have low ability. If a follower's overall willingness value W at any one time is equal to or less than $th1_w$, they are considered to have low willingness.

The second threshold, $th2$, is created more pragmatically by choosing a value so high that a medical expert's performance (a follower in level R4) with respect to both ability and willingness would never cross. These thresholds are procedure specific, and working examples of how they are used is available in section 7.5.

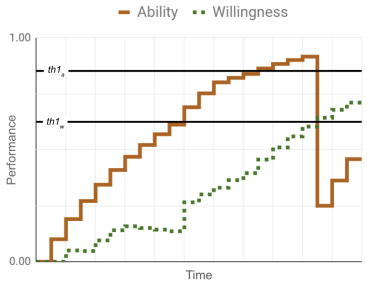
Adhering to Situational Leadership®

As mentioned at the beginning of this section, equations 6.1, 6.2, and 6.3 are not enough on their own in order to adhere to the rules set by the original work on SL®. Followers should not skip readiness levels (followers should not move directly between R1 and R3, R2 and R4, or R1 and R4) (Bosse et al. 2017; Hersey et al. 1988). To combat this issue, a method of artificially lowering or raising v values is implemented as needed, inspired by dynamic range compression (Kates 2005). This allows for one domain's overall performance to "wait" while the other domain's performance catches up.

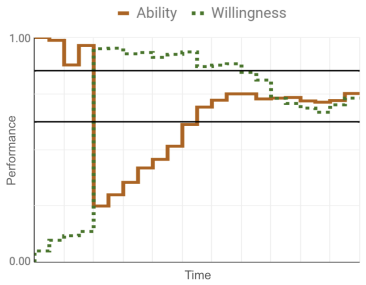
There are several of instances in which the ability or willingness value would need to be artificially changed, some of which are shown in Figure 6.4.

When a follower's ability or willingness values are changed in this manner, each criterion's performance value v is changed. The values that the performance values v become are procedure-specific and are therefore explored in section 7.5.

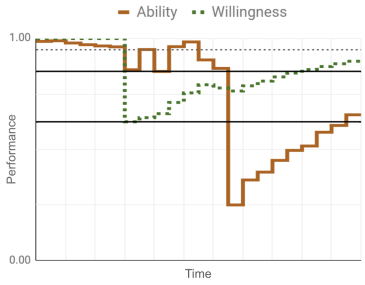
By lowering the ability and willingness values as needed, the system acts similarly to dynamic range compression in which signals are limited once they reach a certain threshold (Kates 2005). Unlike compression, however, the values can be lowered in order to allow the follower to progress through to the next readiness level. This is how we ensure that followers progress from one readiness level to the next without skipping levels.



(a) An R1 follower whose A rises faster than their W will need their ability criteria's v values lowered when Ow crosses the willingness threshold $th1_w$ to ensure they move from R1 to R2.



(b) An R3 follower whose A drops below $th1_a$ but whose W is still low will need their willingness criteria's v values raised to ensure they move from R3 to R2.



(c) An R4 follower whose A or W drops below $th2$ will need the other domain's value lowered to either $th1_a$ or $th1_w$ in order to ensure that the follower moves from R4 to R3 and from R3 to R2 smoothly.

Figure 6.4 – Three instances in which criteria performance values in the ability or willingness domain must be modified.

Now that the model for computing a user's readiness level and the method of ensuring our agent model adheres to the original model of SL® have been thoroughly explained, we move onto the model for computing the most appropriate leadership style for the agent.

6.4 Leadership style calculator

The leadership style calculator is shown as module three in Figure 6.1. This module belongs in the intent planner, which is discussed further in section 8. As discussed in Section 2.1.1, the original work on SL® specifies that generally readiness level and leadership style are linked - that is to say that if readiness level is R1, then leadership style should be style S1 (directing), etc. (Hersey et al. 1988). However, followers are only one factor that should affect leadership style. Other factors include the job demands and decision time (Hersey et al. 1988).

In this section, the stimuli that determine the leadership style of the agent are outlined and the model is thoroughly explained. Later, section 7.5 demonstrates how leadership style changes throughout the procedure with several simulations.

6.4.1 Stimuli

The inputs, or stimuli, that determine leadership style are all things that present themselves during a procedure, regardless of the type of procedure, that are relevant when determining leadership style. Therefore, the stimuli chosen for determining leadership style consist of:

— Readiness level;

— Procedure changes.

A user's *readiness level* has already been discussed and is calculated by equations 6.1, 6.2, and 6.3 as specified in section 6.3.

Procedure changes refers to the number of orders coming from the team of human experts standing by and monitoring the procedure. In our agent model, it is possible (although optional) for people outside of the human-agent interaction to monitor the situation and intervene if necessary. Orders that come from these individuals modify the formalized procedure. These changes may be the addition and/or the deletion of one or more actions. Any additions may be actions that the follower is not familiar with, even if the follower is familiar with the formalized procedure. Therefore, in our agent model, we assume that the user needs more direction when there has been a procedure change.

In the following section, the process of using user readiness level and the procedure changes to determine the most appropriate leadership style are detailed.

6.4.2 Determining leadership style

The pseudo code below demonstrates how procedure changes affect the agent's leadership style. Readiness levels R1-R4 and leadership styles S1-S4 are represented by the integers 1-4.

```
if procedure changes.Count > 0
    if action to do == first procedure change:
        if readiness level == 2 | 3 | 4:
            leadership style = 2
        else:
            leadership style = readiness level
```

As shown above, if the next action in the procedure belongs to a task that has been changed, the leadership style is automatically S1 or S2 (directing or coaching) depending

on the readiness level. This is to ensure that the agent is as directive as possible while still allowing room for follower autonomy when possible. Throughout the procedure, it is better for the agent to be more directive than not directive enough in order to prevent confusion and errors (Sims et al. 2009).

During a procedure, if there are no procedure changes made, then the agent's leadership style will always be equal to the user's readiness level.

In the following section, the last module in Figure 6.1 is briefly discussed.

6.5 Communicative intentions planner

The fourth module from Figure 6.1 determines the agent's communicative intentions. A communicative intention must accomplish two things: one, it must define the semantic units associated with a communicative event and two, it should allow the annotation of these units with properties or functions that further describe communicative function (Kopp et al. 2006).

The communicative intention is used to create a communicative action, which is the verbal signal for what the agent will say. This communication action, the sentence uttered by the agent, takes different forms based on the speech rules we have established in Table 5.15.

To accomplish agent speech that communicates the same intention but differs in style depending on leadership style, we identify a number of structures based on our speech rules covered in Table 5.16, and each characteristic is put into a list. Items are randomly pulled from these lists to create a sentence that the agent will speak. The goal is to increase followers' ability and willingness throughout the procedure.

These lists and the implementation of communicative intentions into the agent framework is discussed in more detail in section 8.2.

6.6 Conclusions

This chapter has explained how SL® can be modeled in a human-agent relationship when the agent assumes the role of a leader. The model is not specific to agents in the medical domain. Sections 6.1 and 6.2 provided the architecture we use in our agent model in order to result in an agent system that can use any number of user behaviors to calculate that user's readiness level, subsequently compute the agent's leadership style,

and ultimately output leadership style as a series of verbal and nonverbal signals, the implementation of which is discussed in detail in chapter 8.

Section 6.3 has been dedicated to explaining how caregiver readiness level is computed in real-time during the procedure. An algorithm with flexible parameters has been created in order to handle a variety of different tasks. This algorithm has been built to emulate SL®. Section 6.4 has explained how leadership style is chosen in real-time during the procedure based on follower readiness level and the number of procedure changes. Section 6.5 briefly introduces how communicative intentions are generated.

Key points from Chapter 6

— An agent model based on Situational Leadership® uses multiple equations that monitor user behavior and proposes the most appropriate leadership style;
— Four parameters allow for user behaviors to be evaluated in terms of a user's ability and willingness;
— Two stimuli (readiness level and procedure changes) are used to determine the most appropriate leadership style.

Chapter 7

Application of Computational Situational Leadership® to a Medical Procedure

Contents

In chapter 6, we detailed our computational model for an agent system that uses Situational Leadership® (SL®) to lead a human user through any kind of procedure. In this chapter, we detail how this model is applied to the scenario driving this thesis: a medical procedure. As such, this chapter acts as an addendum to chapter 6 when the context of the human-agent interaction is a medical procedure. Therefore, we focus on additions to the model discussed in chapter 6.

In section 7.1, we briefly introduce the specific medical procedure used in this thesis. In section 7.2, the criteria that are triggered by user behaviors are outlined. Section 7.3 explains an addition to the generic model that involves each criterion's persistence value. In section 7.4, we discuss the addition of the patient's state as a variable that influences the agent's leadership style. Section 7.5 provides several working examples of readiness level and leadership style progression using our computational model. Finally, section 7.6 contains final thoughts and key points from this chapter.

7.1 Medical procedure

The medical procedure that this thesis deals with is the diagnosis of abdominal pain and is formalized in the agent framework which is explained later in chapter 8. Each task in the procedure consists of one or more actions, which describe a singular step toward completing the task's goal. For example, in the task *Inquiring*, there is an action for *asking the patient for their level of pain*. The eight tasks and actions are available in Appendix G. Each action has a set of properties that include the resources required (if any), the expected duration of the action, and the role of the person who is supposed to complete the action (explained in detail in section 8.1).

In this scenario, there are four total roles: the agent, the human follower, the human patient, and the medical experts. The medical experts exist as a source of information to and from the agent in the form of further instructions and various environmental factors of the procedure (patient state, follower state, etc.) respectively. The patient's job is simply to provide information to the caregivers when necessary, and other than that, the patient exists as the subject of the medical procedure itself.

However, the human-agent interaction involves only two of those roles: the agent (the leader) and the user (the follower). The agent's job is to effectively communicate the procedure steps to the caregiver. The caregiver must cooperate with the agent, follow the procedure, and provide information when necessary.

Agent: Because the agent is expected to lead a caregiver who may or may not be familiar with the procedure or individual actions at hand, the agent also takes a pedagogical role as a tutor. The agent must be able to keep track of the current stage of the procedure and the information state of the user. When providing instructions to the user, the agent must be able to relay those instructions with varying degrees of task (directions) and relationship (socio-emotional) behavior (Hersey et al. 1988).

User: The user in this case is the human caregiver. Their role as the follower is to listen to what the agent tells them to do and follow its instructions. The user must be willing to communicate to the agent throughout the interaction and provide information to the agent when requested to do so. When the user is unsure of what to do, they should ask for help from the agent.

In the following section, we discuss the criteria chosen that indicate a follower's inability and unwillingness during a medical procedure.

7.2 Criteria chosen

As discussed in sections 6.2 and 6.3, user behaviors that are indicative of their inability and unwillingness during the specific procedure at hand are identified and referred to as criteria. In this section, we delve into the criteria chosen for a medical procedure. Their parameters are specific to each task rather than the procedure as a whole, and so the parameter values for each criterion are covered further in section 7.5.

Bosse et al. identified thirty-three behaviors that indicated a follower's performance with respect to their ability or willingness. However, this previous work focused on the working relationship between a graduate student and their supervisor over the course of months or years. These behaviors, while serving as a foundation for the behaviors identified in this thesis, are naturally quite different to the behaviors that a follower might perform during a medical procedure. Many of these behaviors were either irrelevant to a medical scenario (such as *feeling over-obligated* and *lacking self-esteem*) or were impossible to compute within our virtual environment without an activity monitor (such as *defensive behavior* and *discomfort in body language*).

Previous research has also analyzed user behavior such as facial expressions and head movements to better understand the user's information state (Dermouche and Pelachaud 2019). However, a human activity monitor is not utilized in VR-Mars. While these kinds of nonverbal behaviors are certainly useful, we chose to prioritize the caregiver's ability to move around the environment without worrying about whether their faces were reliably detected. Instead, we focus on human behavior that can be monitored and input from a keyboard.

Using the existing list of behaviors from Bosse et al. as a foundation, we created our own list of nine behavior criteria, shown in Table 7.1. The criteria each act as a demonstration of a lack of ability or willingness:

1. *Error: action in task*: The follower has chosen to do an action that exists in the task but is out of order;
2. *Error: action outside task*: The follower has chosen to do an action that does not exist in the current task;
3. *Wrong resource chosen*: The follower has taken the wrong resource for the action;
4. *(No) resource chosen*: The follower has neglected to take a resource when one is required *or* tries to take a resource when none is required;

	Criterion Name	Domain	Set of criteria
1	Error: action in task	ability	C_a
2	Error: action outside task	ability	C_a
3	Wrong resource chosen	ability	C_a
4	(No) resource chosen	ability	C_a
5	Action duration too short	ability	C_a
6	Action duration too long	ability	C_a
7	Question for help	ability	C_a
8	Hesitation	willingness	C_w
9	Question for reassurance	willingness	C_w

Table 7.1 – The nine criteria used in this thesis along with their values for *domain*. The values for *persistence* and *weight* are task-specific and *error* is follower-specific, and so these values are not fixed.

5. *Action duration too short*: The action duration is less than a certain threshold (for demonstration purposes, we have chosen the value of a duration < 0.9 * expected duration, although this value should be validated by experimentation);
6. *Action duration too long*: The action duration is more than a certain threshold (for demonstration purposes, we have chosen the value of a duration > 1.1 * expected duration, although this value should be validated by experimentation); ;
7. *Question for help*: The follower has asked a question because they do not know how to proceed;
8. *Hesitation*: The follower has hesitated for a period of time before beginning an action (for demonstration purposes, we have chosen the value of 5 seconds, although this value should be validated by experimentation);
9. *Question for reassurance*: The follower has asked a question to ensure what they are doing is correct.

These criteria have been chosen due to their relevance to the medical procedure described in section 7.1. However, more criteria could be gathered from follower behavior using the same four parameters depending on the needs of the procedure and the capabilities of the agent system, as explained in chapter 8.

As mentioned in section 6.2, the performance values v for each criterion are calculated only when the user has an opportunities to trigger a criterion. For example, the error value e of *Error: action outside task* only changes when the follower has the opportunity to make or not make another error (i.e., when it is time for the caregiver to begin a new

action). The performance values v are only calculated when e changes. The A and W values are updated when the individual v values change.

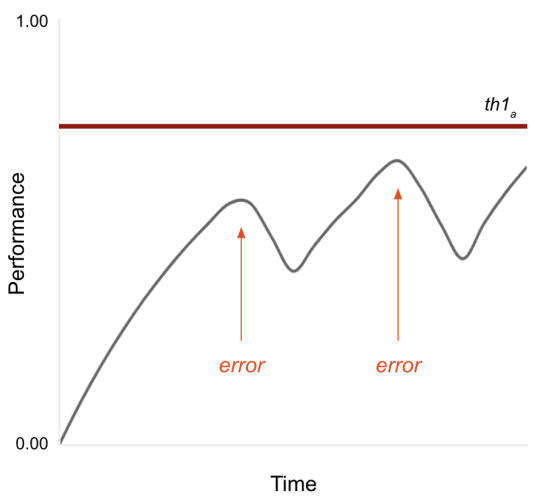
7.3 Criteria persistence values

Using the nine criteria and the model explained in chapter 6.3.1, the agent system is able to determine a follower's readiness level. There is one change that we make to the model's persistence value because of the specific application to a medical procedure: Each criterion in the ability domain has two different values for persistence: one for followers who have low ability according to the model calculation and one for those who have high ability according to the model calculation (persistence values for criteria in the willingness domain do not change).

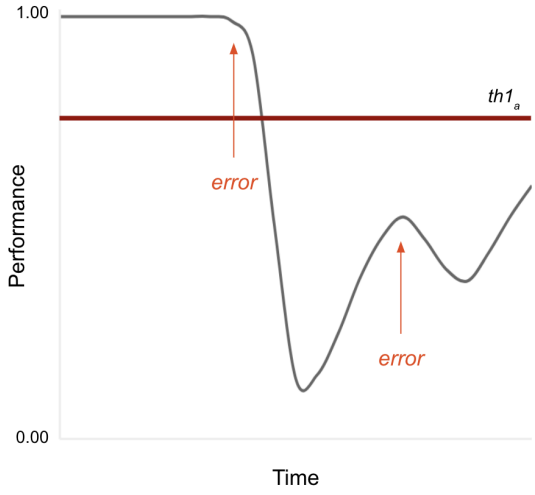
The method of modifying the ability criteria persistence values based on follower ability is done to differentiate between a novice who does everything correctly and an expert who makes one or two mistakes. During a medical procedure, even one small mistake can lead to serious consequences (Laposata 2014). When a follower displays high-ability behavior, the mistakes they make should have more impact on the performance value because when in doubt, the follower should be considered to have lower ability than they may have (Laposata 2014). When a follower triggers a criterion, that doubt about their ability is introduced. If the persistence value remains the same for both low- and high-ability followers, then the criteria trigger may not alert the system that the follower actually does not understand what they are doing. In other words, the gravity of the mistakes should depend on the follower's ability.

Figure 7.1 displays two curves representing our expectations of how the performance of two users with respect to one criterion would change over the course of a procedure. In Figure 7.1a, the follower has begun as a novice. In Figure 7.1b, the follower has begun as an expert. About halfway through the procedure, both followers make an error with respect to the criterion, and later, they make the same error. As shown, we expect that the errors should have a greater impact on the high-ability follower because once a high-ability follower makes an error, doubt about that follower's ability has been introduced. Note that the threshold $th1_a$ in Figures 7.1a and 7.1b represents the threshold for the overall ability A and not for this criterion in particular, but $th1_a$ is included in the figures to demonstrate when the persistence values would change.

When the overall ability value A drops below the ability threshold $th1_a$, then the



(a) The expected progression of the performance value v for one criterion when the follower begins with a default v value of 0 and the follower makes two errors over the course of the procedure. The red line represents the ability threshold $th1_a$.



(b) The expected progression of the performance value v for one criterion when the follower begins with a default v value of 1 and the follower makes two errors over the course of the procedure. The red line represents the ability threshold $th1_a$.

Figure 7.1 – Two graphs that represent the expectation of performance value progression due to varying persistence values when low- and high-ability followers make the same two errors during a procedure. The red line represents the ability threshold.

Persistence: low values are used instead. *Persistence: low* values are also used for followers who begin the procedure with low ability or willingness.

In the following section, we discuss patient state, a variable that is added to the generic model explained in chapter 6 that influences the agent's leadership style.

7.4 Patient state

In addition to follower behavior, there are other factors that should help determine the agent's leadership style. During an emergency medical procedure, there are instances in which the demands of the caregiver and reduced decision time should affect leadership style. In fact, previous research has demonstrated that there are many more cases in which an agent's leadership style should not directly correspond to a follower's profile (Sims et al. 2009). For example, if the patient's health suddenly takes a turn for the worse, it may be necessary for the agent to adopt a manner of communication that is more directive than the follower would ordinarily need in order to prevent possible errors but still allow for

user autonomy when appropriate (Goddard 2002; Hersey et al. 1988; Searle 1979). When the patient's state is very critical, leadership style should be as directive as possible (Sims et al. 2009).

For this reason, patient state is something that is added to the generic computational SL® model. The architecture shown in Figure 6.1 is shown again in Figure 7.2 with patient state added. As shown, patient state is considered dynamic input because it is monitored continuously throughout the procedure. This is discussed in more detail in chapter 8.

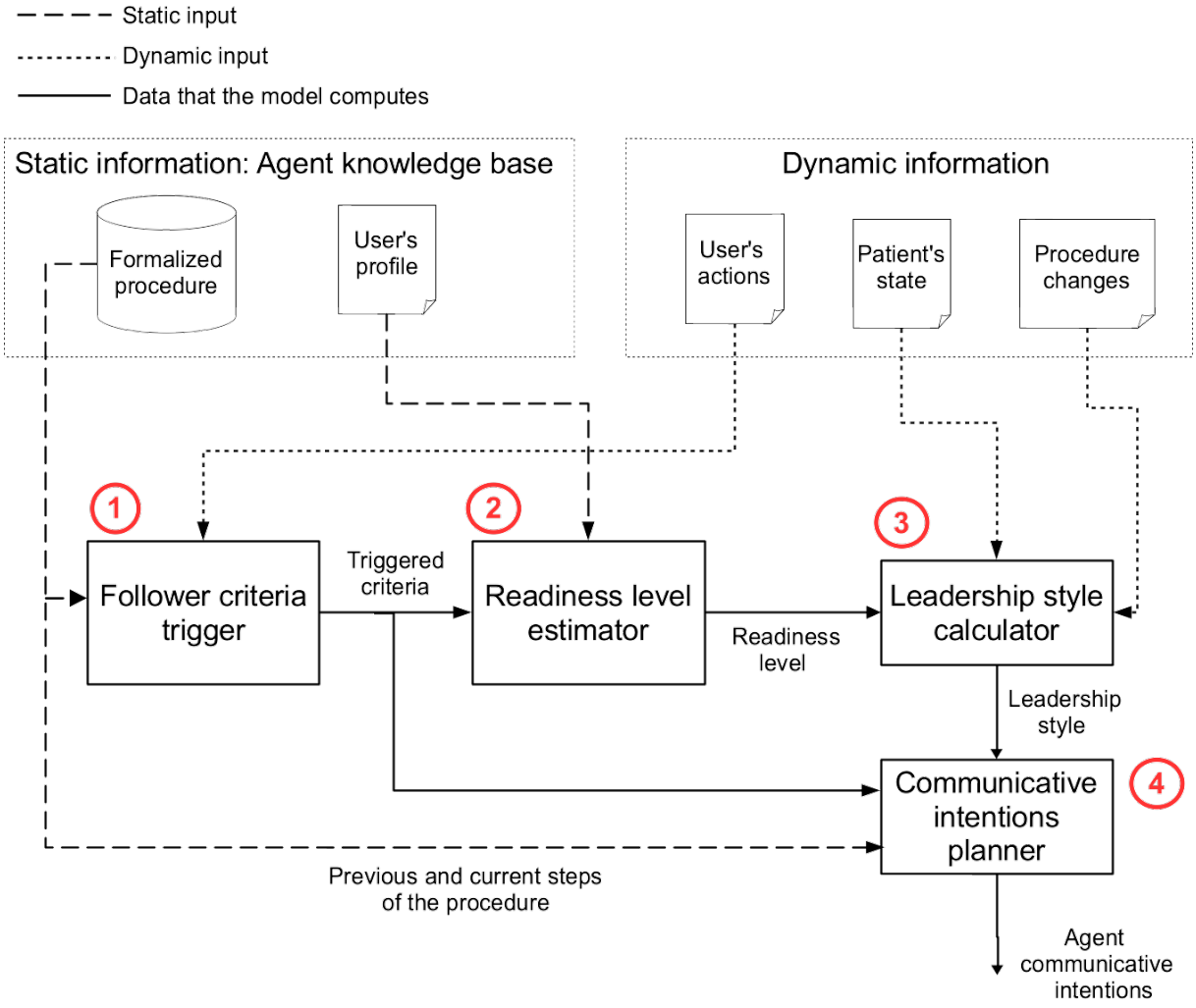


Figure 7.2 – The general architecture and flow of information within the agent framework.

Patient state is measured by the patient's heart rate, respiratory rate, blood pressure, and any other relevant measurements in comparison to the patient's normal and healthy

values. These values will have been developed with the help of medical doctors and will be patient-specific. Recall that this medical procedure will be performed during a remote expedition to Mars among a team of astronauts. These astronauts' medical information will be detailed in the system so that values other than their normal ones can be noted.

Patient state is denoted with a float in the interval $[0, 1]$, with values closer to 0 indicating low criticality, values closer to 1 indicating high criticality, and values in the middle indicating medium criticality. Thresholds between 0 and 1 are defined for each level of criticality: TCM for medium criticality and TCH for high criticality. Values for these thresholds should determined with the help of experts or through empirical studies. The specifics of how patient state is assigned during the procedure is outside the scope of this thesis. However, the patient state variable is still presented here to demonstrate how varying it affects leadership style.

If the patient's state is below TCM and/or there are procedure changes made, then the agent's leadership style does not directly correspond to readiness level.

Patient state at time $t - 1$ is set to 0 before the procedure begins. This value is also procedure-specific. Using just the patient state, several rules have been created to choose leadership style, described by the pseudo code below. Note that in the code, the numbers 1-4 refer to either leadership styles S1-S4 or readiness levels R1-R4.

```
if patient state >= TCH:
    leadership style = 1
else if patient state >= TCM:
    if readiness level > 1:
        leadership style = readiness level - 1
    else:
        leadership style = readiness level
else:
    leadership style = readiness level
```

According to the rules above, when patient state is above TCM, the agent's leadership style is always going to be more directive than the follower technically needs in order to combat any possible errors that they may make (Sims et al. 2009). When the patient state is above TCH, the leadership style is always directing. When the patient state is above TCM, the leadership style should be slightly more directive than the follower would ordinarily need. If the patient state is below TCM, then the situation does not require any

change in leadership style, and so leadership style directly corresponds to the matching readiness level of the caregiver.

Note that unlike readiness level, patient state is calculated on a continuous basis. Only after patient state has been taken into account can procedure changes are examined.

In the following section, we demonstrate the computational model by examining two example followers throughout the medical procedure.

7.5 Simulations

In this section, we show how followers' readiness levels and the agent's corresponding leadership style changes based on a variety of factors. This section is split into two parts: one for examining how followers' readiness level might progress and one for examining how the agent's leadership style might adapt over the course of the procedure.

7.5.1 Progressions of readiness level

In this section, we explain how each criterion's parameters have been chosen for the tasks in the medical procedure and demonstrate how followers' overall ability A and willingness W values and readiness levels react to various follower behaviors.

Dr. Eimear Wall, a medical doctor in the Health Service Executive in Ireland, was consulted when creating the values for each parameter and performance cutoff in the procedure used in this thesis. It was important to create the model with the guidance of a medical professional to ensure caregivers' correct readiness levels were returned during a variety of scenarios. While these values have not been validated by other health professionals or by experimentation, they have been validated by Dr. Wall. Then, even if these values cannot be considered scientifically obtained, we can use them as acceptable values to exhibit the principles of the model as well as the general tendencies of temporal evolution that it implies.

As a reminder, the full list of each task and action in the procedure is available in Appendix G.

Task 1: inquiring

The first task in the procedure involves the user asking the patient questions in order to help the agent diagnose the cause of the patient's pain and also answering questions

about the patient's behavior.

The criteria, parameter values, and performance value cutoffs for the inquiring task are listed in Table 7.2. As mentioned in chapter 6, the values for each parameter and each criterion will change depending on the specific task as certain things become more or less important depending on the individual actions and goals. Dr. Wall was consulted when these values were created in order to ensure that the correct readiness level was reflected for a caregiver who performs these criteria.

	Criterion Name	**Domain**	**High-ability persistence**	**Low-ability persistence**	**Weight**	**Performance cutoff**
1	Error: action in task	ability	0.7	0.85	0.05	0
2	Error: action outside task	ability	0.1	0.85	0.3	1
3	Wrong resource chosen	ability	*n/a*	*n/a*	0	*n/a*
4	No resource chosen	ability	0.1	0.85	0.28	1
5	Action duration too short	ability	0.7	0.85	0.05	0
6	Action duration too long	ability	0.7	0.85	0.05	1
7	Question for help	ability	0.4	0.85	0.27	1
8	Hesitation	willingness	0.9	0.9	0.5	0.5
9	Question for reassurance	willingness	0.9	0.9	0.5	0.75

Table 7.2 – The behavior criteria used for computing readiness level and each of their parameters for the task *inquiring*. The *persistence* values change depending on whether the follower has low or high ability. In this task, there are no resources to be used so error *Wrong resource used* is inapplicable.

For the *inquiring* task, the most important ability criteria is *Error: action outside task* as a follower performing a new action before the agent has received all the information needed within the *inquiring* task could be detrimental to the patient's diagnosis. This criterion is given the highest weight of 0.3. Criterion *No resource chosen* is also given comparably high weight of 0.28. Because this task does not contain any resources, a follower taking a resource indicates that the follower does not understand how to complete that action. Finally, criterion *Question for help* is also given a high weight of 0.27 because it indicates that the follower does not understand how to complete the action and is therefore another good indication of low ability. If the follower is considered to be high-ability, these the first two errors change their performance values considerably with low

persistence values of 0.1, and the last error is given a lower-than-average persistence of 0.4 since asking a question is not as serious the follower's history of asking questions for help can give more clues to their overall ability than the history does for *Error: action outside task* and *No resource chosen*.

Error: action in task, *Action duration too short*, and *Action duration too long* have the lowest weights. Errors out of order do not matter so much as long as the task is completed. Because of their low weight, short or long action duration times on their own do not generally affect the overall ability performance unless they happen repeatedly throughout the procedure, and this is done purposefully because at this stage, it does not matter so much if the caregiver and patient spend shorter or longer than normal discussing the patient's symptoms.

Wrong resource chosen has a weight of zero because in a task with no resources, it is not possible for a follower to select a wrong resource. The rest of this errors parameters are also inapplicable.

The persistence for low-ability followers is set to 0.85 for each ability criterion in order for the computation to better remember the follower's history. This is important as a low-ability follower is generally more likely to make mistakes, and so that history is important so that the agent does not communicate in a manner that assumes more ability than the follower has.

Finally, *Hesitation* and *Question for reassurance* are weighted equally in the willingness domain. Each of these are equally indicative of a follower's willingness. Additionally, the persistence for both is set quite high at 0.9 because a follower's willingness should be calculated from their history rather than from a single task.

Threshold creation Thresholds $th1_a$ and $th1_w$ are created using the performance cutoff values and weights from equations 6.4 and 6.5. The calculation is shown in Table 7.3. Threshold $th1_a$ 0.82, and the $th1_w$ is 0.6.

Threshold $th2$ is set to 0.95. This value was devised after consultations with Dr. Wall by examining example follower behavior and determining which were able to self-lead without the leader's help, and thus it serves as a pragmatic value. Followers with both an ability and a willingness value equal to or greater than 0.95 are considered to be in readiness level R4. However, this value should be validated through experimentation.

	Criterion Name	Domain	Equation 2 calculation	
1	Error: action in task	ability	$0.1 * 0.05$	$= 0.005$
2	Error: action outside task	ability	$0.9 * 0.3$	$= 0.27$
3	Wrong resource chosen	ability	n/a	
4	No resource chosen	ability	$0.9 * 0.28$	$= 0.252$
5	Action duration too short	ability	$0.1 * 0.05$	$= 0.005$
6	Action duration too long	ability	$0.9 * 0.05$	$= 0.045$
7	Question for help	ability	$0.9 * 0.27$	$= 0.243$
		ability threshold 1:		**0.82**
8	Hesitation	willingness	$0.5 * 0.5$	$= 0.25$
9	Question for reassurance	willingness	$0.7 * 0.5$	$= 0.35$
		willingness threshold 1:		**0.6**

Table 7.3 – The calculation of the thresholds for ability and willingness using equations 6.4 and 6.5.

Scenarios In order to demonstrate how followers' readiness levels might progress during the inquiring task, we examine two scenarios from different followers: Follower A and Follower B. Follower A begins the task in readiness level R1, and Follower B begins the task in readiness level R4, and therefore they start with the appropriate default performance values v of 0 or 1. Both followers trigger the criterion *Error: action in task* at every action in the inquiring task. Both followers' A and W values at the end of each action are presented in Table 7.4 and a visualization of their progression is found in Figure 7.3.

Note that in Figure 7.3, the x-axis displays the action number, and each segment represents the completion of the previous task. The errors may not occur immediately after these points because the followers are waiting for the agent to finish speaking the next order and/or because they are hesitating for less than five seconds before choosing the action they will do. Recall that v for each criterion and therefore A and W are updated whenever the follower triggers a criterion or has the opportunity to trigger a criterion, as explained in section 8.1.3.

Because the weight of the criterion *Error: action in task* is only 0.05, it does not have a large effect on A of either follower. As shown in Table 7.4 and Figure 7.3, Follower A reaches level R2 (because their W value has surpassed $th1_w$) at the end of action 9. Follower B remains in level R4 throughout the task.

Action	Error made	Follower A		Follower B	
		A	*W*	*A*	*W*
default v values		0	0	1	1
1	Error: action in task	0.1425	0.1	0.985	1
2	Error: action in task	0.2636	0.19	0.9745	1
3	Error: action in task	0.3666	0.271	0.9672	1
4	Error: action in task	0.4541	0.3439	0.9620	1
5	Error: action in task	0.5285	0.4095	0.9584	1
6	Error: action in task	0.5917	0.4686	0.9559	1
7	Error: action in task	0.6455	0.5217	0.9541	1
8	Error: action in task	0.6911	0.5695	0.9529	1
9	Error: action in task	0.7300	0.6126	0.9520	1

Table 7.4 – Two examples of readiness level progression during the inquiring task for Followers A and B. At each action, they each try to do an action that exists in the task but is out of order. The values in the table are the last A and W values after the action is completed.

Task 2: palpation

Following the questioning of the patient's pain during the inquiring task, the caregiver is then instructed to palpate the patient. Table 7.5 presents each criterion's parameters, which differ from those presented in Table 7.2. Again, Dr. Wall was consulted when creating these values to ensure a variety of follower behaviors reflected an accurate readiness level.

Because this task requires that the caregiver follow each action exactly, *Error: action in task*, *Error: action outside task*, and *Resource chosen* are all set at the same weight of 0.17. This task is more urgent in nature also, and a caregiver who completes an action in too little time or too much time may not have completed the action correctly, hence why *Action duration too short* and *Action duration too long* are both set to the same weight of 0.17. *Question for help* is also an important indicator of inability, which is why its weight is set to 0.15.

Because the follower must adhere to the actions in this task, the ability persistence values for low-ability followers are slightly lower than they were in the questions task. This ensures that any errors have a greater impact on the overall ability value and reflect the follower's true ability over the course of the task.

Threshold creation To compute $th1_a$ and $th1_w$, the performance cutoffs in Table 7.5 are used in equations 6.4 and 6.5, resulting in a $th1_a$ value of 0.932 and a $th1_w$ value of

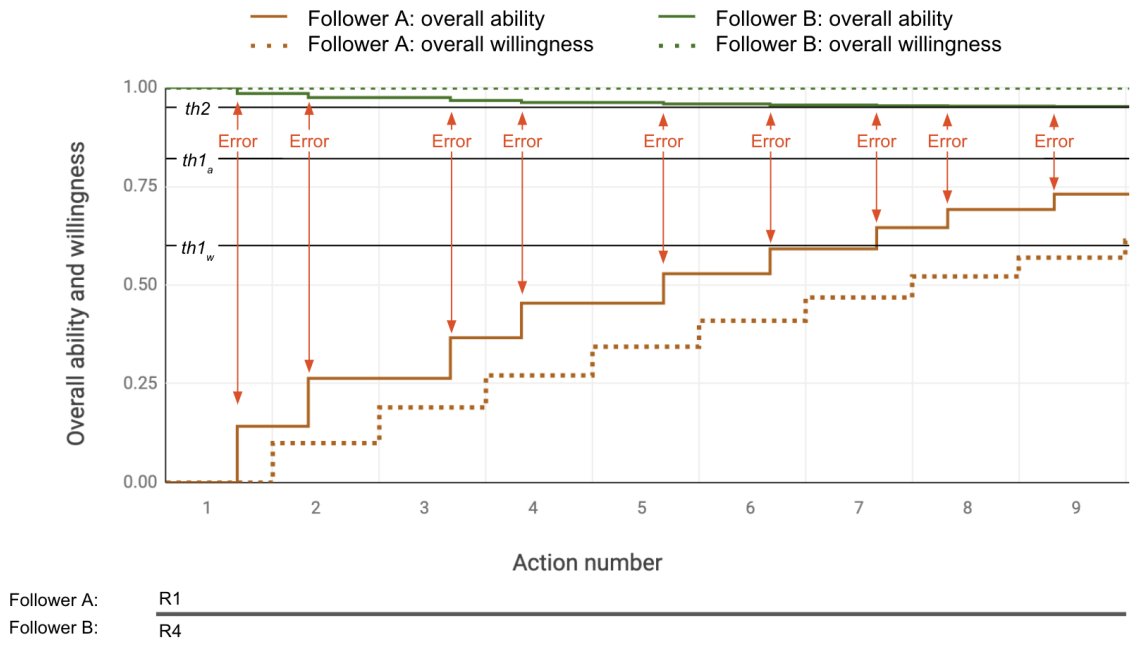


Figure 7.3 – Two examples of readiness level progression during the inquiring task for Followers A and B. At each action, they each try to do an action that exists in the task but is out of order.

	Criterion Name	**Domain**	**High-ability persistence**	**Low-ability persistence**	**Weight**	**Performance cutoff**
1	Error: action in task	ability	0.3	0.7	0.17	1
2	Error: action outside task	ability	0.3	0.7	0.17	1
3	Wrong resource chosen	ability	*n/a*	*n/a*	0	*n/a*
4	No resource chosen	ability	0.3	0.7	0.17	1
5	Action duration too short	ability	0.7	0.7	0.17	0.8
6	Action duration too long	ability	0.7	0.7	0.17	0.8
7	Question for help	ability	0.4	0.7	0.15	1
8	Hesitation	willingness	0.9	0.9	0.5	0.8
9	Question for reassurance	willingness	0.9	0.9	0.5	0.75

Table 7.5 – The behavior criteria used for computing readiness level and each of their parameters for the task *palpation*. In this task, there are no resources to be used so error *Wrong resource used* is irrelevant.

0.775.

Scenarios Followers A and B are examined again, but each follower begins this next task with the ending v values for each criterion from the inquiring task (because this is a continuation of the same procedure). In this task, each follower makes a variety of errors. A visualization of their progression is found in Figure 7.4.

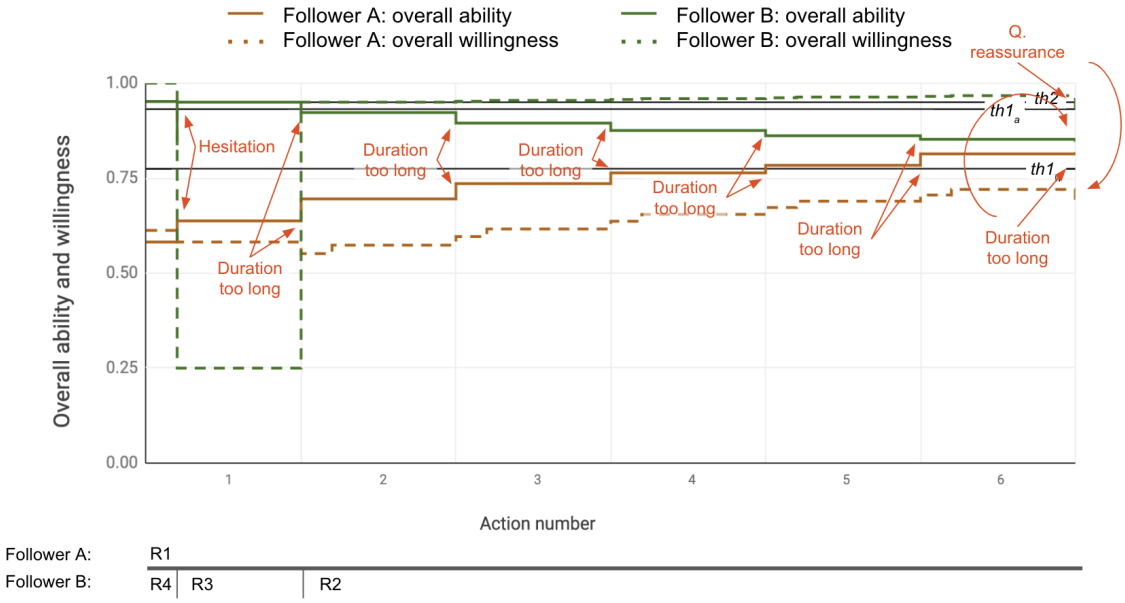


Figure 7.4 – Two An examples of readiness level progression during the palpating task for Followers A and B. They each perform a variety of errors.

When we compare Followers A and B, we see that A decreases when the follower begins in readiness level R4 yet increases when the follower begins in readiness level R1. Because only one criterion is being triggered, and that criterion has a weight of 0.17, both overall ability values will converge towards 0.83 (1-0.17 = 0.83).

As shown in Figure 7.4, Follower A's overall A and W values continue to rise, despite taking longer than expected for each action and asking two questions for reassurance after having completed an action. However, their A and W values never cross the first thresholds of 0.932 for ability and 0.775 for willingness, and so the follower remains in readiness level R1, as indicated by the bar at the bottom of Figure 7.4.

For Follower B, due to the varying persistence and weight values, the overall ability value immediately drops to 0.8428. Because this value is below both $th2$ and $th1_a$, the follower is technically in level R2 at this point. However, followers cannot skip levels, so

instead the follower must pass through level R3 first. Thus at the beginning of action 1, two things happen: the overall ability value is artificially raised to 0.95 and the overall willingness value is artificially lowered to 0.25. In this way, the overall ability value is above the first threshold, and the overall willingness value is below the first threshold, meaning that the follower is in readiness level R3.

The follower then takes longer than expected to complete action 1, and the overall ability value drops below the first threshold again. Even though the follower has asked a question for reassurance at the same time, the overall willingness value is then raised to ensure the follower passes through readiness level R2.

Tasks 3-8: measuring the patient's vital signs

Tasks 3-8 in the medical procedure all involve measuring the patient's vital signs, and so due to their similar nature, the same criteria parameters and performance cutoff values are used in tasks 3-8 (the rest of the procedure). Table 7.6 presents each criterion's parameters and performance cutoffs. Note that *Wrong resource chosen*'s parameters are applicable in tasks 3-8 because resources are required. The persistence values are exactly the same as they were for the palpation task.

	Criterion Name	**Domain**	**High-ability persistence**	**Low-ability persistence**	**Weight**	**Performance cutoff**
1	Error: action in task	ability	0.3	0.7	0.15	0.95
2	Error: action outside task	ability	0.3	0.7	0.15	0.95
3	Wrong resource chosen	ability	0.3	0.7	0.15	1
4	No resource chosen	ability	0.3	0.7	0.15	1
5	Action duration too short	ability	0.7	0.7	0.15	0.8
6	Action duration too long	ability	0.7	0.7	0.15	0.8
7	Question for help	ability	0.4	0.7	0.1	1
8	Hesitation	willingness	0.9	0.9	0.5	0.8
9	Question for reassurance	willingness	0.9	0.9	0.5	0.75

Table 7.6 – The behavior criteria used for computing readiness level and each of their parameters for tasks 3-8.

Threshold creation To compute $th1_a$ and $th1_w$, the performance cutoffs in Table 7.5 are used in equations 6.4 and 6.5, resulting in a $th1_a$ value of 0.9025 and a $th1_w$ value of 0.775. Threshold $th2$ remains at 0.95.

Scenarios Followers A and B are examined again, and each follower begins this next task with the ending v values for each criterion from the palpating task because they are part of the same procedure. In this task, each follower makes a variety of errors. Visualizations of their progression are found in Figures 7.5 and 7.6. The progression for Follower A and B is shown in separate graphs for ease of reading.

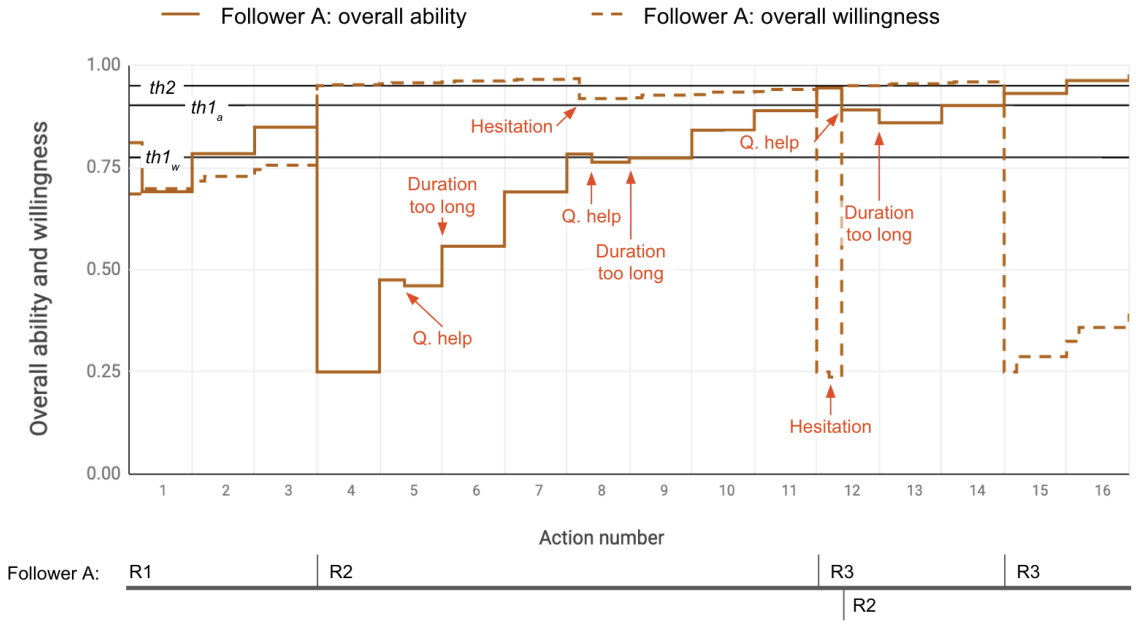


Figure 7.5 – Follower A's readiness level progression during tasks 3-8. They perform a variety of errors. The values in the table are the last A and W values after the action is completed.

As shown in Figures 7.5 and 7.6, the overall ability and willingness values change immediately after the procedure starts due to the changed persistence and weight values in the new task. Follower A's ability rises above $th1_a$ at the end of action 3. Because this would normally place the follower in readiness level R3 (having skipped level R2), the performance values for all ability criteria are lowered to 0.25 and those for all willingness criteria are raised to 0.95 to force the follower into level R2.

At the end of action 11, Follower A's overall A value again rises above $th1_a$, which

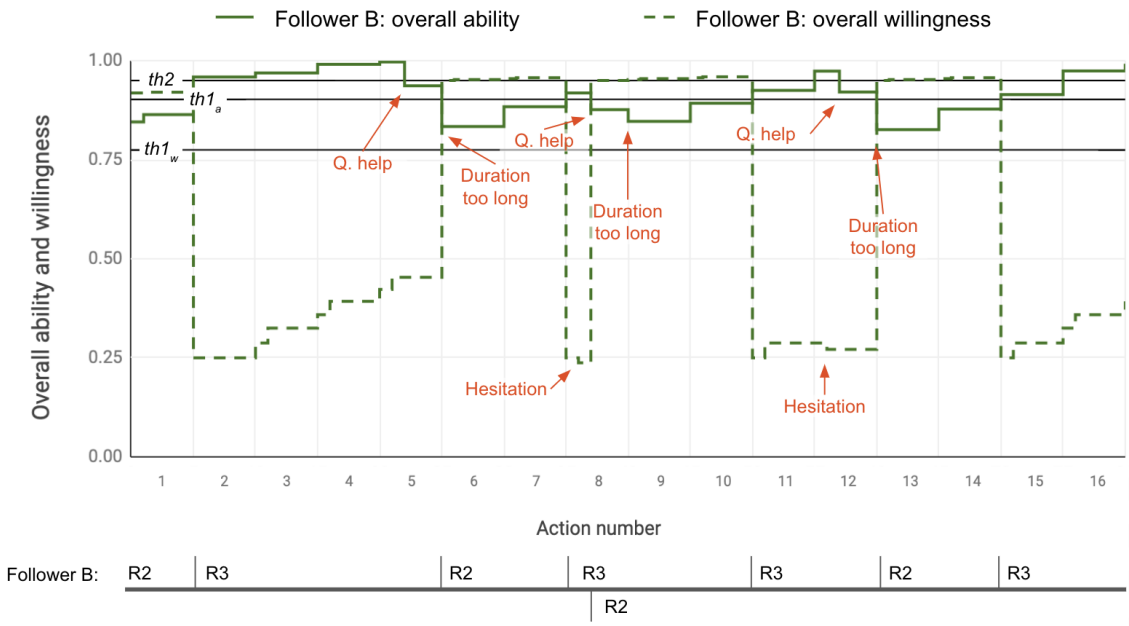


Figure 7.6 – Follower B's readiness level progression during tasks 3-8. They perform a variety of errors.

would normally place the follower in readiness level R4. In order to force the follower to first pass through R3, the willingness criteria performance values v are lowered to 0.25. At action 12, A drops below $th1_a$, and the follower moves back down to readiness level R2. However, at the end of action 14, Follower A's overall A value rises above $th1_a$ for the third time. In order to force the follower to pass through level R3, all v values for each willingness criterion is lowered to 0.25. The follower finishes the medical procedure in level R3.

In Figure 7.6, we see the progression of A and W for Follower B. At the end of action 2, the follower's overall A rises above $th1_a$, placing the follower in readiness level R4. In order to force the follower to pass through level R3 instead, the willingness criteria performance values v are lowered to 0.25.

At the end of action 5, A drops below $th1_a$. In order to force the follower to pass through readiness level R2, the willingness criteria performance values v are raised to 0.95. As shown in Figure 7.6, A crosses $th1_a$ several times, and so the willingness criteria performance values v are modified several times as well to facilitate the smooth transition of the follower through the readiness levels R2 and R3.

Similarly to the scenario for the R1 follower (Figure 7.5), this follower ends the proce-

dure in readiness level R3. In fact, in examining both Figures 7.5 and 7.6, we see a lot of similarities. Remember that these two followers have completed the procedure exactly the same, despite one starting in R1 and one starting in R4. It makes sense that eventually, their overall ability and willingness values would become nearly the same. If the procedure were longer, eventually these values would converge completely.

Now that we have demonstrated how the computational model works for two followers beginning in different readiness levels, we demonstrate how leadership style can change as well.

7.5.2 Progressions of leadership style

In order to demonstrate how leadership style reacts to various events during a medical procedure, we examine the scenario of Follower B from the previous section. Figure 7.7 displays the follower's readiness level progression and the agent's corresponding leadership style given that the patient's state remains below TCM and there are no procedure changes sent.

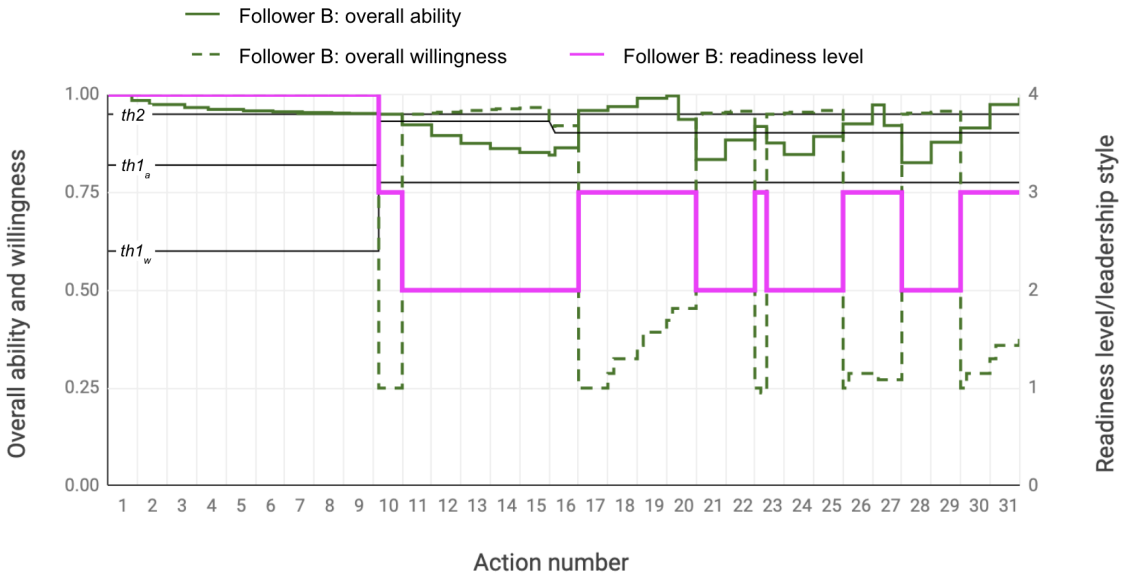


Figure 7.7 – Follower B's readiness level progression and the agent's leadership style progression during the entire medical procedure. In this scenario, the patient state remains uncritical and there are no procedure changes, so the leadership style is equal to readiness level.

Because the patient's state remains below TCM (uncritical) throughout the entire

procedure and there are no procedure changes, the leadership style is the same as readiness level. When *readiness level/leadership style* is 4 in Figure 7.7, this means that the readiness level is R1 and leadership style is S1 (directing). According to SL®, leadership style directly corresponds to readiness level unless there are situational factors at play (Hersey et al. 1988). In the next scenario, we demonstrate how a varying patient state and procedure changes can also influence the agent's leadership style.

Figure 7.8 displays the follower's readiness level progression and the agent's corresponding leadership style given that the patient's state varies and there is a procedure change sent by the medical experts.

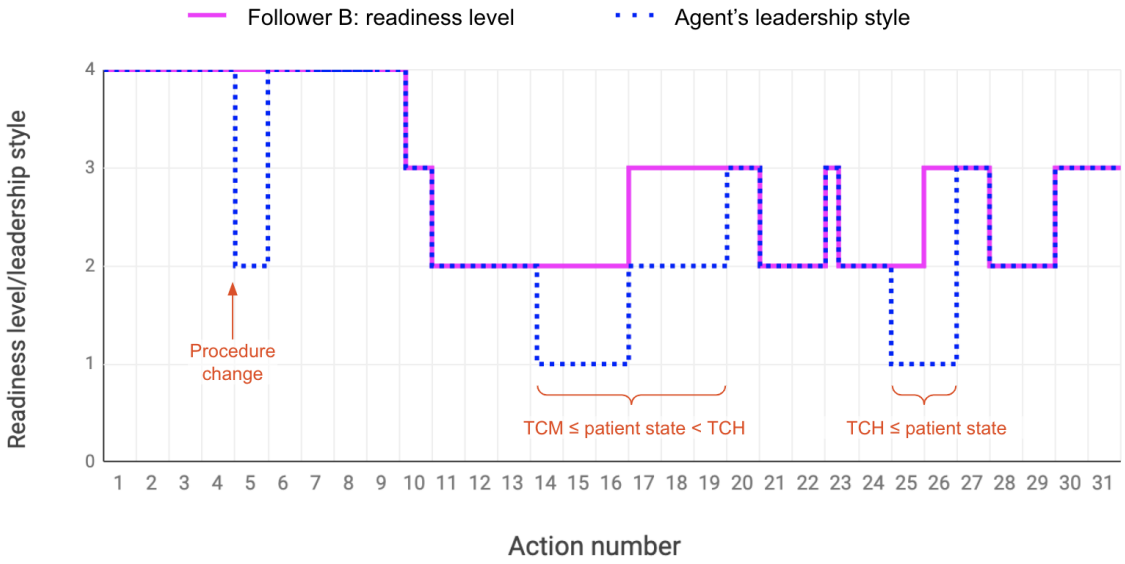


Figure 7.8 – Follower B's readiness level progression and the agent's leadership style progression during the entire medical procedure. In this scenario, the patient state varies and there are several procedure changes sent by the medical experts, so the leadership style does not directly correspond to readiness level.

As shown in Figure 7.8, a procedure change is sent for action 5. Because the readiness level is R4, the agent's leadership style becomes S2 (coaching), according to the rules laid out in section 6.4.2. Between actions 14 and 19, the patient's state is between the medium threshold TCM and the high threshold TCH. According to the rules in section 7.4, the the agent's leadership style is $readinesslevel - 1$. Then, in actions 25 and 26, the patient's state is above TCH, and so the agent's leadership is S1 (directing).

7.6 Conclusions

In this chapter, we have explained how the model of computational SL® explained in chapter 6 can be applied to a medical procedure, which requires some additions. In section 7.1, the medical procedure was briefly discussed. In section 7.2, the criteria chosen for use in this thesis were explained. These criteria were chosen in order to work with our agent system that does not utilize an activity monitor. Section 7.3 provides an explanation of the varying criteria persistence values based on follower ability. This allows the system to better preserve the patient's health by aiming to never provide less guidance to the follower than necessary. In section 7.4, we discuss how the role of the patient influences the agent's leadership style.

Section 7.5 includes working examples of how followers' readiness level might progress and how the agent's leadership style progresses as a consequence of various factors. Performance value v calculations are provided, the method of computing overall ability and willingness values is explained, the method of creating thresholds is detailed, and changes to overall ability A and willingness W are demonstrated for different situations. The implementation of our computational model of SL® is discussed in chapter 8.

Key points from Chapter 7

- Nine criteria indicate a follower's inability and unwillingness throughout the medical procedure;
- Persistence values for each criterion in the ability domain change depending on the follower's ability;
- Patient state influences the agent's leadership style;
- Working examples of the computational model are demonstrated.

Chapter 8

Implementation

Contents

Now that the model for choosing a follower's readiness level and the agent's leadership style has been detailed, it is time to put all of these elements together by implementing the agent model and the agent's behavior. In this chapter, the final chapter of part III, we detail how the agent model works.

We utilize a SAIBA-compliant framework, which was first discussed in section 2.6.1. A SAIBA framework allows for a human to communicate with an agent in real-time by enabling the agent system to process environmental information and respond accordingly (Cafaro, H. H. Vilhjálmsson, et al. 2014). Previous work that also create SAIBA-compliant virtual agent systems to accomplish a wide variety of tasks can be found in section 2.6.1.

Our model is implemented within Mascaret, a metamodel for an informed intelligent virtual environment in which an embodied virtual human can interact with a user (Querrec, Taoum, et al. 2018; H. Vilhjálmsson et al. 2007). Mascaret was chosen specifically for its past use in pedagogical scenarios in which an agent must lead a human follower through a series of steps. All the virtual and augmented reality rendering is done through Unity.

Within Mascaret, procedures such as medical procedures can be formalized and simulated, action by action, in the virtual environment. The user can follow the procedure in the virtual environment, and they can interact with virtual humans who provide assistance. Virtual humans in Mascaret are ECAs compliant with the SAIBA framework (H. Vilhjálmsson et al. 2007). Their high-level communicative intentions are translated in multimodal behavioral signals which are transformed in animation. Communicative intentions are discussed in more detail in section 8.2.

The Mascaret framework permits the modeling of semantic, structural, geometric, and topological properties of the entities in the virtual environment and their behaviors. Mascaret also defines the notion of a virtual agent by their behaviors, their communications, and their organization.

In this chapter, we describe the implementation of the agent. In section 8.1, we describe the implementation of the agent framework using Mascaret, and in section 8.2, we describe in detail how the agent's communicative intentions are created. Our conclusions are listed in section 8.3.

8.1 Agent platform in Mascaret

This section discusses the implementation of the model discussed in chapters 6 and 7 and includes the existing Mascaret framework, the additions made by our work, the specifics of the criteria implementation, and the specifics of the intent planner.

8.1.1 Existing Mascaret work

Figure 8.1 provides a description of the framework that used Mascaret *before* the work involved in this thesis. The virtual world is populated by agents (class `Agent`). Some of these agents are embodied (class `EmbodiedAgent`). Agents can own behaviors (class `Behavior`) which are executed cyclically. Among these behaviors, `ProceduralBehavior` is particularly important because it allows the agent to go through a known formalized procedure (class `Procedure`). A formalized procedure is a set of actions (class `Action`) that can be done by different agents, each one with a specific role (class `Role`) in the procedure, and by using different resources (class `Ressource`). Each resource is associated with an entity (class `Entity`) in the virtual world).

An example of what the procedure looks like is in Figure 8.2.

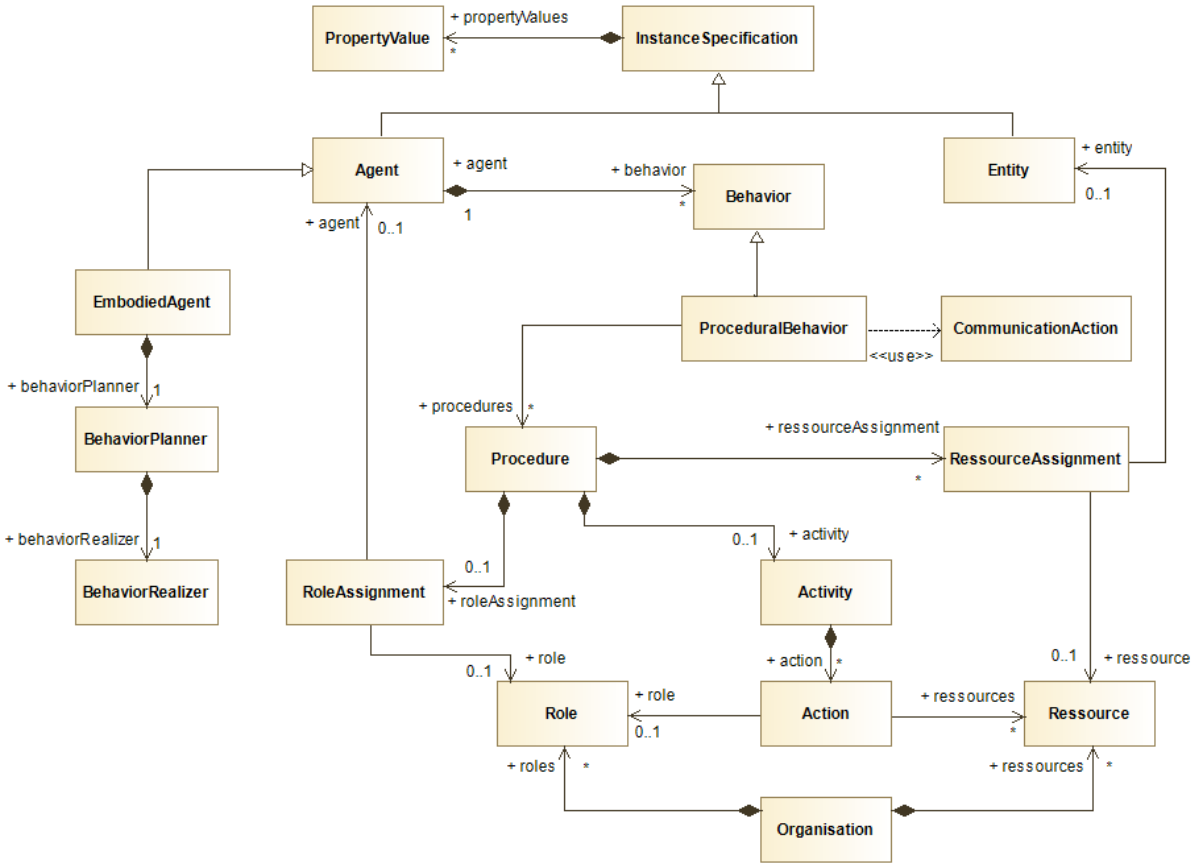


Figure 8.1 – A class diagram of the agent framework in Mascaret before the work in this thesis was completed.

Thus an agent in Mascaret can interact with and utilize the procedure through its procedural behavior to retrieve any kind of information about it, such as the current action, the next action to do, the resources needed to complete an action, which agent has which role in the procedure, and which agent should perform a certain action. For any agent or resource involved or used in the procedure, information can be also retrieved. In Mascaret, resources are associated with virtual entities (class `Entity`) and both agents and entities come from `InstanceSpecification`, which means that they each own a set of properties.

For example, when a medical procedure unfolds, the agent knows this procedure at all times thanks to its procedural behavior. The agent can ask for properties of any role in the procedure and any resource. In particular, the patient is seen as having a role in the procedure, so their current properties can be interrogated. The patient's properties can include information that describes their current health state such as the temperature, the

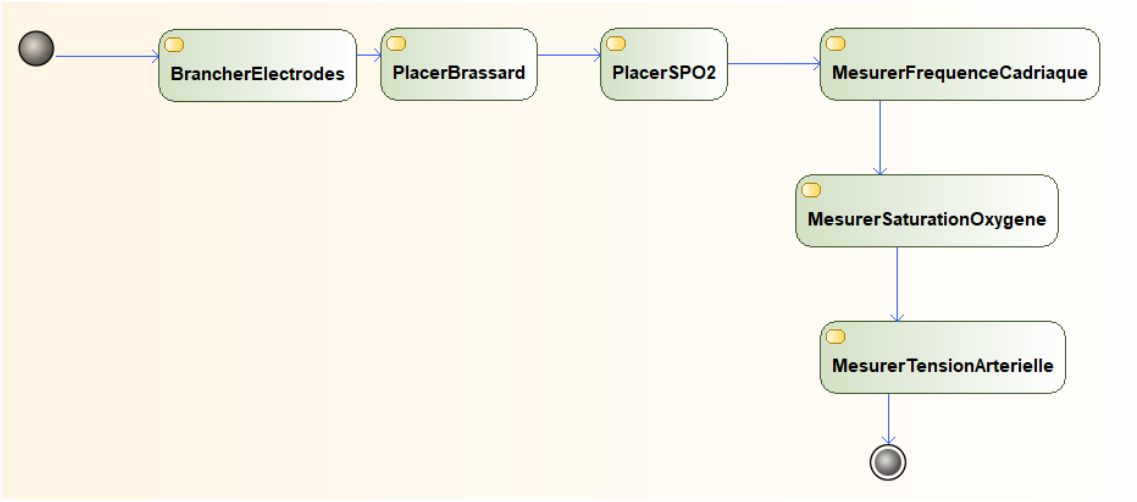


Figure 8.2 – An example of how the actions work together within the formalized procedure.

blood pressure, the blood oxygenation level, etc.

The medical experts standing by on Earth are able to monitor the virtual environment in what would normally be real-time (but instead involves a twenty-minute or more time lag due to the distance between Earth and Mars) and edit the formalized procedure within Mascaret. These edits can involve changes to any of the procedural information: changes, additions, or removals of resources; re-ordering of tasks or actions; and additions or removals of tasks and actions. When a change is sent, it is also sent with meta-data specifying that it is a change and not part of the original procedure. Because these changes are done directly to the procedure itself, they can then be sent to the agent just like the rest of the procedure actions.

Each step of the procedure could generate a communication action (class `CommunicationAction`) intended for the `User`, for example, to announce the next action to do. This communication action is sent to the agent which transforms the message into a textual or an audio signal. When the agent is embodied and owns a behavior planner (class `BehaviorPlanner`), some basic nonverbal behaviors are added to the speech and animated on a virtual agent tanks to its behavior realizer (class `BehaviorRealizer`).

8.1.2 Additions by this thesis

The work completed in this thesis builds on this existing work in Mascaret by adding additional classes to handle an agent that uses SL® according to the model described in sections 6.3 and 6.4. We integrated our proposed architecture (from Figure 6.1) result-

ing in the class diagram in Figure 8.3. Note that not everything included in this image has been implemented. Due to Covid-19 and the fact that this thesis was meant to be theoretical in nature, we chose to focus on agent speech, creating a robust algorithm for determining readiness level and leadership style (sections 6.3.1, and 6.4.2) and validating the model with simulations (section 7.5) than focus on the implementation. However, this architecture has been thoroughly planned out, and many aspects have been implemented. Individual criteria classes, along with the computation of performance values `v`, as well as communicative intentions have been implemented in the agent framework.

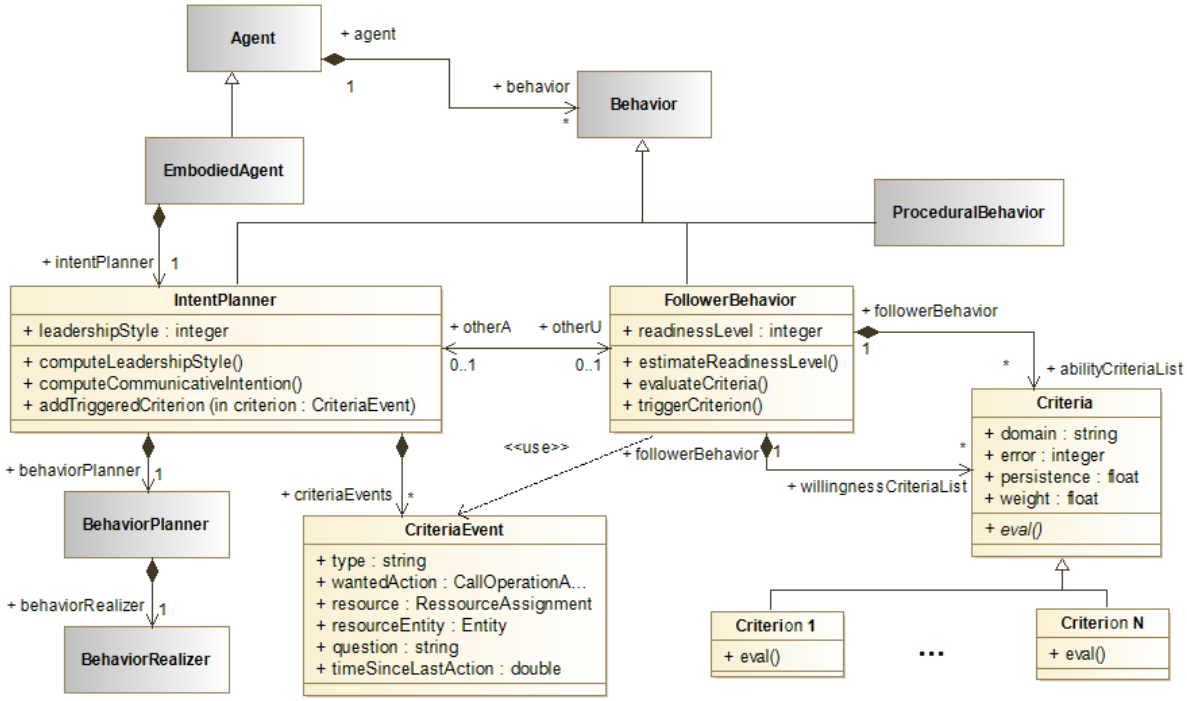


Figure 8.3 – A class diagram demonstrating this thesis' additions to Mascaret. The classes represented with grey boxes are existing, and the classes represented in yellow are new additions.

One class that has been added to the framework is `FollowerBehavior`. The FollowerBehavior class implements the first two modules of the general architecture in both Figure 6.1 and 7.2: the follower criteria triggered and the readiness level estimator.

8.1.3 Criteria

FollowerBehavior is provided to the agent representing the user. The FollowerBehavior owns two lists of criteria (class `Criteria`), one linked to the ability and the other to the

willingness of the follower. The Criteria class is specialized through heritage in a number of derived classes, one for each criterion taken into account in this work and shown later in section 7.2. These classes are referred to as `Criterion1`, etc. in Figure 8.3. Any number of criteria can easily be added or removed. However, as was explained in chapter 7, we have chosen nine criteria that indicate a user's ability and willingness during a medical procedure specifically. These criteria are shown in the class diagram in Figure 8.4.

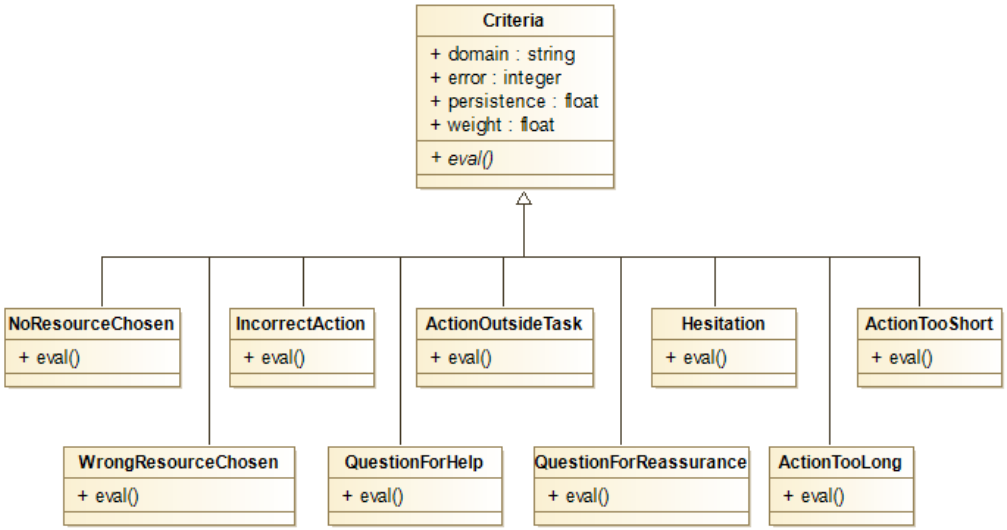


Figure 8.4 – The nine criteria developed in this thesis for use in a medical scenario are developed as individual Criteria classes.

Each criterion class inherits all the methods from the base class Criteria. In the Criteria class, the `eval()` function method is abstract, but in the individual criterion classes, the eval() function is implemented. Thus each subclass (Criterion1 or NoResourceChosen, etc.) overrides the eval() function of base class (Criteria). These subclasses know how to check if the `error` is triggered and how to compute the performance value `v`.

The follower behavior cyclically evaluates all the criteria by invoking `evaluateCriteria()`. This function calls the eval() function of each criterion and computes its performance value. As shown in Figure 8.3, each derived class of the Criteria class provides its own implementation of the eval() function. In fact, each one of them knows how to check if a certain criterion is triggered and can compute the error value and consequently the performance value as explained in section 6.3. Within FollowerBehavior, the performance value for each criterion in the ability domain is multiplied by its `weight` and then added together. The same is done for the willingness

criteria. Thus FollowerBehavior houses the overall ability and willingness values.

As mentioned in chapter 6, the computational model and implementation are designed to be flexible to allow any number of criteria to be added or removed as necessary. The only criteria parameter that is dependent on the total number of criteria is weight as all the criteria weights must add up to equal 1. These weights can be chosen intelligently, practically, and manually according to criterion importance (as explained in section 6.3.1). However, in order to make the system more flexible, all weights can be set to the same value of $1/n$ where n represents the total number of criteria. In section 9.3, we discuss the possibility of an algorithm which more intelligently but automatically chooses criteria weights based on criterion importance.

The performance value of each criterion is used by the `estimateReadinessLevel()` function to estimate the follower's readiness level.

In Mascaret, the Criteria class knows the FollowerBehavior that the criterion belongs to. Thus the Criteria class knows when a user's overall ability `A` or overall willingness `W` need to be adjusted according to the rules outlined in section 6.3.1, and therefore the individual criteria's performance values can be adjusted.

Thanks to the ProceduralBehavior owned by the agent (that represents the user and that owns the FollowerBehavior, too), each criterion (from class Criteria) can interrogate everything concerning a procedure (the action to do, resources to use, duration spent during an action, etc.).

Whenever a criterion appears (for example, when the user makes a mistake), the `triggerCriterion()` function is called and its role consists in informing the IntentPlanner of the virtual assistant (known thanks to the association between the FollowerBehavior and the IntentPlanner) by invoking its function `addTriggeredCriterion()` which receives a `CriteriaEvent`. A criteria event is an enum containing the needed information about a criterion, such as its type and other information. For example, for a criterion triggered for incorrect actions, additional information are the name of the incorrect action and the name of the correct action.

All user actions are currently taken into account with input from a keyboard and are not automatically detected with an activity monitor.

Each criterion is explained in more detail regarding its implementation below. If a criterion is triggered, then the error value is equal to 1. If a criterion is not triggered, then the error value is equal to 0. The eval() function is called regardless so that overall ability A or willingness W can be computed. Note that there are ten criteria specified here to

demonstrate the difference between *resource chosen* and *no resource chosen*, even though one criterion class handles both of those.

— Error: action in task: When the user chooses an action to do, that chosen action is compared to the correct next action. If there is no match, the user's chosen action is searched for in the list of all actions for the current task. If the chosen action is found in that list, then the criterion is triggered. The incorrect action that the user has chosen is sent to the intent planner via a criteria event. The criteria event is deleted when the user performs the correct action;
— Error: action outside task: When the user chooses an action to do, that chosen action is compared to the correct next action. If there is no match, the user's chosen action is searched for in the list of all actions for the current task. If the chosen action is not found in that list, then the criterion is triggered. The incorrect action that the user has chosen is sent to the intent planner via a criteria event. The criteria event is deleted when the user performs the correct action;
— Wrong resource chosen: When the follower chooses a resource, Mascaret compares the chosen resource to the correct resource for the next action. If they do not match, the criterion is triggered. The incorrect resource that the user has chosen and the correct resource that the user should have chosen are sent to the intent planner via a criteria event. The criteria event is deleted when the user chooses the correct resource;
— No resource chosen: When the follower begins an action but does not choose a resource, the system checks whether there is a resource for the next action. If there is a resource that should be used, the criterion is triggered. The correct resource that the user should have chosen is sent to the intent planner via a criteria event. The criteria event is deleted when the user chooses the correct resource;
— Resource chosen: When the follower begins an action and chooses a resource, the system checks whether there is a resource for the next action. If there is no resource that should be used, the criterion is triggered. The correct resource that the user should have chosen and the incorrect resource chosen are sent to the intent planner via a criteria event. The criteria event is deleted when the user puts down resource and continues working on the action;
— Action duration too short: After the user has completed, the time between the user starting the action and finishing the action is compared to the expected duration. If the actual action duration is less than the expected duration or another threshold

set by the system, then the criterion is triggered. Note that for testing purposes, the threshold we use for an action duration being too short is an action duration that is less than 0.9 times the expected duration. However, this value should be validated further with experimentation. Only the type of criteria event is sent to the intent planner. The criteria event is deleted when the agent communicates to the user (depending on the leadership style);

— Action duration too long: After the user has completed, the time between the user starting the action and finishing the action is compared to the expected duration. If the actual action duration is greater than the expected duration or another threshold, then the criterion is triggered. For testing purposes, the threshold we use for an action duration being too long is an action duration that is greater than 1.1 times the expected duration. However, this value should be validated further with experimentation. Only the type of criteria event is sent to the intent planner. The criteria event is deleted when the agent communicates to the user (depending on the leadership style);

— Question for help: When the follower asks a question, the content is evaluated. If the content involves asking for help, then the criterion is triggered. Only the type of criteria event is sent to the intent planner. The criteria event is deleted when the agent communicates to the user;

— Hesitation: After the agent provides an order, a timer starts. After a certain period of time without the user beginning the action, the criterion is triggered. The criterion is then triggered discretely after a certain time increment elapses. In this work, the initial period of time that must elapse before the criterion is triggered is set to five seconds; however, this value should be validated with experimentation. The criterion is then triggered every second after the first five seconds have passed without user activity. The time since the user's last completed action is sent to the intent planner via a criteria event. The criteria event is deleted when the agent communicates to the user (depending on the leadership style);

— Question for reassurance: When the follower asks a question, the content is evaluated. If the content involves asking for instructions or clarification about an action that is already complete, then the criterion is triggered. Only the type of criteria event is sent to the intent planner. The criteria event is deleted when the agent communicates to the user (depending on the leadership style).

8.1.4 Intent planner

The `IntentPlanner` class is the other behavior that we have introduced in Mascaret. It implements the second two modules of the general architecture in Figures 6.1 and 7.2, which is the leadership style calculator and the communicative intentions generator. This behavior is provided to the embodied agent representing the virtual assistant. Note that an embodied agent must own an intent planner among all its behaviors.

The intent planner is a cyclic behavior, so it cyclically computes the leadership style by invoking the `computeLeadershipStyle()` function (addressing module 3 from Figures 6.1 and 7.2). This function has access to the readiness level of the user thanks to the association between the IntentPlanner class and the FollowerBehavior class. Moreover, through the ProceduralBehavior owned by the base class (`Agent`), the IntentPlanner can interrogate everything concerning a procedure: changes or information about resources, actions and roles, and the patient's health state.

Readiness level and leadership style are each stored inside the IntentPlanner and are represented by an integer on the interval $[1, 4]$, with 1 representing R1 or S1, 2 representing R2 or S2, etc. Patient state, also stored inside the IntentPlanner is represented by a float. Procedure changes are sent directly to the formalized procedure. Within the IntentPlanner, an if-else statement is made to determine leadership style according to the if-else statements in sections 6.4.2 and 7.4, with the patient's state always being taken into account before procedure changes.

The IntentPlanner also generates the communicative intentions, addressing module four from Figures 6.1 and 7.2. To do that, cyclically it checks both the list of criteria and the ProceduralBehavior to find the next action of the procedure. The criteria event list contains something if FollowerBehavior has triggered any criterion. Once a `communicationAction` has been created for a criteriaEvent, the event is removed from the list. Communicative intentions are discussed in the next section.

8.2 Communicative intentions planner

In this section, we discuss how both verbal and nonverbal behavior (first covered in chapters 4 and 5) are used to create communicative intentions and actions within our agent framework.

8.2.1 Verbal communication

As mentioned in section 2.6.1, the intent planner within a SAIBA-compliant framework computes the agent's communicative intentions (Kopp et al. 2006). The intent planner relays information regarding communicative and expressive intent, without any reference to physical behavior, to the behavior planner with FML (function markup language).

To create agent speech, the intent planner generates the communicative intentions of the agent using FML, a behavior planner translates communicative intentions (what the agent wants the human follower to do) into verbal signals, and a behavior realizer transforms these signals into animation.

In our agent framework, communicative intentions are created from the current procedure action, the criteria events, and the leadership style computed within the intent planner. As explained in section 8.1, a criteria event is created every time user behavior triggers one of the criteria. When this occurs, the criteria event is stored in a list within the intent planner. Using Mascaret, communicative intentions using any number of entities from the procedure can be created.

In the `action()` function of the intent planner, the size of list of the triggered criteria is checked. If it is higher then zero, then it means that an error was performed and must be addressed. Otherwise, the list of following actions in the procedure is checked to see if the user has to do another action. In this way, errors are communicated before actions to do. According to this information and the leadership level, the agent must decide what to do.

At the end of chapter 5, we listed the rules for verbal communication in each leadership style in Table 5.16. These rules were validated by both dataset analysis and a participant evaluation and did not always match the original definitions of each leadership style (Hersey et al. 1988). For example, it was found that participants with readiness level R3 may work best under a leader that provides higher task behavior. Using these rules regarding how information is communicated (particularly whether the agent should be communicating low or high task behavior), we define rules for when a communication action is generated, as shown in Table 8.1.

When the leadership style is directing or coaching (S1 or S2), the follower is considered to have low ability, and therefore needs a lot of guidance. Leadership styles supporting and delegating (S3 and S4) are instead paired with followers who self-lead (Hersey et al. 1988). Therefore, the agent should not communicate every error and action that must be done when the leadership style is supporting or delegating. However, our results from our

Directing/coaching	— a `communicationAction` is generated every time there is an action to do — a `communicationAction` is generated every time there is a criteria event
Supporting	— a `communicationAction` is generated every time there is an action to do — a `communicationAction` is generated every time there is a criteria event *except* when the criteria event is hesitation
Delegating	— a `communicationAction` is generated only when the user tries to perform an action that is not in the task or an action out of order or asks a question

Table 8.1 – The rules for each leadership style for communicative intentions created in the intent planner.

speech evaluation (section 5.4.5 and Table 5.16) suggested that followers in R3, who are paired with supporting leaders, benefit from detailed instruction, too.

A detailed description of the action is available by using `actionToDo.Description`, and a less detailed description is available by using `actionToDo.Activity.Description`. Resources that the user wants to use can be called by using `criteriaEvent.resourceEntity`, and resources that are correct for the current action can be called using `actionToDo.getIncomingObjectNode()[0].Description`. Using the rules in Table 5.16, we store various types of information, such as structures, keywords, and phrases, in lists and then pull from those lists to create communication actions.

As an example, consider a sentence that is created for the agent using directing leadership. The lists of possible structures :

```
List<string> moodLstLS1 = new List<string>
 {"imperative", "interrogative", "indicative"};

List<string> intLstLS1 = new List<string>
 {"Can you", "Could you", "Would you"};

List<string> indLstLS1 = new List<string>
 {"I need you to", "I'd like you to",
```

```
 "I want you to", "We need to"};
```

Thus the sentence would be created like this:

```
mood = moodLstLS1[rnd.Next(0,
 moodLstLS1.Count)]

if(mood == "imperative"):
   commAction.naturalContent =
   actionToDo.Description);
elseif(mood == "interrogative"):
   commAction.naturalContent = intLstLS1
    [rnd.Next(0, intLstLS1.Count)] +
    actionToDo.Description) + "?";
else:
   commAction.naturalContent = intLstLS1
    [rnd.Next(0, indLstLS1.Count)] +
    actionToDo.Description) + ".";
```

There is no list for imperative sentences because `actionToDo.Description` begins with an infinitive verb (e.g., "Open", "Check", etc.).

Note that unstructured dialogue between the caregiver and the agent is not possible. The follower is only allowed to ask certain questions, and the agent's speech is limited to only information within the procedure. Therefore, the agent's responses to questions are fixed because the questions themselves are fixed. Using natural language only to create communication actions prevents unstructured dialogue which in turn better preserves the patient's health (Bickmore, Trinh, et al. 2018). When a criteria event is triggered at the same time that a new action is needed, all current criteria events are used to create a communicative intention first.

8.2.2 Nonverbal behavior

The taxonomy covered in chapter 3 is an important part of the agent's communication. A performative enum called `ACLPerformative` is created that includes the speech acts in our taxonomy. By attaching the relevant speech act to each communicative intention, we are better able to match the agent's speech with an appropriate nonverbal behavior.

The communicative intention is generated with FML and is sent to the BehaviorPlanner. The modality for agent communication is both verbal and nonverbal behavior, so communication actions (verbal behavior) are paired with appropriate nonverbal behavior in the behavior planner. By using a SAIBA-compliant framework, the agent can receive feedback from the environment and display behavior in real time.

The gestures discussed in chapter 4 are created using xml and are housed inside a gestuary. Inside the behavior planner, BML (Behavior Markup Language) is used to define behavior that should be performed by the agent. Each BML statement is composed of the natural content communication action from the intent planner as well as an appropriate matching gesture from the gestuary.

While this piece has not been implemented, we envision that gestures will be chosen intelligently with an algorithm, as previous works mentioned in section 2.6.2 did, or will be organized into relevant lists and randomly pulled from those, as we did with the natural content of the communication actions.

8.3 Conclusions

In this chapter, we have explained how the model components explained in chapters 6 and 7 have been implemented using Mascaret in a SAIBA-compliant framework (covered in section 8.1). Additionally, our method of creating communicative intentions using the formalized procedure, triggered criteria, and the chosen leadership style was detailed (covered in section 8.2). In our SAIBA-compliant agent framework, communicative intentions are created and translated with BML to a behavior planner and sent to a behavior realizer to be animated by the virtual agent. The behavior realizer is implemented at the CERV (Centre Européen de Réalité Virtuelle) on the ENIB campus.

To conclude, the pieces of the agent framework that have been implemented include:

— The computation of a user's readiness level, including individual classes for each criterion;

— The intent planner, including the computation of leadership style and the creation of communication actions from the procedure and the criteria events.

Additionally, the theoretical architecture here has been designed as a part of this thesis.

Key points from Chapter 8

— The computation of a user's readiness level is handled with individual classes for each criterion;
— The `IntentPlanner` includes the computation of leadership style and the creation of communication actions from the procedure and the criteria events;
— A flexible architecture allows for real-time interaction based on user behavior, patient state, and the formalized medical procedure.

Part IV

Conclusion

Chapter 9

Conclusion

Contents

In the work presented in this paper, we propose a multi-agent system that incorporates Situational Leadership® SL®) and thus allows an agent to lead a user through a medical procedure. Previous work on agents as tutors and agents in the medical domain have informed our work. Our agent facilitates a learning interaction, but more importantly, it aims to preserve the health of an injured individual during a potentially high-stress situation. To our knowledge, an agent acting as a leader of a medical situation has not been developed or studied before, and so our work presents a closure in the gap of current research.

In this chapter, we conclude this work by summarizing our contributions in section 9.1, identifying the limitations of our work in section 9.2, and presenting proposed future work in this field of study in section 9.3.

9.1 Summary of Contributions

In chapter 1, we presented the research questions that drove the work in this thesis. Those questions were:

1. How can an agent system effectively identify a follower's correct readiness level during a procedure?
2. How can an agent system identify the agent's most appropriate leadership style during a procedure?
3. How can an agent perform different styles of leadership through multimodal behavior?
4. Under what leadership style do caregivers of each type perform better?

In order to find the answers to each of these research questions, we embarked on several different projects that resulted in the following contributions (listed in the order they appear in this thesis):

1. Development of a taxonomy for an agent leading a medical procedure. Research was done into existing taxonomies for both human medical professionals and virtual agents. Because there has been no prior work on an agent leading a medical procedure, there was no taxonomy specific to this scenario. Therefore, in order to better structure agent behavior and plan an interaction with a user that results in a healthy patient, it was necessary to create a new taxonomy for our agent.

The taxonomy presented in this thesis combines some of the non-technical skills and elements that are included in existing medical leader taxonomies with relevant communicative intentions developed in our work and speech acts from existing agent taxonomies. Even though this taxonomy was developed to help manage the human-agent interaction in our system, it can also be used for human-human interaction in the medical domain, especially when more structure beyond what currently exists is needed.

This contribution addresses research question 3.

2. The identification of appropriate nonverbal behavior for a medical leader in each leadership style. Research was done into existing work on nonverbal behavior of both humans and agents in various positions of leadership. Research was also conducted into various nonverbal behaviors that could be applied to an individual in a position of authority. Through this research, a compilation of behaviors that could be used by an agent emulating different leadership styles was created. This compilation is applicable to not only to human-agent interaction but to human-human interaction as well.

These behaviors were grouped by low and high task behavior and were also grouped by behaviors to avoid and behaviors that should generally always be used. This compilation can be found in Appendix B.

This contribution addresses research question 3.

3. Identification of medical leader speech belonging to each leadership style so that the agent can use the most appropriate speech in different situations. A dataset of medical leader speech was compiled, annotated with leadership style by four annotators, and then used in a number of analyses to extract the linguistic and semantic properties of each utterance:

a. Statistical analysis provides linguistic rules for each leadership style regarding grammatical mood, utterance length, whether the utterance is direct or indirect, and the speech acts used most;
b. Clustering on sequences of words and sequences of parts-of-speech tags resulted in sequences that belong to each leadership style;
c. Utterances were dependency parsed to find the latent dependencies between different parts of each whole sentence in the dataset, resulting in dependency triples, and clustered upon to find which kinds of latent structures belong to each leadership style;
d. Utterances were chunked to find specific kinds of phrases present in each leadership style.

These linguistic rules were compiled in Table 5.16. These rules were then used to create utterances in each leadership style and were evaluated by eighty-seven participants in an experiment. Analysis of the results of this evaluation validated most of the linguistic rules found from the analysis in the list above while others were disproved. For example, participants as a whole disagreed on what relationship behavior looks like in speech and the sequence "it looks like" that belonged in the rules for delegating speech was not responded to well by participants.

Work on the linguistics that should belong to each leadership style has not been done before, and so this part is applicable to not only human-computer relationships but to human-human relationships as well.

This contribution addresses research question 3.

4. Identification of the most appropriate leadership style for each type of follower. With further analysis of the experiment data from contribution 5 above, we have found the leadership style under which that each type of follower performs best. In this thesis, "best" refers to a the leadership style that results in a follower that feels capable and willing and remains calm. We found that each readiness level was best matched with its corresponding leadership style in terms of speech, and therefore we validate the SL® model as well as our linguistic rules.

This contribution addresses research question 4.

5. Creation and implementation of an algorithm that uses follower behavior and outputs readiness level. A mathematical algorithm using several equations was developed using multiple parameters to compute readiness level in real time. The algorithm was designed so that the number of criteria could be flexible. It is also designed to keep in line with SL® in that followers can only progress or regress from one readiness level to the next. Also, when in doubt, the follower is assumed to have less ability than they may have in order to preserve the patient's health.

After careful study of previous work, several follower behaviors (referred to as criteria) were identified that would present themselves during a medical emergency and are able to be recorded in real-time from an isolated location like Mars. By working with a medical professional, the values of the different parameters for each criterion were created and then validated for each task in the medical procedure multiple times over multiple follower scenarios, many of which were not included in this thesis for the sake of page length. However, the scenarios in section 7.5 do reflect that the algorithm works in the sense that the most appropriate readiness level is reflected at all times given the user's behavior.

This contribution addresses research question 1.

6. Development of a method that determines the most appropriate leadership style. The study of previous work on SL® and leadership in the medical domain resulted in an understanding of what factors should determine an agent's leadership style while leading a medical procedure. These factors were the follower's readiness level, the follower's familiarity with the current procedure, and the patient's state of health. An algorithm was developed to compute the general criticality of the situation using the patient's state, using flexible parameters similar to those used in the computation of readiness level. A

follower's familiarity with the procedure is used to determine their starting readiness level, and when a procedure change is sent, the system automatically assumes the follower is not familiar with the changed action (again, when in doubt, we assume the follower is less capable than they are in order to preserve the patient's health).

Then, using readiness level, patient state, and procedure changes, we develop a method of computing the most appropriate leadership style for the situation in real time. This method is also validated over multiple different medical scenarios with the guidance of a medical professional.

This contribution addresses research question 2.

7. Design of an agent framework that uses SL® to allow an agent to lead a follower through a medical procedure. We chose Mascaret as the interaction framework because it allows procedures to be formalized and simulated, action by action, in the virtual environment. Mascaret also allows the user to follow the procedure in the virtual environment and interact with a virtual agent who guides them. The system allows for the collection of data from the patient the user. The system also allows for communication between the control center on Earth and the caregivers on Mars.

This thesis contributes to existing Mascaret frameworks by one, developing several class structures to handle incoming user behavior and compute readiness level from that behavior; and two, developing an intent planner that computes leadership style and generates communicative intentions and communication actions from the procedure itself and user behavior. The system was built in such a way so that the number of criteria is flexible and thus criteria can be added or removed as needed.

This contribution addresses research questions 1, 2, and 3.

8. The design of an experiment that validates the computational model of SL®. An experiment was designed to validate the computational model described in contributions 5 and 6. The purpose of this experiment is to validate the method of choosing the readiness level and the method of choosing the leadership style by (1) analyzing whether the correct model parameters have been chosen and therefore whether the right readiness level has chosen, and (2) analyzing whether the most appropriate leadership style was chosen. The experiment evaluates the parameter values within the model, compares the

calculated readiness level to the participants' actual readiness levels, assesses how the participants perceive the agent's leadership style, and examines which leadership style leads to the best outcomes for the medical procedure. The experiment design is available in Appendix H.

This contribution would address research questions 1-4.

In the following section, we discuss some limitations that occurred during this thesis work.

9.2 Limitations

The contributions presented in this thesis are not without limitations. In this section, we discuss the limitations of our experimental study and limitations regarding the implementation of the agent system.

9.2.1 Experimental limitations

Our perception of speech experimentation was conducted online, and so we had no control over the environment in which they were performing the study. While we presented a thorough introduction that explained the task, we could not control whether participants thoroughly read and understood this task. These factors could have influenced the interest and perceptions of the participants.

In the initial phase of this thesis work, we intended to recruit participants to interact with the agent in a laboratory setting to further test how they might trigger criteria events and also react to the nonverbal behaviors we planned to include. Due to pandemic constraints, we had to adjust our settings to accommodate online research, which resulted in a pivot toward a focus on agent speech as analysis and experimentation regarding perceptions of speech were much easier to collect with online studies.

9.2.2 Agent limitations

As mentioned throughout this thesis, our agent utilizes a text-to-speech module, which means that speech elements such as prosody cannot be taken into account. Therefore, our work on agent speech was entirely text-based with the exception of the creation of the dataset which was based on speech in videos.

Despite the limitations, our work branched into several areas and thus opened up many opportunities for further work. In the following section, we discuss research that could continue with regard to this project.

9.3 Future Work

In this section, opportunities for future work are provided to better explore different aspects of this project. This thesis was proposed as a theoretical foundation to lay the groundwork for an agent leading a medical procedure. The following propositions would be valuable contributions to this work but would involve an entirely new PhD thesis to resolve.

9.3.1 Computing readiness level

First, there is contradicting research regarding what questions mean in terms of a follower's ability and/or willingness. In section 2.4, we discussed several works that either determined that questions asked by followers was a sign of ability (McKellar 1986), a sign of inability (Bosse et al. 2017), or a sign of willingness (Andersson and Edberg 2010; Delaney 2003; Jewell 2013). After consultations with Dr. Wall, the criterion for asking questions was split into two categories: one for questions for help and one for questions for reassurance. However, experimentation and further study should be conducted to verify whether the act of asking a question is a sign of ability, inability, or willingness. It is possible that the answer to this question is unique to each individual follower.

Second, when developing values for each parameter for each task in the medical procedure, only one medical professional was consulted to ensure the correct readiness level was reflected given a variety of scenarios and criteria triggered (Dr. Wall). However, it would be valuable to consult multiple medical professionals for the parameter values to ensure that there is an agreement on what readiness level a follower should have during different scenarios.

Third, each action has an expected duration, set within the formalized medical procedure in Mascaret. Our definition of an action that has taken too long is the same for each action in the procedure and is a duration that is longer than 1.1 times the expected duration of the action. The definition of an action that is too short is also the same for each action and is defined as a duration that is shorter than 0.9 times the expected action

duration. "Long" and "short" actions should be defined more intelligently with the help of medical professionals and should also be validated with experimentation.

Fourth, the weights of each criterion in the ability domain must add up to equal 1, and the same is true for those in the willingness domain. These weight values can be created manually. However, when automatically creating new criteria classes in the agent system, this means that all the existing criteria weights must be revisited. A method around this is to set each weight to $1/n$ where n is the total number of criteria in that domain. In order to preserve weight as an importance variable, in which it describes the most important and least important criteria in each domain, a new algorithm could be created that allows for the weight parameter to be automatically set depending on the importance of each criterion.

Fifth, and lastly, we did not explore interactions between criteria. However, interactions should be explored and could lead to a more intelligent understanding of how a follower's ability in one area affects their ability in another area. For example, when a user triggers the criterion *wrong resource chosen*, perhaps the criterion *no resource chosen* should be affected even if it was not directly triggered.

9.3.2 Agent behavior

Our research into agent speech largely involved the criteria events triggered by the user's behavior. Whether criteria events are communicated depends entirely on the leadership style. However, this could be changed. For example, the communication of a criteria event could also depend on that criterion's weight in the task. If, for example, the criterion *error: action in task* has a very low weight as it does during the inquiring task, perhaps a communication action does not need to be generated in self-led leadership styles supporting and delegating.

Within the intent planner, a communication action is generated for each criteria event. However, it could be interesting to see whether a single communication action could be generated with two criteria events occur very close together.

Additionally, we do not generate communication actions for procedure changes or patient state specifically. However, experimentation could be conducted to evaluate whether these kinds of communication actions help followers of each readiness level.

Lastly, nonverbal behavior should be explored more in general, including an intelligent way of pairing it to communicative intentions. Nonverbal behavior is discussed further in the next section.

9.3.3 Further experimentation

As mentioned above, nonverbal behavior should be explored further. This should involve an experimentation to evaluate the perceptions of different behaviors with and without speech to see which behaviors should be included in the gestuary. Nonverbal behavior should also be tested in the virtual environment during the procedure to examine whether the presence of behaviors affects user performance or feelings.

More experimentation is also needed to understand how individuals perceive relationship behavior and how varying levels of task and relationship behavior influence a follower's performance during a task. This is most easily done by inviting participants to perform the procedure in the virtual environment. Appendix H contains the details of an experiment designed to validate our model with participants.

It could also be interesting to design the agent's speech based on just one participant's preferences. As shown by the work on agent speech covered in chapter 5, there were many linguistic characteristics that annotators and participants did not agree on. By developing a highly-personalized agent, we may be able to test whether an agent personalized to a user enables that user to perform better than if the agent uses the general linguistic rules found in Table 5.16.

Generally, our computational model of SL® including our criteria events should be validated with a final experimentation in which participants of different readiness levels perform the procedure.

9.3.4 Multiple caregivers

Finally, something that was not explored in our work was how the agent would lead the procedure if there were multiple caregivers. This would involve a procedure that includes multiple roles for users, research into the dynamics between caregivers, and a new model of interaction in which the agent communicates to each caregiver and the group of caregivers as well.

Bibliography

Akoglu, Haldun (2018), « User's guide to correlation coefficients », *in*: *Turkish Journal of Emergency Medicine* 18 (3), pp. 91–93, ISSN: 2452-2473, DOI: https://doi.org/10.1016/j.tjem.2018.08.001.

Allwood, Jens (1976), *Linguistic Communication as Action and Cooperation: A Study in Pragmatics*, Göteborg, Sweden: Department of Linguistics, University of Göteborg.

Allwood, Jens, Joakim Nivre, and Elisabeth Ahlsen (1992), « On the Semantics and Pragmatics of Linguistic Feedback », *in*: *Journal of Semantics* 9 (1), DOI: 10.1093/jos/9.1.1.

Andersson, Petra and Anna-Karin Edberg (2010), « The Transition From Rookie to Genuine Nurse: Narratives From Swedish Nurses 1 Year After Graduation », *in*: *Journal of continuing education in nursing* 41 (4), pp. 186–92, DOI: 10.3928/00220124-20100326-05.

Anikina, Tatiana and Ivana Kruijff-Korbayova (Sept. 2019), « Dialogue Act Classification in Team Communication for Robot Assisted Disaster Response », *in*: *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, SIGdial '19, Stockholm, Sweden: Association for Computational Linguistics, pp. 399–410, DOI: 10.18653/v1/W19-5946.

Araszewski, Daniele, Michele Bianca Bolzan, Juliana Helena Montezeli, and Aida Maris Peres (2014), « The Exercising Of Leadership In The View Of Emergency Room Nurses », *in*: *Cogitare Enferm* 19 (1), pp. 41–47, DOI: 10.5205/1981-8963-v11i4a15242p1709-1715-2017.

Babur, Önder and Loek Cleophas (Jan. 2017), « Using n-grams for the Automated Clustering of Structural Models », *in*: *International Conference on Current Trends in Theory and Practice of Informatics*, SOFSEM '17, Limerick, Ireland: Springer, pp. 510–524, DOI: 10.1007/978-3-319-51963-0_40.

Baylor, Amy L., Soyoung Kim, Chanhee Son, and Miyoung Lee (2009), « Designing nonverbal communication for pedagogical agents: When less is more », *in*: *Computers in Human Behavior* 25 (2), pp. 450–457, ISSN: 0747-5632, DOI: 10.1016/j.chb.2008.10.008.

Ben-Asher, Noam, Jin-Hee Cho, and Sibel Adal (June 2018), « Adaptive Situational Leadership Framework », *in*: *Proceedings of the 2018 Conference on Cognitive and Computational Aspects of Situation Management*, CogSIMA '18, Boston, MA, USA: IEEE, pp. 63–9, DOI: 10.1109/COGSIMA.2018.8423977.

Bennett, Casey C. and Kris Hauser (2013), « Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach », *in*: *Artificial Intelligence in Medicine* 57 (1), pp. 9–19, DOI: 10.1016/j.artmed.2012.12.003.

Berber Sardinha, Tony and Marcia Veirano Pinto (2021), « A linguistic typology of American television », *in*: *International Journal of Corpus Linguistics* 26 (1), pp. 127–160, DOI: 10.1075/ijcl.00039.ber.

Bergmann, Kirsten, Sebastian Kahl, and Stefan Kopp (Aug. 2013), « Modeling the semantic coordination of speech and gesture under cognitive and linguistic constraints », *in*: *Proceedings of the Thirteenth International Conference on Intelligent Virtual Agents*, IVA '13, Edinburgh, Scotland: ACM, pp. 203–216, DOI: 10.1007/978-3-642-40415-3_18.

Bernard, Denys and Alexandre Arnold (2019), « Cognitive interaction with virtual assistants: From philosophical foundations to illustrative examples in aeronautics », *in*: *Computers in Industry* 107 (1), pp. 33–49, DOI: 10.1016/j.compind.2019.01.010.

Bevacqua, Elisabetta, Ken Prepin, Etienne de Sevin, Radosław Niewiadomski, and Catherine Pelachaud (May 2009), « Reactive behaviors in SAIBA architecture », *in*: *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '09, Budapest, Hungary: ACM, URL: https://www.enib.fr/~bevacqua/site/papers/aamasWorkshop09.pdf.

Biancardi, Beatrice, Angelo Cafaro, and Catherine Pelachaud (Nov. 2017), « Analyzing first impressions of warmth and competence from observable nonverbal cues in expert-novice interactions », *in*: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI '17, Glasgow, Scotland: ACM, pp. 341–349, DOI: 10.1145/3136755.3136779.

Bickmore, Timothy W., Reza Asadi, Aida Ehyaei, Harriet Fell, Lori Henault, Stephen Intille, Lisa Quintiliani, Ameneh Shamekhi, Ha Trinh, Katherine Waite, Christopher Shanahan, and Michael K. Paasche-Orlow (2015), « Context-awareness in a persistent hospital companion agent », *in*: *Proceedings of the 15th International Conference on Intelligent Virtual Agents*, IVA '15, Delft, The Netherlands: ACM, pp. 332–342, DOI: 10.1007/978-3-319-21996-7_35.

Bickmore, Timothy W., Ha Trinh, Stefan Olafsson, Teresa K. O'Leary, Reza Asadi, Nathaniel M. Rickles, and Ricardo Cruz (2018), « Patient and Consumer Safety Risks When Using Conversational Assistants for Medical Information: An Observational Study of Siri, Alexa, and Google Assistant », *in*: *Journal of Medical Internet Research* 20 (9), DOI: 10.2196/11510.

Black, L.A., Michael F. Mctear, Norman D Black, Roy Harper, and M. Lemon (July 2005), « Appraisal of a conversational artefact and its utility in remote patient monitoring », *in*: *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems*, CBMS '05, IEEE, pp. 506–8, DOI: 10.1109/CBMS.2005.33.

Blanchard, K., P. Zigarmi, and D. Zigarmi (1999), *Leadership and the One Minute Manager: Increasing Effectiveness Through Situational Leadership*, The one minute manager library, HarperCollins, ISBN: 9780688039691, URL: https://books.google.ie/books?id=geebCcCyLI8C.

Blanchard, Kenneth H. (1967), « College Boards of Trustees: A Need for Directive, Leadership », *in*: *Academy of Management Journal* 10 (4), pp. 409–417, DOI: 10.5465/255273.

Blanchard, Kenneth H., Drea Zigarmi, and Robert B. Nelson (1993), « Situational Leadership® after 25 years: a retrospective », *in*: *Journal of Leadership Organizational Studies* 1 (1), pp. 21–36, DOI: 10.1177/107179199300100104.

Bosse, Tibor, Rob Duell, Zulfiqar Ali Memon, Jan Treur, and Natalie van der Wal (2017), « Computational model-based design of leadership support based on situational leadership theory », *in*: *SIMULATION: Transactions of The Society for Modeling and Simulation International* 93 (7), DOI: 10.1177/0037549717693324.

Brody, Shmuel (2005), « Cluster-Based Pattern Recognition in Natural Language Text », Master's thesis, The Hebrew University of Jerusalem.

Buche, Cédric, Ronan Querrec, Pierre De Loor, and Pierre Chevaillier (2004), « MASCARET: A pedagogical multi-agent system for virtual environment for training », *in*: *Journal of Distance Education Technologies* 2 (4), pp. 41–61, DOI: 10.1109/CYBER.2003.1253485.

Bunt, Harry (Jan. 2009), « The DIT++ taxanomy for functional dialogue markup », *in*: *Proceedings of 8th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '09, Budapest, Hungary: ACM, URL: https://www.researchgate.net/publication/254786679_The_DIT_taxanomy_for_functional_dialogue_markup.

Burgoon, Judee K., Joseph A. Bonito, B. Bengtsson, Carl Cederberg, M. Lundeberg, and L. Allspach (2000), « Interactivity in human–computer interaction: a study of credibility, understanding, and influence », *in*: *Computers in Human Behavior* 16 (6), pp. 553–574, ISSN: 0747-5632, DOI: 10.1016/S0747-5632(00)00029-7.

Burgoon, Judee K. and Norah Dunbar (Jan. 2006), « Nonverbal Expressions of Dominance and Power in Human Relationships », *in*: *The Sage handbook of nonverbal communication*, Sage Publications, Inc, pp. 279–297, DOI: 10.4135/9781412976152.n15.

Bush, Tony (2003), *Theories of Educational Leadership and Management*, Thousand Oaks, California, USA: SAGE Publications, ISBN: 9781526432131.

Cafaro, Angelo, Merijn Bruijnes, Jelte van Waterschoot, Catherine Pelachaud, Mariet Theune, and Dirk Heylen (Aug. 2017), « Selecting and expressing communicative functions in a SAIBA-compliant agent framework », *in*: *Proceedings of the 17th International Conference on Intelligent Virtual Agents*, IVA '17, Stockholm, Sweden: ACM, pp. 73–82, DOI: 10.1007/978-3-319-67401-8_8.

Cafaro, Angelo, Hannes Högni Vilhjálmsson, Timothy W. Bickmore, Dirk Heylen, and Catherine Pelachaud (2014), « Representing Communicative Functions in SAIBA with a Unified Function Markup Language », *in*: *Proceedings of the 14th International Conference on Intelligent Virtual Agents*, IVA '14, Boston, Mass., USA: ACM, pp. 73–82, DOI: 10.1007/978-3-319-09767-1_11.

Carney, Dana, Amy Cuddy, and Andy J Yap (2010), « Power posing: brief nonverbal displays affect neuroendocrine levels and risk tolerance », *in*: *Psychological science* 21 (10), pp. 1363–8, DOI: 10.1177/0956797610383437.

Carney, Dana, Judith A. Hall, and Lavonia Smith LeBeau (2005), « Beliefs about the nonverbal expression of social power », *in*: *Journal of Nonverbal Behavior* 29 (2), pp. 105–123, DOI: 10.1007/s10919-005-2743-z.

Cassell, Justine, Hannes Vilhjálmsson, and Timothy W. Bickmore (Aug. 2001), « BEAT: the behavior expression animation toolkit », *in*: *Proceedings of the 28th annual conference on computer graphics and interactive techniques*, SIGGRAPH '01, Los Angeles, Cali., USA: ACM, pp. 477–486, DOI: 10.1145/383259.383315.

Chase, Francis S. and Egon G. Guba (1955), « Administrative roles and behavior », *in*: *Review of Educational Research* 24 (4)), pp. 281–298, DOI: 10.3102/00346543025004281.

Chaudhry, Noureen and Manzoor Arif (2012), « Teachers' Nonverbal Behavior and Its Impact on Student Achievement », *in*: *International Education Studies* 5 (4), pp. 56–64, DOI: 10.5539/ies.v5n4p56.

Chetty, Girija and Matthew White (Aug. 2019), « Embodied Conversational Agents and Interactive Virtual Humans for Training Simulators », *in*: *The 15th International Conference on Auditory-Visual Speech Processing*, AVSP '19, Melbourne, Australia: ISCA, pp. 73–7, DOI: 10.21437/AVSP.2019-15.

Chiu, Chung-Cheng, Louis-Philippe Morency, and Stacy Marsella (Aug. 2015), « Predicting Co-verbal Gestures: A Deep and Temporal Modeling Approach », *in*: *The 15th International Conference on Intelligent Virtual Agents*, IVA '15, Delft, The Netherlands: ACM, pp. 152–166, DOI: 10.1007/978-3-319-21996-7_17.

Chollet, Mathieu, Magalie Ochs, and Catherine Pelachaud (Aug. 2014), « From Nonverbal Signals Sequence Mining to Bayesian Networks for Interpersonal Attitudes Expression », *in*: *The 14th International Conference on Intelligent Virtual Agents*, IVA '14, DOI: 10.1007/978-3-319-09767-1_15.

Chung, Hyunji and Sangjin Lee (2018), « Intelligent Virtual Assistant knows Your Life », arXiv: 1803.00466.

Cisneros, Fanny, Victoria I. Marín, Martha Lucia Orellana, Nancy Peré, and Dolores Zambrano (June 2019), « Design Process of an Intelligent Tutor to Support Researchers in Training », *in*: *Proceedings of EdMedia and Innovate Learning*, Amsterdam, Netherlands: AACE, pp. 1079–1084, URL: https://www.researchgate.net/publication/334260936_Design_Process_of_an_Intelligent_Tutor_to_Support_Researchers_in_Training.

Ciuffani, Bianca Malaika (July 2017), « Non-verbal Communication and Leadership: The impact of hand gestures used by leaders on follower job satisfaction », *in*: *9th IBA Bachelor Thesis Conference*, IBA, Enschede, The Netherlands: University of Twente, DOI: 10.3389/fcomm.2022.869084.

Colace, Francesco, Massimo De Santo, Luca Greco, and Paolo Napoletano (2014), « Text Classification Using a Few Labeled Examples », *in*: *Computational Human Behavior* 30, pp. 689–697, ISSN: 0747-5632, DOI: 10.1016/j.chb.2013.07.043.

Cuban, Larry (1998), *The Managerial Imperative and the Practice of Leadership in Schools*, SUNY series, Educational Leadership, Albany, New York: SUNY Press, ISBN: 9780887065941.

Darioly, Annick and Marianne Mast (Jan. 2014), « The role of nonverbal behavior for leadership: An integrative review », *in*: *Leader interpersonal and influence skills: The soft skills of leadership*, ed. by R. E. Riggio S. J. Tan, Routledge/Taylor Francis Group, pp. 73–100, ISBN: 9781135018177.

Delaney, Colleen (2003), « Walking a Fine Line: Graduate Nurses' Transition Experiences During Orientation », *in*: *Journal of Nursing Education* 42 (10), pp. 437–443, DOI: 10.3928/0148-4834-20031001-05.

Dermouche, Soumia and Catherine Pelachaud (Oct. 2019), « Engagement Modeling in Dyadic Interaction », *in*: *Proceedings of the 21st International Conference on Multimodal Interaction*, ICMI '19, Suzhou, Jiangsu, China: ACM, pp. 440–445, DOI: 10.1145/3340555.3353765.

Descartin, Karina S., Richard P. Menger, and Sharmila D. Watkins (2015), *Application of Advances in Telemedicine for Long-Duration Space Flight*, NASA, URL: https://humanresearchroadmap.nasa.gov/gaps/closureDocumentation/1-Karina-S_TM-2015-218562.pdf.

Economidou-Kogetsidis, Maria (2016), « Variation in evaluations of the (im)politeness of emails from L2 learners and perceptions of the personality of their senders », *in*: *Journal of Pragmatics* 106 (2), pp. 1–19, ISSN: 0378-2166, DOI: https://doi.org/10.1016/j.pragma.2016.10.001.

Ensink, Titus and Christoph Sauer (2003), « Social-functional and cognitive approaches to discourse interpretation », *in*: *Framing and Perspectivising in Discourse*, ed. by Titus Ensink and Christoph Sauer, John Benjamins Publishing, pp. 1–21, ISBN: 9789027296641, DOI: 10.1075/pbns.111.02ens.

Ensmenger, Nathan (2012), « The Digital Construction of Technology: Rethinking the History of Computers in Society », *in*: *Technology and Culture* 53 (4), pp. 753–76, DOI: 10.2307/41682741.

Felice, Rachele and Paul Deane (2012), « Identifying speech acts in e-mails: toward automated scoring of the TOEIC® e-mail task », *in*: *ETS Research Report Series* 2012 (2), pp. 1–62, DOI: 10.1002/j.2333-8504.2012.tb02298.x.

Ferreira, Rafael, Rafael Lins, Steven Simske, Fred Freitas, and Marcelo Riss (2016), « Assessing Sentence Similarity through Lexical, Syntactic and Semantic Analysis », *in*: *Computer Speech and Language* 39 (C), pp. 1–28, DOI: 10.1016/j.csl.2016.01.003.

Fiore, Arlene M., Vaishali Naik, and Eric M. Leibensperger (2015), « Air Quality and Climate Connections », *in*: *Journal of the Air Waste Management Association* 65 (6), pp. 645–685, DOI: 10.1080/10962247.2015.1040526.

Flin, Rhona, Rona Elizabeth Patey, R. Glavin, and N. Maran (2010), « Anaesthetists' non-technical skills », *in*: *British Journal of Anaesthesia* 105 (1), pp. 38–44, DOI: 10.1093/bja/aeq134.

Følstad, Asbjørn and Marita Skjuve (Aug. 2019), « Chatbots for Customer Service: User Experience and Motivation », *in*: *Proceedings of the 1st International Conference on Conversational User Interfaces*, CUI '19, Dublin, Ireland: Association for Computing Machinery, pp. 1–9, DOI: 10.1145/3342775.3342784.

Forster, Alan J., Heather D. Clark, Alex Menard, Natalie Depuis, Robert Chernish, Natasha Chandok, Asmat Khan, Megal Letourneau, and Carl van Walraven (2005), « Effect of a nurse team coordinator on outcomes for hospitalized medicine patients », *in*: *The American Journal of Medicine* 118 (10), pp. 1148–1153, DOI: 10.1016/j.amjmed.2005.04.019.

Frechette, Casey and Roxana Moreno (2010), « The Roles of Animated Pedagogical Agents' Presence and Nonverbal Communication in Multimedia Learning Environments », *in*: *Journal of Media Psychology* 22 (2), pp. 61–72, DOI: 10.1027/1864-1105/a000009.

Goddard, Cliff (2002), « Directive speech acts in Malay (Bahasa Melayu): an ethnopragmatic perspective », *in*: *Cahiers de praxématique* 38 (1), pp. 113–143, DOI: 10.4000/praxematique.582.

Goleman, Daniel (2000), *Leadership that gets results*, Harvard business review, ISBN: 978-1633692626.

Grammar Wiz (2021), *Parts of a Sentence*, https://www.grammarwiz.com/parts-of-a-sentence.html.

Greenleaf, Robert K. (2015), *The Servant as Leader*, The Greenleaf Center for Servant Leadership, ISBN: 978-0982201220.

Greven, Tim (July 2017), « The Influence of Non-Verbal Behaviour on Meeting Effectiveness and Pro-Active Behaviour: A Video Observational Study », *in*: *9th IBA Bachelor Thesis Conference*, IBA '17, University of Twente, URL: https://essay.utwente.nl/72781/1/Greven_BA_BMS.pdf.

Grimaldi, Michele and Catherine Pelachaud (June 2021), « Generation of Multimodal Behaviors », *in*: *Proceedings of the 21st International Conference on Intelligent Virtual Agents*, IVA '21, ACM, pp. 98–100, DOI: 10.1145/3472306.3478368.

Guillaume, Demary, Stéphane Dubourdieu, Coralie Berenguer, Jean-Claude Martin, Laurence Bolot, Francis Beguec, Benoit Frattini, and Virginie Demulier (2018), « Comportements non verbaux pour des subordonnées virtuels passifs vs. proactifs d'une équipe médicale : analyse de vidéos de simulation MOTS-CLEFS », preprint.

Gumpert, Raymond A. and Ronald K. Hambleton (1979), « How Xerox managers fine-tune managerial styles to employee maturity and task needs », *in*: *Management Review* 68 (12), pp. 8–12.

Gümüş, Sedat, Mehmet Bellibaş, Murat Esen, and Emine Gümüş (2018), « A systematic review of studies on leadership models in educational research from 1980 to 2014 », *in*: *Educational Management Administration Leadership* 46 (1), pp. 25–48, DOI: 10.1177/1741143216659296.

Haag, Kathrin and Hiroshi Shimodaira (Sept. 2016), « Bidirectional LSTM Networks Employing Stacked Bottleneck Features for Expressive Speech-Driven Head Motion Synthesis », *in*: *Proceedings of the 16th International Conference on Intelligent Virtual Agents*, IVA '16, Los Angeles, CA, USA: ACM, pp. 198–207, DOI: 10.1007/978-3-319-47665-0_18.

Hamoudi, Yassine (2016), « Extracting RDF triples using the Stanford Parser », PhD thesis, University of Lyon.

Hansen, Miriam, Sabine Fabriz, and Sebastian Stehle (2015), « Cultural Cues in Students' Computer-Mediated Communication: Influences on E-mail Style, Perception of the Sender, and Willingness to Help », *in*: *Journal of Computer-Mediated Communication* 20 (3), pp. 278–294, ISSN: 1083-6101, DOI: 10.1111/jcc4.12110.

Harrigan, Jinni A., Karen S. Lucic, Denise Kay, Anne McLaney, and Robert Rosenthal (1991), « Effect of expresser role and type of self-touching on observers' perceptions », *in*: *Journal of Applied Social Psychology* 21 (7), pp. 585–609, DOI: 10.1111/j.1559-1816.1991.tb00538.x.

Harris, Jonathan and Paula Mayo (2018), « Taking a case study approach to assessing alternative leadership models in health care », *in*: *British Journal of Nursing* 27 (11), pp. 608–613, DOI: 10.12968/bjon.2018.27.11.608.

Hasegawa, Dai, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi (Nov. 2018), « Evaluation of Speech-to-Gesture Generation Using Bi-Directional

LSTM Network », *in*: *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, IVA '18, Sydney, NSW, Australia: ACM, pp. 79–86, DOI: 10.1145/3267851.3267878.

Henrickson, Parker, Rhoda Flin, A. McKinley, and S. Yule (2013), « The Surgeons' Leadership Inventory (SLI): a taxonomy and rating system for surgeons' intraoperative leadership skills », *in*: *BMJ Simulation and Technology Enhanced Learning* 205 (6), pp. 745–751, DOI: 10.1016/j.amjsurg.2012.02.020.

Hersey, Paul, Kenneth H. Blanchard, and Dewey E. Johnson (1988), « Situational Leadership », *in*: *Management of Organizational Behavior: Leading Human Resources*, 5th ed., Prentice-Hall, pp. 169–201, ISBN: 978-0135512685.

Hjortdahl, Magnus, Amund H Ringen, Anne-Cathrine Naess, and Torben Wisborg (2009), « Leadership is the essential non-technical skill in the trauma team - results of a qualitative study », *in*: *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicinevolume* 17 (48), DOI: 10.1186/1757-7241-17-48.

Hoegen, Rens, Deepali Aneja, Daniel McDuff, and Mary Czerwinski (2019), « An End-to-End Conversational Style Matching Agent », arXiv: 1904.02760.

Huang, Ann, Pascal Knierim, Francesco Chiossi, Lewis Chuang, and Robin Welsch (Apr. 2022), « Proxemics for Human-Agent Interaction in Augmented Reality », *in*: *Proceedings of the ACM International Conference on Human Factors in Computing Systems*, CHI 2022, New Orleans, LA, USA: ACM, pp. 1–13, DOI: 10.1145/3491102.3517593.

Huang, Yazhou, Andrew Feng, Marcelo Kallmann, and Ari Shapiro (Nov. 2012), « An Analysis of Motion Blending Techniques », *in*: *Proceedings of the 5th International Conference on Motion In Games*, MIG 2012, Rennes, France: Springer-Verlag, pp. 232–243, DOI: 10.1007/978-3-642-34710-8_22.

Huang, Yuyun, Emer Gilmartin, and Nick Campbell (Sept. 2016), « Conversational Engagement Recognition Using Auditory and Visual Cues », *in*: *Proceedings of the 22nd European Conference on Speech Communication and Technology*, Interspeech '16, San Francisco, CA, USA: ISCA, pp. 590–594, DOI: 10.21437/Interspeech.2016-846.

Hudlicka, Eva (2013), « Virtual Training and Coaching of Health Behavior: Example from Mindfulness Meditation Training », *in*: *Patient Educ Couns* 92 (2), pp. 160–6, DOI: 10.1016/j.pec.2013.05.007.

Hussein, Hadher, Abbood Ad-darraji, Thomas Chow, Voon Foo, Shaik Abdul, and Malik Mohamed Ismail (2012), « Offering as a Commissive and Directive Speech Act: Con-

sequence for Cross-Cultural Communication », *in*: *International Journal of Scientific and Research Publications* 2 (3), ISSN: 2250-3153.

Ishikawa, Hirono, Hideki Hashimoto, Makoto Kinoshita, Shin Fujimori, Teruo Shimizu, and Eiji Yano (2006), « Evaluating medical students' non-verbal communication during the objective structured clinical examination », *in*: *Medical Education* 40 (12), pp. 1180–1187, DOI: 10.1111/j.1365-2929.2006.02628.x.

Jelodar, Hamed, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao (2019), « Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a Survey », *in*: *Multimedia Tools Applications* 78 (11), pp. 15169–211, DOI: 10.1007/s11042-018-6894-4.

Jewell, Andrea (2013), « Supporting the novice nurse to fly: A literature review », *in*: *Nurse education in practice* 13 (4), pp. 323–7, DOI: 10.1016/j.nepr.2013.04.006.

Judge, Timothy A. and Ronald F. Piccolo (2004), « Transformational and transactional leadership: A meta-analytic test of their relative validity », *in*: *Journal of Applied Psychology* 89 (5), p. 755, DOI: 10.1037/0021-9010.89.5.755.

Jurafsky, Daniel and James Martin (Feb. 2008), *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, vol. 2, London, UK: Pearson, ISBN: 978-0130950697.

Kates, James M. (2005), « Principles of Digital Dynamic-Range Compression », *in*: *Trends in Amplification* 9 (2), pp. 45–76, DOI: 10.1177/108471380500900202.

Kearns, William R., Neha Kaura, Myra Divina, Cuong Vo, Dong Si, and Teresa M Ward Weichao Yuwen (Apr. 2020), « A Wizard-of-Oz Interface and Persona-Based Methodology for Collecting Health Counseling Dialog », *in*: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, Hawai'i, USA: ACM, pp. 1–9, DOI: 10.1145/3334480.3382902.

Kendon, Adam (2004), *Gesture: Visible Action as Utterance*, Cambridge, UK: Cambridge University Press, ISBN: 19781316264935.

Key, Mary Ritchie (Apr. 1973), « Nonverbal Behavior in Speech Acts », *in*: *Proceedings of the 1973 Conference on Sociology of Language and Theory of Speech Acts*, Bielefeld, Germany: University of Bielefeld, URL: https://files.eric.ed.gov/fulltext/ED086005.pdf.

Keyton, Joann (2018), « Interaction Process Analysis (IPA) », *in*: *The Cambridge Handbook of Group Interaction Analysis*, ed. by Elisabeth Brauner, Margarete Boos, and

Michaela Kolbe, Cambridge Handbooks in Psychology, Cambridge, UK: Cambridge University Press, pp. 441–450, DOI: 10.1017/9781316286302.024.

Khoury, Richard (2012), « Sentence Clustering Using Parts-of-Speech », *in*: *International Journal of Information Engineering and Electronic Business* 4 (1), DOI: 10.5815/ijieeb.2012.01.01.

Kocaballi, A. Baki, Shlomo Berkovsky, Juan Carlos Quiroz, and Liliana Laranjo (2019), « The Personalization of Conversational Agents in Health Care: Systematic Review », *in*: *Journal of Medical Internet Research* 11 (21), DOI: 10.2196/15360.

Kondrak, Grzegorz (Oct. 2005), « N-Gram Similarity and Distance », *in*: *Proceedings of the 12th Conference on String Processing and Information Retrieval*, SPIRE '05, Buenos Aires, Argentina: Springer, pp. 115–26, DOI: 10.1007/11575832_13.

Kopp, Stefan, Brigitte Krenn, Stacy Marsella, Andrew Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn Thórisson, and Hannes Vilhjálmsson (2006), « Towards a Common Framework for Multimodal Generation: The Behavior Markup Language », *in*: IVA '06, Marina Del Rey, CA, USA: ACM, pp. 205–17, DOI: 10.1007/11821830_17.

Krämer, Nicole, Gale Lucas, Lea Schmitt, and Jonathan Gratch (2018), « Social snacking with a virtual agent – On the interrelation of need to belong and effects of social responsiveness when interacting with artificial entities », *in*: *International Journal of Human-Computer Studies* 109 (4), pp. 112–21, DOI: https://doi.org/10.1016/j.ijhcs.2017.09.001.

Krämer, Nicole, Nina Simons, and Stefan Kopp (Sept. 2007), « The Effects of an Embodied Conversational Agent's Nonverbal Behavior on User's Evaluation and Behavioral Mimicry », *in*: *Proceedings of the 7th International Conference on Intelligent Virtual Agents*, IVA '07, Paris, France, pp. 238–51, DOI: 10.1007/978-3-540-74997-4_22.

Krathwohl, David R. (2002), « A Revision of Bloom's Taxonomy: An Overview », *in*: *Theory Into Practice* 41 (4), pp. 212–218, DOI: 10.1207/s15430421tip4104_2.

Krishna, Sooraj (2021), « Modelling communicative behaviours for different roles of pedagogical agents », PhD thesis, University of Sorbonne.

Kruijff-Korbayova, Ivana, Francis Colas, Mario Gianni, Fiora Pirri, Joachim Greeff, Koen Hindriks, Mark Neerincx, Petter Ogren, Tomás Svoboda, and Rainer Worst (2015), « TRADR Project: Long-Term Human-Robot Teaming for Robot Assisted Disaster Response », *in*: *Künstliche Intelligenz* 29 (2), pp. 193–201, DOI: 10.1007/s13218-015-0352-5.

Kulms, Philipp and Stefan Kopp (Sept. 2016), « The Effect of Embodiment and Competence on Trust and Cooperation in Human–Agent Interaction », *in*: *Proceedings of the 16th International Conference on Intelligent Virtual Agents*, IVA '16, Los Angeles, CA, USA: ACM, DOI: 10.1007/978-3-319-47665-0_7.

Kulms, Philipp, Nikita Mattar, and Stefan Kopp (Aug. 2015), « An Interaction Game Framework for the Investigation of Human–Agent Cooperation », *in*: *Proceedings of the 15th International Conference on Intelligent Virtual Agents*, Delft, The Netherlands: ACM, pp. 399–402, DOI: 10.1007/978-3-319-21996-7_43.

Lakoff, George (2002), *Moral Politics: How Liberals and Conservatives Think*, Chicago, Illinois, USA: University of Chicago Press, ISBN: 978-0226467719.

Lance, Brent and Stacy C. Marsella (Sept. 2007), « Emotionally Expressive Head and Body Movement During Gaze Shifts », *in*: *Proceedings of the 7th International Conference on Intelligent Virtual Agents*, IVA '07, Paris, France: ACM, pp. 72–85, DOI: 10.1007/978-3-540-74997-4_8.

Landis, J. Richard and Gary G. Koch (1977), « The Measurement of Observer Agreement for Categorical Data », *in*: *Biometrics* 33 (1), pp. 159–74, DOI: 10.2307/2529310.

Laposata, Michael (2014), « Errors in clinical laboratory test selection and result interpretation: commonly unrecognized mistakes as a cause of poor patient outcome », *in*: *Diagnosis: Official Journal of the Society to Improve Diagnosis in Medicine* 1 (1), pp. 85–8, DOI: doi:10.1515/dx-2013-0010.

Laranjo, Liliana, Adam G Dunn, Huong Ly Tong, and A. Baki Kocaballi (2018), « Conversational agents in healthcare: A systematic review », *in*: *Journal of the American Medical Informatics Association* 25 (9), pp. 1248–58, DOI: 10.1093/jamia/ocy072.

Lee, Jina and Stacy Marsella (Aug. 2006), « Nonverbal Behavior Generator for Embodied Conversational Agents », *in*: *Proceedings of the 6th International Conference on Intelligent Virtual Agents*, IVA '06, Marina Del Rey, CA, USA: ACM, pp. 243–55, DOI: 10.1007/11821830_20.

Lee, Jina and Stacy C. Marsella (2010), « Predicting Speaker Head Nods and the Effects of Affective Information », *in*: *IEEE Transactions on Multimedia* 12 (6), pp. 552–62, DOI: 10.1109/TMM.2010.2051874.

Lee, Sun Kyong, Pavitra Kavya, and Sarah C. Lasser (2021), « Social interactions and relationships with an intelligent virtual agent », *in*: *International Journal of Human-Computer Studies* 150, DOI: https://doi.org/10.1016/j.ijhcs.2021.102608.

Lee-Kelley, Liz (2002), « Situational leadership: Managing the virtual project team », *in*: *Journal of Management Development* 21 (6), pp. 461–76, DOI: 10.1108/02621710210430623.

Levenshtein, V. I. (1966), « Binary Codes Capable of Correcting Deletions, Insertions and Reversals », *in*: *Soviet Physics Doklady* 10 (8), pp. 707–710, URL: https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf.

Lewin, Kurt and Ronald Lippitt (1938), « An Experimental Approach to the Study of Autocracy and Democracy: A Preliminary Note », *in*: *Sociometry* 1 (3), pp. 292–300, URL: https://www.jstor.org/stable/2785585.

— (1939), « Patterns of aggressive behavior in experimentally created social climates », *in*: *The Journal of Social Psychology* 10 (2), pp. 269–299, URL: https://www.tandfonline.com/doi/abs/10.1080/00224545.1939.9713366.

Liao, Soohyun Nam, Sander Valstar, Kevin Thai, Christine Alvarado, Daniel Zingaro, William G. Griswold, and Leo Porter (2019), « Behaviors of Higher and Lower Performing Students in CS1 », *in*: *Proceedings of the 2019 Conference on Innovation and Technology in Computer Science Education*, ITiCSE '19, Aberdeen, UK: ACM, DOI: 10.1145/3304221.3319740.

Lin, Dekang and Xiaoyun Wu (2009), « Phrase Clustering for Discriminative Learning », *in*: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, ACL '09, Suntec, Singapore: Association for Computational Linguistics, pp. 1030–8, URL: https://aclanthology.org/P09-1116.pdf.

Litman, Diane, Susannah Paletz, Zahra Rahimi, Stefani Allegretti, and Caitlin Rice (Nov. 2016), « The Teams Corpus and Entrainment in Multi-Party Spoken Dialogues », *in*: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, EMNLP '16, pp. 1421–31, DOI: 10.18653/v1/D16-1149.

Lourdeaux, Domitile, Zoubida Afoutni, Marie-Hélène Ferrer, Nicolas Sabouret, Virginie Demulier, Jean-Claude Martin, Laurence Bolot, Vincent Boccara, and Romain Lelong (July 2019), « VICTEAMS: A Virtual Environment to Train Medical Team Leaders to Interact with Virtual Subordinates », *in*: *Proceedings of the 19th International Conference on Intelligent Virtual Agents*, IVA '19, ACM, pp. 241–3, DOI: 10.1145/3308532.3329418.

Lucas, Gale, Albert Rizzo, Jonathan Gratch, Stefan Scherer, Giota Stratou, Jill Boberg, and Louis-Philippe Morency (2017), « Reporting Mental Health Symptoms: Breaking

Down Barriers to Care with Virtual Human Interviewers », *in*: *Frontiers in Robotics and AI* 4 (51), DOI: 10.3389/frobt.2017.00051.

Mahabal, Abhijit, Jason Baldridge, Burcu Karagol Ayan, Vincent Perot, and Dan Roth (2020), « Text Classification with Few Examples using Controlled Generalization », arXiv: 2005.08469.

Malandrakis, Nikolaos, Minmin Shen, Anuj Goyal, Shuyang Gao, Abhishek Sethi, and Angeliki Metallinou (2020), « Controlled Text Generation for Data Augmentation in Intelligent Artificial Agents », arXiv: 1910.03487.

Malik, Usman, Mukesh Barange, Julien Saunier, and Alexandre Pauchet (Nov. 2018), « Performance Comparison of Machine Learning Models Trained on Manual vs ASR Transcriptions for Dialogue Act Annotation », *in*: *Proceedings of the 30th International Conference on Tools with Artificial Intelligence*, ICTAI '18, Volos, Greece: IEEE, pp. 1013–7, DOI: 10.1109/ICTAI.2018.00156.

Mann, William (1980), *Toward a Speech Act Theory for Natural Language Processing*, tech. rep. ISI/RR-79-75, Los Angeles, CA, USA: University of Southern California Marina del Rey Information Sciences Institute, URL: https://apps.dtic.mil/sti/pdfs/ADA087250.pdf.

Manser, Tanja (2009), « Teamwork and patient safety in dynamic domains of healthcare: a review of the literature », *in*: *Acta Anaesthesiologica Scandinavica* 53 (2), pp. 143–51, DOI: 10.1111/j.1399-6576.2008.01717.x.

Maricchiolo, Fridanna, Augusto Gnisci, Marino Bonaiuto, and Gianluca Ficca (2009), « Effects of different types of hand gestures in persuasive speech on receivers' evaluations », *in*: *Language, Cognition and Neuroscience* 24 (2), pp. 239–266, DOI: 10.1080/01690960802159929.

Mariooryad, Soroosh and Carlos Busso (2012), « Generating Human-Like Behaviors Using Joint, Speech-Driven Models for Conversational Agents », *in*: *IEEE Transactions on Audio Speech and Language Processing* 20 (8), pp. 2329–40, DOI: 10.1109/TASL.2012.2201476.

Marsella, Stacy, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapirok (July 2013), « Virtual Character Performance From Speech », *in*: *Proceedings of the 12th Eurographics Symposium on Computer Animation*, SCA '13, Anaheim, CA, USA: ACM, pp. 25–35, DOI: 10.1145/2485895.2485900.

Matsumoto, David and Hyisung C. Hwang (2012), « Cultural Similarities and Differences in Emblematic Gestures », *in*: *Journal of Nonverbal Behavior* 37 (1), pp. 1–27, DOI: 10.1007/s10919-012-0143-8.

McKellar, Nancy A. (1986), « Behaviors Used in Peer Tutoring », *in*: *The Journal of Experimental Education* 54 (3), pp. 163–7, DOI: 10.1080/00220973.1986.10806416.

Mignault, Alain and Avi Chaudhuri (2003), « The Many Faces of a Neutral Face: Head Tilt and Perception of Dominance and Emotion », *in*: *Journal of Nonverbal Behavior* 27 (2), pp. 111–132, DOI: 10.1023/A:1023914509763.

Miner, Adam S, Arnold Milstein, Stephen Schueller, and Roshini Hegde (2016), « Smartphone-Based Conversational Agents and Responses to Questions About Mental Health, Interpersonal Violence, and Physical Health », *in*: *JAMA Internal Medicine* 176 (5), pp. 619–25, DOI: 10.1001/jamainternmed.2016.0400.

Moher, David, Ashley Lynn Weinberg, Richard Hanlon, and Kathryn Runnalls (1992), « Effects of a medical team coordinator on length of hospital stay », *in*: *Canadian Medical Association journal / journal de l'Association medicale canadienne* 146 (4), pp. 511–5, URL: https://pubmed.ncbi.nlm.nih.gov/1737315/.

Montenegro, César, Asier López Zorrilla, Javier Mikel Olaso, Roberto Santana, Raquel Justo, Jose A. Lozano, and María Inés Torres (2019), « A Dialogue-Act Taxonomy for a Virtual Coach Designed to Improve the Life of Elderly », *in*: *Multimodal Technologies and Interaction* 3 (3), p. 52, DOI: 10.3390/mti3030052.

Morineau, Thierry, Pascal Chapelain, Marion Le Courtois, and Jean-Marc Gac (2017), « Fields of promoted actions for facilitating multitasking activity during a medical emergency », *in*: *BMJ Simulation and Technology Enhanced Learning* 3 (2), pp. 70–2, DOI: 10.1136/bmjstel-2016-000182.

Morineau, Thierry, Pascal Chapelain, and Philippe Quinio (2016), « Task management skills and their deficiencies during care delivery in simulated medical emergency situation: A classification », *in*: *Intensive and Critical Care Nursing* 34, pp. 42–50, DOI: 10.1016/j.iccn.2015.11.001.

Moss, Jacqueline, Yan Xiao, and Siti Zubaidah (2002), « The Operating Room Charge Nurse: Coordinator and Communicator », *in*: *Journal of the American Medical Informatics Association* 9 (6 Suppl 1), S70–4, DOI: 10.1197/jamia.M1231.

Murell, K. L. (1997), « Emergent theories of leadership for the next century: Towards relational concepts », *in*: *Organization Development Journal* 15 (3), pp. 35–42.

Nakhal, Bilal (2017), « Generation of communicative intentions for virtual agents in an intelligent virtual environment : application to virtual learning environment », PhD thesis, École Nationale d'Ingénieurs de Brest.

Navarro, Joe and Marvin Karlins (2008), *What Every Body Is Saying: An Ex-FBI Agent's Guide to Speed-Reading People*, 1st ed., New York City, NY, USA: William Morrow Paperbacks, ISBN: 9780061755668.

Nguyen, Truong-Huy D., Elin Carstensdottir, Nhi Ngo, Magy Seif El-Nasr, Matt Gray, Derek Isaacowitz, and David Desteno (Aug. 2015), « Modeling Warmth and Competence in Virtual Characters », *in*: *Proceedings of the 15th International Conference on Intelligent Virtual Agents*, IVA '15, Delft, The Netherlands: ACM, DOI: 10.1007/978-3-319-21996-7_18.

Nicogossian, Arnauld E. (2016), *Space Physiology and Medicine*, 4th ed., New York, NY, USA: Springer-Verlag, ISBN: 978-1-4939-6650-9.

Northouse, Peter (2007), *Leadership: Theory and Practice*, New York, NY, USA: Sage, ISBN: 978-1483317533.

Oliva, Jesús, José Ignacio Serrano, María Dolores del Castillo, and Ángel Iglesias (2011), « SyMSS: A syntax-based measure for short-text semantic similarity », *in*: *Data Knowledge Engineering* 70 (4), pp. 390–405, DOI: https://doi.org/10.1016/j.datak.2011.01.002.

Özateş, Şaziye, Arzucan Ozgur, and Dragomir Radev (May 2016), « Sentence Similarity based on Dependency Tree Kernels for Multi-document Summarization », *in*: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC '16, Portoroz, Slovenia: European Language Resources Association (ELRA), pp. 2833–8, ISBN: 978-2-9517408-9-1.

Papworth, Mark, Derek Milne, and George Boak (2009), « An exploratory content analysis of situational leadership », *in*: *Journal of Management Development* 28 (7), pp. 593–606, DOI: 10.1108/02621710910972706.

Pasquariello, Stefano and Catherine Pelachaud (2002), « Greta: A Simple Facial Animation Engine », *in*: *Soft Computing and Industry*, pp. 511–25, DOI: 10.1007/978-1-4471-0123-9_43.

Patterson, Miles L. (1990), « Functions of non-verbal behavior in social interaction », *in*: *Handbook of Language and Social Psychology*, Hoboken, NJ, USA: Wiley, pp. 101–20, ISBN: 9780471924814.

Pecune, Florian, Angelo Cafaro, Magalie Ochs, and Catherine Pelachaud (Sept. 2010), « Evaluating Social Attitudes of a Virtual Tutor », *in*: *Proceedings of the 16th International Conference on Intelligent Virtual Agents*, IVA '10, Los Angeles, CA, USA: ACM, pp. 245–55, DOI: 10.1007/978-3-319-47665-0_22.

Philip, Pierre, Stéphanie Bioulac, Alain Sauteraud, Cyril Chaufton, and Jérôme Olive (2014), « Could a Virtual Human Be Used to Explore Excessive Daytime Sleepiness in Patients? », *in*: *Presence* 23 (4), pp. 369–76, DOI: 10.1162/PRES_a_00197.

Philip, Pierre, Lucile Dupuy, Marc Auriacombe, Fuschia Serre, Etienne de Sevin, Alain Sauteraud, and Jean-Arthur Micoulaud Franchi (2020), « Trust and acceptance of a virtual psychiatric interview between embodied conversational agents and outpatients », *in*: *NPJ Digital Medicine* 3 (1), DOI: 10.1038/s41746-019-0213-y.

Philip, Pierre, Jean-Arthur Micoulaud Franchi, Patricia Sagaspe, Etienne de Sevin, Jérôme Olive, Stephanie Bioulac, and Alain Sauteraud (2017), « Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders », *in*: *Scientific Reports* 7 (1), DOI: 10.1038/srep42656.

Pluwak, Agnieszka (2016), « Towards Application of Speech Act Theory to Opinion Mining », *in*: *Cognitive Studies | Études cognitives* 16, pp. 33–44, DOI: 10.11649/cs.2016.004.

Poggi, Isabella (Sept. 2001), « The Lexicon and the Alphabet of Gesture, Gaze, and Touch », *in*: *Proceedings of the 9th International Conference on Intelligent Virtual Agents*, IVA '09, Madrid, Spain: ACM, pp. 235–6, DOI: 10.1007/3-540-44812-8_20.

Popat, Shraddha, Pramod Deshmukh, and Vishakha Metre (Oct. 2017), « Hierarchical document clustering based on cosine similarity measure », *in*: *Proceedings of the 1st International Conference on Intelligent Systems and Information Management*, ICISIM '17, Maharashtra, India: IEEE, pp. 153–9, DOI: 10.1109/ICISIM.2017.8122166.

Querrec, Ronan, Cédric Buche, Éric Maffre, and Pierre Chevaillier (2004), « Multiagent Systems for Virtual Environment for Training. Application to Fire-Fighting », *in*: *Advanced Technology for Learning* 1 (1), DOI: 10.2316/Journal.208.2004.1.202-1453.

Querrec, Ronan, Joanna Taoum, Bilal Nakhal, and Elisabetta Bevacqua (Nov. 2018), « Model for Verbal Interaction between an Embodied Tutor and a Learner in Virtual Environments », *in*: *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, IVA '18, Sydney, NSW, Australia: ACM, pp. 197–202, DOI: 10.1145/3267851.3267895.

Ravenet, Brian, Angelo Cafaro, Beatrice Biancardi, Magalie Ochs, and Catherine Pelachaud (Aug. 2015), « Conversational Behavior Reflecting Interpersonal Attitudes in Small Group Interactions », *in*: *Proceedings of the 15th International Confefence on Intelligent Virtual Agents*, IVA '15, Delft, The Netherlands: ACM, pp. 375–88, DOI: 10.1007/978-3-319-21996-7_41.

Ravenet, Brian, Magalie Ochs, and Catherine Pelachaud (Aug. 2013), « From a User-created Corpus of Virtual Agent's Non-verbal Behavior to a Computational Model of Interpersonal Attitudes », *in*: *Proceedings of the 13th International Conference on Intelligent Virtual Agents*, IVA '13, Edinburgh, UK: ACM, DOI: 10.1007/978-3-642-40415-3_23.

Ravenet, Brian, Catherine Pelachaud, Chloé Clavel, and Stacy Marsella (July 2018), « Automating the Production of Communicative Gestures in Embodied Characters », *in*: *Frontiers in Psychology* 9, DOI: 10.3389/fpsyg.2018.01144.

Rosenberg, Andrew and Julia Hirschberg (2009), « Charisma perception from text and speech », *in*: *Speech Communication* 51 (7), pp. 640–55, DOI: https://doi.org/10.1016/j.specom.2008.11.001.

Ruane, Elayne, Sinead Farrell, and Anthony Ventresque (Nov. 2020), « User Perception of Text-Based Chatbot Personality », *in*: *Proceedings of the 2020 International Workshop on Chatbot Research and Design*, CONVERSATIONS '20, Amsterdam, The Netherlands: Springer, pp. 32–47, DOI: 10.1007/978-3-030-68288-0_3.

Ruttkay, Zsofia (2001), « Constraint-Based Facial Animation », *in*: *Constraints* 6 (1), pp. 85–113, DOI: 10.1023/A:1009853410360.

Ruttkay, Zsófia, Job Zwiers, Herwin Welbergen, and Dennis Reidsma (Aug. 2006), « Towards a Reactive Virtual Trainer », *in*: *Proceedings of the 6th International Conference on Intelligent Virtual Agents*, IVA '06, Marina Del Rey, CA, USA: ACM, pp. 292–303, DOI: 10.1007/11821830_24.

Ryu, Jeeheon and Amy Baylor (2005), « The psychometric structure of pedagogical agent persona », *in*: *Cognition and Learning* 2, pp. 291–314, URL: https://www.researchgate.net/publication/228356035_The_psychometric_structure_of_pedagogical_agent_persona.

Sacavem, Antonio, Rui Cruz, Maria Sousa, Albérico Rosário, and João Salis Gomes (2019), « An Integrative Literature Review on Leadership Models for Innovative Organizations », *in*: *Journal of Reviews on Global Economics* 8, pp. 1741–51, DOI: 10.6000/1929-7092.2019.08.156.

Sadoughi, Najmeh and Carlos Busso (2017), « Speech-Driven Animation with Meaningful Behaviors », arXiv: 1708.01640.

Saerbeck, Martin, Tom Schut, Christoph Bartneck, and Maddy D. Janse (Apr. 2010), « Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor », *in*: *Proceedings of the 28th International Conference on Human Factors in Computing Systems*, CHI '10, Atlanta, GA, USA: ACM, pp. 1613–1622, DOI: 10.1145/1753326.1753567.

Samter, Wendy (2006), « Nonverbal Communication », *in*: *Explaining Communication: Contemporary Theories and Exemplars*, Mahwah, NJ, USA: Lawrence Erlbaum Associates, pp. 39–62, ISBN: 9780805839586.

Schug, Joanna, David Matsumoto, Yutaka Horita, Toshio Yamagishi, and Kemberlee Bonnet (2010), « Emotional expressivity as a signal of cooperation », *in*: *Evolution and Human Behavior* 31 (2), pp. 87–94, DOI: 10.1016/j.evolhumbehav.2009.09.006.

Schyns, Birgit and Gisela Mohr (2004), « Nonverbal Elements of Leadership Behaviour », *in*: *German Journal of Human Resource Management / Zeitschrift für Personalforschung* 18 (3), pp. 289–305, DOI: 10.1177/239700220401800303.

Searle, John R. (1979), *Expression and Meaning: Studies in the Theory of Speech Acts*, Cambridge, UK: Cambridge University Press, ISBN: 9780521313933.

Shamir, Boas and Jane M. Howell (1999), « Organizational and contextual influences on the emergence and effectiveness of charismatic leadership », *in*: *The Leadership Quarterly* 10 (2), pp. 257–83, DOI: 10.1016/S1048-9843(99)00014-4.

Sims, Henry P., Samer Faraj, and Seokhwa Yun (2009), « When should a leader be directive or empowering? How to develop your own situational theory of leadership », *in*: *Business Horizons* 52 (2), pp. 149–158, DOI: 10.1016/j.bushor.2008.10.002.

Smith, Howard A. (1979), « Nonverbal behavior aspects in teaching », *in*: *Review of Educational Research* 49 (4), pp. 631–72, DOI: 10.2307/1169988.

Song, Peng, Yun Jin, Li Zhao, and Minghai Xin (2014), « Speech Emotion Recognition Using Transfer Learning », *in*: *IEICE Transactions on Information and Systems* E97.D (9), pp. 2530–2, DOI: 10.1587/transinf.2014EDL8038.

Spillane, James P. (2005), « Distributed Leadership », *in*: *The Educational Forum* 69 (2), pp. 143–150, DOI: 10.1080/00131720508984678.

Stevenson, Mark and Mark Greenwood (2009), « Dependency Pattern Models for Information Extraction », *in*: *Research on Language and Computation* 7 (1), pp. 13–39, DOI: 10.1007/s11168-009-9061-2.

Straßmann, Carolin, Astrid Rosenthal von der Pütten, Ramin Yaghoubzadeh, Raffael Kaminski, and Nicole Krämer (2016), « The Effect of an Intelligent Virtual Agent's Nonverbal Behavior with Regard to Dominance and Cooperativity », *in*: *Proceedings of the 16th International Conference on Intelligent Virtual Agents*, IVA '16, Los Angeles, CA, USA: ACM, pp. 15–28, DOI: 10.1007/978-3-319-47665-0_2.

Tanaka, Hiroki, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura (2017), « Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders », *in*: *PLoS ONE* 8 (12), e0182151, DOI: doi.org/10.1371/journal.pone.0182151.

Taoum, Joanna, Anaïs Raison, Elisabetta Bevacqua, and Ronan Querrec (June 2018), « An Adaptive Tutor to Promote Learners' Skills Acquisition during Procedural Learning », *in*: *Proceedings of the 14th International Conference on Intelligent Tutoring Systems*, ITS '18, Montréal, Quebec, Canada: IEEE, URL: http://ceur-ws.org/Vol-2354/w4paper5.pdf.

The Five Grammatical Moods (2022), https://osuwritingcenter.okstate.edu/blog/2020/11/6/the-five-grammatical-moods.

Thompson, Geir and Lars Glasø (2018), « Situational leadership theory: a test from a leader-follower congruence approach », *in*: *Leadership Organization Development Journal* 39 (5), pp. 574–91, DOI: 10.1108/LODJ-01-2018-0050.

Thórisson, Kristinn (2002), « Natural Turn-Taking Needs No Manual: Computational Theory and Model, from Perception to Action », *in*: *Multimodality in Language and Speech Systems*, Berlin, Germany: Springer, pp. 173–207, DOI: 10.1007/978-94-017-2367-1_8.

Trapnell, P.D. and R.H. Broughton (2006), *The Interpersonal Questionnaire (IPQ): Duodecant Markers of Wiggins' Interpersonal Circumplex*, The University of Winnipeg. Unpublished data, URL: http://www.paultrapnell.com/measures/IPQ-revised.pdf.

Trastek, Victor, Neil Hamilton, and Emily Niles (2014), « Leadership Models in Health Care - A Case for Servant Leadership », *in*: *Mayo Clinic proceedings* 89 (3), pp. 374–81, DOI: 10.1016/j.mayocp.2013.10.012.

Uhl-Bien, Mary (2003), « The Future of Leadership Development », *in*: *The future of leadership development*, Mahwah, NJ, USA: Lawrence Erlbaum Associates Publishers, chap. Relationship Development as a Key Ingredient for Leadership Development, pp. 129–47.

Urbain, Jerome, Radoslaw Niewiadomski, Elisabetta Bevacqua, Thierry Dutoit, Alexis Moinet, Catherine Pelachaud, Benjamin Picart, Joölle Tilmanne, and Johannes Wagner (2010), « AVLaughterCycle », *in*: *Journal on Multimodal User Interfaces* 4 (1), pp. 47–58, DOI: 10.1007/s12193-010-0053-1.

van Diggele, Christie, Annette Burgess, Chris Roberts, and Craig Mellis (2020), « Leadership in healthcare education », *in*: *BMC Medical Education* 20 (2), DOI: 10.1186/s12909-020-02288-x.

van Waterschoot, Jelte, Merijn Bruijnes, Jan Flokstra, Dennis Reidsma, Daniel Davison, Mariët Theune, and Dirk Heylen (Nov. 2018), « Flipper 2.0: A Pragmatic Dialogue Engine for Embodied Conversational Agents », *in*: *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, IVA '18, Sydney, NSW, Australia: ACM, pp. 43–50, DOI: 10.1145/3267851.3267882.

Vilhjálmsson, Hannes, Nathan Cantelmo, Justine Cassell, Nicolas Chafai, Michael Kipp, Stefan Kopp, Maurizio Mancini, Stacy Marsella, Andrew Marshall, Catherine Pelachaud, Zsófia Ruttkay, Kristinn Thórisson, Herwin Welbergen, and Rick Werf (Sept. 2007), « The Behavior Markup Language: Recent Developments and Challenges », *in*: *Proceedings of the 7th International Conference on Intelligent Virtual Agents*, IVA '07, Paris, France: ACM, pp. 99–111, DOI: 10.1007/978-3-540-74997-4_10.

Vosoughi, Soroush and Deb Roy (May 2016), « Tweet Acts: A Speech Act Classifier for Twitter », *in*: *Proceedings of the 10th AAAI Conference on Weblogs and Social Media*, ICWSM '16, Cologne, Germany: AAAI, pp. 711–4, URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14821.

Walton, Douglas (2007), « The Speech Act Of Clarification In A Dialogue Model », *in*: *Studies in Communication Sciences* 7 (2), pp. 165–97, URL: https://www.researchgate.net/publication/346657878_THE_SPEECH_ACT_OF_CLARIFICATION_IN_A_DIALOGUE_MODEL.

Wang, Isaac, Jesse Smith, and Jaime Ruiz (May 2019), « Exploring Virtual Agents for Augmented Reality », *in*: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, Glasgow, Scotland Uk: ACM, pp. 1–12, DOI: 10.1145/3290605.3300511.

Wang, Rui and Günter Neumann (June 2007), « Recognizing Textual Entailment Using Sentence Similarity based on Dependency Tree Skeletons », *in*: *Proceedings of the*

ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, RTE '07, Prague, Czech Republic: ACL, pp. 36–41, DOI: 10.3115/1654536.1654546.

Wang, Wenya and Sinno Pan (2019), « Syntactically-Meaningful and Transferable Recursive Neural Networks for Aspect and Opinion Extraction », *in*: *Computational Linguistics* 45 (4), pp. 1–32, DOI: 10.1162/COLI_a_00362.

Weisser, Martin (2018), *How to Do Corpus Pragmatics on Pragmatically Annotated Data: Speech Acts and Beyond*, Amsterdam, The Netherlands: John Benjamins Publishing Company, ISBN: 9789027264299.

White, Ralph and Ronald Lippitt (1960), *Autocracy and Democracy: An Experimental Inquiry*, New York, NY, USA: Harper, ISBN: 9780837157108.

Wierzbicka, Anna (1987), *English Speech Act Verbs: A Semantic Dictionary*, Cambridge, MA, USA: Academic Press, ISBN: 9780123128102.

Wilson, Deirdre and Dan Sperber (2012), « Mood and the analysis of non-declarative sentences », *in*: *Meaning and Relevance*, Cambridge, UK: Cambridge University Press, pp. 210–229, DOI: 10.1017/CBO9781139028370.013.

Yalçin, Özge Nilay (2020), « Empathy framework for embodied conversational agents », *in*: *Cognitive Systems Research* 59 (7), pp. 123–32, DOI: https://doi.org/10.1016/j.cogsys.2019.09.016.

Yang, Ping-Jing and Wai-Tat Fu (Oct. 2016), « Mindbot: A Social-Based Medical Virtual Assistant », *in*: *Proceedings of the 2016 International Conference on Healthcare Informatics*, ICHI '16, Chicago, IL, USA: IEEE, p. 319, DOI: 10.1109/ICHI.2016.105.

Yule, Steven, Rhoda Flin, Simon Paterson-Brown, Nikki Maran, and David Rowley (2006), « Development of a rating system for surgeons' non-technical skills », *in*: *Medical Education* 40 (11), pp. 1098–104, DOI: 10.1111/j.1365-2929.2006.02610.x.

Yule, Steven, Rhona Flin, Nikki Maran, David Rowley, George Youngson, and Simon Paterson-Brown (2008), « Surgeons' Non-technical Skills in the Operating Room: Reliability Testing of the NOTSS Behavior Rating System », *in*: *World journal of surgery* 32 (4), pp. 548–56, DOI: 10.1007/s00268-007-9320-z.

Zigarmi, Drea and Taylor Peyton (2017), « A Test of Three Basic Assumptions of the Situational Leadership® II Model and Their Implications for HRD Practitioners », *in*: *European Journal of Training and Development* 41 (3), pp. 241–60, DOI: 10.1108/EJTD-05-2016-0035.

APPENDICES

A Taxonomy for an agent leading a medical procedure

This appendix contains the entire taxonomy of leader behavior. The left-most column contains each required medical coordinator non-technical skill. *Element* describes the subtasks for each non-technical skill. *Communicative intention* is tasks that seek to obtain or provide information from or to followers. *Speech act* refers to the speech act that the agent should use when communicating.

Non-technical skill	**Element**	**Communicative intention**	**Speech act**
Situation Awareness	Updates and reports	gives updates and reports on developments	inform
	Gathering information	requests information	request information
		responds to new information	respond
Decision-making	Selecting and communicating options	communicates selected decision	inform
Task Management	Planning and preparation	communicates plans	inform
	Flexibility/ responding to change	redirects tasks	instruct
	Prioritising	communicates priority of tasks	inform
	Setting and maintaining standards	states standards and expectations	instruct
	Using authority	gives orders	instruct
		states case and provides justification	inform
	Coordinating activities	confirms shared understanding of the group	inform
		ensures caregivers are comfortable with tasks/-capable of completing them	request information
		debriefs followers after procedure	inform
		confirms roles and responsibilities	instruct
		clarifies goals	instruct
		ensures team is working together	instruct
	Identifying and utilizing resources	identifies tools for followers	inform

Team Management	Supporting others	comforts, reassures, encourages	support
		acknowledges concerns of caregivers	support
		thanks team	support
		ensures caregivers are comfortable in general	support
		offers support	offer
		offers suggestions	offer
	Assessing capabilities	intervenes when caregiver does not perform a task to the expected standard	inform
		provides feedback to followers	inform

B Nonverbal leadership behaviors

This appendix displays all the low-level behaviors that should come from different types of leaders. The first column, *Expression Point*, refers to the part of the body where the behavior occurs. The remaining four columns describe purely directive behaviors, purely supportive behaviors, behaviors that all leaders should perform regardless of style, and behaviors that no leader should perform.

Expression Point	**Directive Nonverbal behaviors**	**Supportive Nonverbal behaviors**	**Both Directive and Supportive Nonverbal behaviors**	**Neither Directive or Supportive Nonverbal behaviors**
body		more forward leans	open posture erect posture leaning forward	akimbo posture
eyes	continuous eye contact	wide eyes to support	eye contact	
			gaze towards followers when speaking to them gaze towards objects of importance	
mouth			smile	tense lips
eyebrows				raising eyebrows
head	shake of the head to disagree	more head tilts towards objects (rather than hand gestures)	nodding when listening	tilt upward towards followers
			head tilts toward objects of importance	tilt downward towards followers
hands and arms	ideational gestures	hands together and palms upward to indicate a withdrawal from the task	palms upward	expansive gestures (hands are far from body)
	pointing gestures palms downward to stop or correct a task		ideational gestures steepling of hands	crossing arms hands clasped together
				self-touch

C List of sources for the dataset

The sources used to compile the dataset of medical leader speech are:

— AdvocateHealthCare. (2016, October 22). Trauma Team Crew Training: Advocate Illinois Masonic Medical Center [Video]. YouTube. `https://www.youtube.com/watch?v=t-cMUzMBQQE`. A video simulation depicting team management skills communication and situation awareness in emergency response;

— Goddard, C. (2002). Directive speech acts in Malay (Bahasa Melayu): an ethno-pragmatic perspective. Cahiers de Praxématique, (38), 113-143. A journal which discusses utterances with directive illocutionary force;

— LaheyHealth. (2017, January 9). Simulation Series: Cardiac Arrest [Video]. YouTube. `https://www.youtube.com/watch?v=_wkfOppddDA`. A video simulation depicting the care of a cardiac arrest patient;

— Montgomery College. (2018, March 9). Nursing Simulation Scenario: Opioid Withdrawal [Video]. YouTube. `https://www.youtube.com/watch?v=K4kaB34jSm8`. A video simulation depicting the care of a patient in opioid withdrawal;

— Regina QuAppelle. (2013, November 26). Code Blue [Video]. YouTube. `youtube.com/watch?v=U1zq4T7MEWw`. A video simulation depicting a code blue trauma response.

— ResusCouncilUK. (2017, April 21). RC (UK) ABCDE assessment demo [Video]. YouTube. `https://www.youtube.com/watch?v=KNqoXboSVUI`. A video simulation depicting the assessment of a patient using the ABCDE method;

— ResusCouncilUK. (2017, April 21). RC (UK) Cardiac Arrest Management Demo [Video]. YouTube. `https://www.youtube.com/watch?v=jQYHQr3ebLo`. A video simulation depicting the care of a cardiac arrest patient;

— sparky spacy. (2014, June 13). Primary Survey ATLS Video [Video]. YouTube. `https://www.youtube.com/watch?v=NlYt4rO1B8k`. A video simulation depicting the assessment of a patient using the ABCDE method;

— Swedish. (2016, July 21). Mock Stroke Simulation [Video]. YouTube. `https://www.youtube.com/watch?v=fN7MFkG_4_Q&`. A video simulation depicting the care of a stroke patient;

D Dependency RDF triple definitions

The RDF triples from chapter 5, their definitions, and example sentences are below. These definitions come from Universal Dependencies [1].

— JJ-advmod-RB: indicates an adverb that modifies an adjective (e.g., “infusing here” in “Make sure some normal saline is infusing here.”);

— MD-mark-IN: indicates a preposition or subordinating conjunction marking a clause as subordinate to another clause beginning with a modal verb (e.g., “that you would” in “Put some pressure on that if you would.”);

— NN-case-IN: indicates a noun that acts as a dependent of a preposition or subordinating conjunction (e.g., “like pressure” in “Looks like his blood pressure is quite high.”);

— NN-cop-VBZ: indicates a copula which is the relation of a function word linking a subject to a nonverbal predicate (e.g., “[is] in shock” in “It sounds like an airway problem, sounds like he’s in shock.”);

— VB-aux-MD: indicates a modal verb used with a regular verb (e.g., “can prepare” in “if we can prepare for the administration of 300 milligrams of amiodarone IV push, please.”);

— VB-ccomp-VB: indicates a verb acting as a causal component, a dependent clause, of another verb (e.g., “let’s administer” in “Okay, let’s administer 500 mls of normal saline, run that wide open, and if we can prepare for the administration of 300 milligrams of amiodarone IV push, please.”);

— VB-discourse-UH: indicates an interjection used with a verb phrase where the interjection is not clearly linked to the structure of the sentence, except in an expressive way (e.g., “okay” and “get” in “Okay, can someone get a hold of the family and inform the attending physician as well for me, please?”);

— VB-dobj-NN: indicates a noun that is the direct object of a verb (e.g., “get the heart rate” in “Get the heart rate and the blood pressure.”);

— VB-dobj-PRP: indicates a personal pronoun that is the direct object of a verb (e.g., “get him” in “We need to get him up in the OR.”);

— VB-mark-IN: indicates a preposition or subordinating conjunction marking a clause as subordinate to another clause beginning with a verb (e.g., “if” and “let” in “Denise, if you can let me know when two minutes have passed, please.”);

1. `https://universaldependencies.org/`

— VB-nsubj-PRP: indicates a verb performed by a pronoun which acts as the utterance subject (e.g., “we prepare” in “Can we prepare to change compressors, please?”);
— VBG-aux-VBP: indicates an auxiliary of a clause, which is a word associated with a verbal predicate and expresses tense, mood, aspect, voice or evidentiality (e.g., “[am] feeling” and “[are] going” in “I’m feeling very concerned about the direction we’re going here.”);
— VBN-auxpass-VBZ: indicates the auxiliary form of “to be” used to construct the passive voice or any tense or in the infinitive (e.g., “[has] only got” in “Alright, and Betsy, it sounds like he’s only got one IV, so when he arrives, what I’d like you to do is put in a second IV in on his other side if he doesn’t have one.”);
— VBN-nsubjpass-PRP: indicates a passive nominal subject, which is a noun phrase acting as the syntactic subject of a passive clause (e.g., “he’s got” in “I see that he’s got partial paresis.”).

E Chunking rules

A list of the chunking rules discussed in chapter 5 are below. These rules use POS tags and regular experessions and are implemented in Python. More information regarding the syntax for these rules can be found in the NLTK package reference document [2].

- `When-Phrase: {<,>?<WRB>+<PRP>?<VBP>?<TO>?<DT>?<NN>?}`
- `While-Phrase: {<,>?<IN>+<PRP>?<VBP>?<VBG>?<DT>?<NN>?<NNS>?<,>+}`
- `VB-Phrase: {<DT><,>*<VB>}`
- `VB-Phrase: {<RB><VB>}`
- `VB-Phrase: {<UH><,>*<VB>}`
- `VB-Phrase: {<UH><,><VBP>}`
- `VB-Phrase: {<PRP><VB>}`
- `VB-Phrase: {<RB>+<,>+<VB>+<PDT>*}`
- `VB-Phrase: {<VB.+>*<JJ.?>}`
- `VB-Phrase: {<VBP.+>*<JJ.?>}`
- `VB-Phrase: {<VB.?><DT.?><NN.?>}`
- `VB-Phrase: {<VB>?<VBP>?<PDT>?<DT>?<NNS>?}`
- `Q-Tag: {<,><MD><RB>*<PRP><.>*}`
- `NN-Phrase: {<DT>?<JJ>*<NN>?<NNS>?<NNP>}`
- `If-Phrase: {<,>+<IN>+<PRP>+<MD>?}`
- `If-Phrase: {<IN>+<PRP>+<MD>?}`

2. `https://www.nltk.org/howto/chunk.html`

F User Evaluation 1: Sentences

This appendix contains a complete list of sentences used in the user evaluation and their leadership styles.

	Style	Sentence
1	Directing	I need you to prepare the patient by disinfecting the abdomen.
2	Directing	I need you to take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left.
3	Directing	I want you to prepare the patient by disinfecting the abdomen.
4	Directing	I want you to take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left.
5	Directing	I'd like you to prepare the patient by disinfecting the abdomen.
6	Directing	I'd like you to take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left.
7	Directing	Prepare the patient by disinfecting the abdomen, please.
8	Directing	Take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left, please.
9	Directing	Take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left.
10	Directing	We need to prepare the patient by disinfecting the abdomen.
11	Directing	We need to take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left.
12	Directing	We will prepare the patient by disinfecting the abdomen.
13	Directing	We will take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left.
14	Coaching	Can you prepare the patient by disinfecting the abdomen, please?
15	Coaching	Can you prepare the patient by disinfecting the abdomen?
16	Coaching	Can you take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left, please?
17	Coaching	Can you take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left?
18	Coaching	Could you prepare the patient by disinfecting the abdomen, please?
19	Coaching	Could you prepare the patient by disinfecting the abdomen?
20	Coaching	Could you take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left, please?
21	Coaching	Could you take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left?
22	Coaching	Would you prepare the patient by disinfecting the abdomen, please?
23	Coaching	Would you prepare the patient by disinfecting the abdomen?

24	Coaching	Would you take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left, please?
25	Coaching	Would you take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left?
26	Supporting	Do you need any help in preparing the patient?
27	Supporting	Let me know if you need any help preparing the patient.
28	Delegating	I see that the patient needs to be prepared.
29	Delegating	It looks like the patient needs to be prepared.
30	Delegating	The next step is to prepare the patient.
31	Delegating	The patient needs to be prepared before the procedure begins.
32	Delegating	We are going to begin the procedure soon, and the patient needs to be prepared.
33	Delegating	We are going to prepare the patient.

G Medical Procedure Tasks and Actions

This appendix contains the complete medical procedure: all tasks, actions, and French translations.

Action	Name
	Task 1: Inquiring
1	Ask the patient for their level of pain from 1 to 10. 10 maximum pain. *Demander au patient l'intensité de la douleur de 1 a 10. 10 douleur maximum.*
2	Ask the patient if they are nauseated or have vomited. *Demander au patient s'il a des nausées ou vomissements.*
3	Ask the patient if they have a headache *Demander au patient s'il a des maux de tête.*
4	Ask the patient where their pain is *Demander au patient la localisation de sa douleur.*
5	Ask the patient when their pain started. *Demander au patient a quelle heure la douleur a débuté.*
6	Ask the patient if they have diarrhea. *Demander au patient si il a des diarrhées.*
7	Ask the patient if they have chills or feel feverish. *Demander au patient s'il a des frissons ou s'il se sent fébrile.*
8	Ask the patient if they have any urinary burning *Demander au patient s'il a des brulures urinaires.*
9	Indicate if the patient seems confused or incoherent? *Indiquez si le patient vous semble confus ou incohérent.*
	Task 2: Palpating
1	Palpate the left side and notify the nurse. *Palper le flanc gauche et notifiez a l'infirmier.*
2	Palpate the umbilical area and notify the nurse. *Palper la zone ombilicale et notifiez a l'infirmier.*
3	Palpate the right flank and notify the nurse. *Palper le flanc droit et notifiez à l'infirmier.*
4	Palpate the right hypochondrium and notify. *Palper l'hypocondre droit et notifiez.*
5	Palpate the lower umbilical area and notify. *Palper la zone sus ombilicale et notifiez.*
6	Palpate the left hypochondrium and notify. *Palper l'hypocondre gauche et notifiez.*
	Task 3: Taking blood pressure

1	Take the patient's blood pressure *Mesurer pression arterielle.*
2	Announce the blood pressure which is on the scope.
	Task 4: Measuring oxygen saturation index
1	Measure the oxygen saturation *Mesurer saturation oxygen.*
2	Announce the value of the SPO2 that is on the scope. *Annoncer la valeur de la SPO2 qui se trouve sur le scope.*
	Task 5: Placing the electrodes
1	Place the yellow electrode on the patient. *Placer l'electrode jaune sur le patient.*
2	Place the red electrode on the patient. *Placer l'electrode rouge sur le patient.*
3	Place the green electrode on the patient. *Placer l'electrode verte sur le patient.*
	Task 6: Connecting the electrodes
1	Connect the yellow cable to the yellow electrode at the level of the right shoulder. *Branchez le cable jaune sur l'electrode jaune au niveau de l'épaule droite.*
2	Connect the red cable to the red electrode at the level of the left shoulder. *Brancher le cable rouge sur l'electrode rouge au niveau de l'épaule gauche.*
3	Connect the green cable to the green electrode at the tip of the heart. *Brancher le cable vert sur l'electrode verte a la pointe du coeur.*
4	Read the stability *Lire la constance*
	Task 7: Placing the SPO2 sensor
1	Place the SPO2 sensor on a patient's finger. *Placer le capteur SPO2 sur un doigt du patient.*
	Task 8: Placing the cuff
1	Place the blood pressure cuff on the patient's arm. *Placer le brassard à tension sur le bras du patient.*
2	Turn on the scope by pressing the ON button. It's the white button at the bottom right. *Allumer le scope en appuyant sur le bouton ON. Il s'agit du bouton blanc en bas à droite.*
3	Press the NIBP menu in purple, bottom left on the scope. *Appuyez sur le menu PNI en violet, en bas a gauche sur le scope.*
4	Start the blood pressure measurement by pressing the start stop button. *Lancer la prise de tension en appuyant sur le bouton début arret.*

H Proposed Experimentation

In this appendix, we outline the experiment we propose and would have conducted had Covid-19 not occurred.

The experiment uses a between subjects design where participants complete a medical procedure twice, in French and guided by a virtual agent, through augmented reality. Participants complete questionnaires before the experiment, before and after each procedure, and after the experiment. Participants are recruited through the institution (ENIB).

H.1 Study design

Individual participants are welcomed into a room containing the human dummy and the AR headset. The organizer explains to the participant that:

1. They will be completing a medical procedure on the dummy patient and that a virtual assistant will be guiding them through the procedure.
2. They are able to ask the assistant two questions: one, what is the action they have to do and two, what is the resource they have to use.
3. A nurse will be standing by to note the information from the procedure.
4. They will be asked to fill out a questionnaire before each procedure and afterward in regards to their own confidence and ability as well as their perceptions of the agent personally and the agent as a medical procedure leader.

Participants are then asked to complete the pre-experiment questionnaire and the first pre-procedure questionnaire. They then complete the procedure the first time (detailed below). After the procedure is finished, they complete the first post-procedure questionnaire and the second pre-procedure questionnaire. Afterward, they complete the procedure for a second time. After the second procedure is finished, they complete the second post-procedure questionnaire and the post-experiment questionnaire.

The procedure involves diagnosing a patient's pain and preparing for abdominal surgery

Participants are randomly placed in four groups:

1. Group 1: fixed leadership style S1 (directing) during procedure 1 and fixed leadership style S3 (supporting) during procedure 2;
2. Group 2: fixed leadership style S1 (directing) during procedure 1 and fixed leadership style S1 (directing) during procedure 2;

3. Group 3: fixed leadership style S3 (supporting) during procedure 1 and fixed leadership style S3 (supporting) during procedure 2;
4. Group 4: fixed leadership style S3 (supporting) during procedure 1 and fixed leadership style S1 (directing) during procedure 2.

When leadership style S1 is assigned, the agent communicates every action that needs to be done as well as addresses any errors that the participant makes with detail. The agent's orders are constructed from random indices from a series of lists which specify the grammatical mood and keywords that can be used.

When leadership style S3 is assigned, the agent communicates only when an error has been made and provides fewer details regarding each error.

When the participant makes an error, the nurse makes note of it. However, the nurse does provide any assistance to the participant.

H.2 Pre-experiment questionnaire

Before the experiment begins, participants are asked to give their age, gender, their medical experience level, and their AR experience level.

H.3 Post-procedure questionnaire

After each procedure, participants are asked to rate a number of items on a 5-point Likert scale:

1. The agent gave me enough information to complete the procedure
2. The agent gave me more information than I needed to complete the procedure
3. I could have done the procedure without the agent's help
4. The agent motivated me during the procedure
5. The agent is helpful
6. The agent is intelligent
7. The agent has all the medical knowledge necessary
8. I like the agent

These questions are adapted from an existing PhD thesis (Krishna 2021), the IPQ-R questionnaire (Trapnell and Broughton 2006), and from previous work utilizing surveys (Ryu and A. Baylor 2005). The questions aim to identify the participants' attitudes towards the agent as well as understand whether the agent's assistance was helpful or not.

H.4 Post-experiment questionnaire

After the second post-procedure questionnaire is filled out, participants are asked whether they noticed a difference in the behavior of the agent between the two procedures and to leave any further comments regarding their experience.

H.5 Hypotheses

1. When leadership style matches readiness level, participants find the agent more helpful, intelligent, and likeable, and participants perform fewer errors.
2. When leadership style is higher than readiness level, participants find the agent less helpful, intelligent, and likeable, and participants perform more errors.
3. When leadership style is lower than readiness level, participants find the agent less helpful, intelligent, and likeable, yet they perform fewer errors.

DOCTORAT BRETAGNE LOIRE / MATHSTIC

ÉCOLE NATIONALE D'INGÉNIEURS DE BREST

Titre : Un modèle d'un assistant de leader virtuel humanoïde pour un soignant en milieu éloigné et isolé

Mot clés : agents conversationnels animés, procédures médicales et assistance médicale, fiabilité et confiance dans le système informatique, leadership situationnel, comportement verbal et non verbal

Résumé : Lors d'une urgence médicale en un lieu isolé tel que Mars dans lequel un soignant est séparé des experts médicaux par l'espace et le temps, un assistant virtuel est nécessaire afin de guider avec succès les soignants expérimentés et inexpérimentés tout au long de la procédure médicale. Une procédure réussie implique (1) l'amélioration de la santé du patient et (2) le maintien d'une relation de confiance entre le soignant et l'assistant virtuel. Afin de gérer la relation entre l'assistant et le soignant, nous proposons un ECA dont le fonctionnement s'appuie sur la théorie de leadership situationnel. Cet agent doit être en mesure d'adopter un comportement multimodal intégrant des expressions verbales adéquates ainsi qu'un comportement non verbal associé. Pour cela, nous créons et annotons un ensemble de données de discours de leaders médicaux et analysons cet ensemble afin d'extraire une série de règles linguistiques pour le discours dédié au leadership situationnel. Ces règles sont ensuite validées par l'éxpérimentation. Nous proposons également un modèle pour analyser le comportement du soignant pendant la procédure, déterminer le style de leadership de l'agent le plus approprié en fonction du comportement du soignant et de l'état du patient, et générer un comportement d'agent approprié pour faire progresser la procédure et maintenir une relation de confiance avec le soignant.

Title: Modelling a humanoid virtual leader assistant for a remote and isolated caregiver

Keywords: embodied conversational agents, medical procedures and assistance data, system trustworthiness, Situational Leadership, verbal and nonverbal behavior

Abstract: In a medical emergency on an isolated point such as Mars in which a caregiver is separated from medical experts by time and space, a virtual assistant is necessary in order to guide both experienced and inexperienced caregivers through the medical procedure successfully. A successful procedure involves (1) improvement of the patient's health and (2) maintenance of a trusting relationship between the caregiver and the virtual assistant. In order to manage the relationship between the assistant and caregiver, we propose an ECA that employs Situational Leadership. We identify multimodal behavior, both nonverbal behavior and speech, that an agent should employ. We create and annotate a dataset of medical leader speech and analyze it with statistical analysis and k-means clustering, resulting in a series of linguistic rules for speech in Situational Leadership that are then validated through experimentation. We also propose an agent system for analysing caregiver behavior during the procedure, determining the most appropriate agent leadership style based on caregiver behavior and the patient state, and generating appropriate agent behavior to progress the procedure and maintain a trusting relationship with the caregiver.