PROPOSED RESEARCH:

MULTIMODAL ARGUMENTATION MINING IN GROUPS ASSISTED BY AN EMBODIED CONVERSATIONAL AGENT

A. Theoretical-conceptual foundations and state-of-the-art that support the proposal

Interactions and Multimodality are crucial in the development of intelligent AI models that can understand human-like communication. Human learning occurs through interactions with the environment and other humans, which involves the integration of information from multiple modalities such as vision, language but also touch and hearing that enable us to understand the subtle social meanings behind communication.

Therefore, to create intelligent machines that can understand human non-verbal communication, it is essential to train them on **multimodal interactions that mimic those of humans to ensure that they can understand and respond appropriately to complex social phenomena**.

The recent computational boom has seen the emergence of seminal studies focusing on Multimodal data (Cho, Lu, Schwenk, Hajishirzi, & Kembhavi, 2020; Hasan et al., 2019; Jaegle et al., 2021; J. Li, Li, Xiong, & Hoi, 2022; J. Wang et al., 2022; Zadeh, Chan, Liang, Tong, & Morency, 2019) and Interactions, whether these ones are textual like OpenIA's InstructGPT or Anthropic's Claude (Bai et al., 2022; Ouyang et al., 2022; Schulman et al., 2022), or multimodal like Google's PaLM (Chowdhery et al., n.d.; Chung et al., 2022; Schick, Lomeli, Dwivedi-yu, & Dessì, 2022) or GPT-4 (Bubeck et al., 2023; OpenAI, 2023; Wu et al., 2023).

These advancements show the potential for machines to learn from multimodal interactions and understand human communication, which could revolutionize the way humans socially interact with machines in the future. **Nevertheless, nowadays generative agents are restraint to unimodal data or not using the full time-series of every modality of a real human-machine social interaction**.

Interaction and multimodality are vital contexts in many social situations. They are also mandatory to make a machine understand the world and get commonsense knowledge, which is essential when tackling human-related complex tasks. Indeed, **humans are social animals** and they interact with one another. In a general way, the integration of more context is the key to a deep understanding of many phenomena, in order to disambiguate a situation or to reinforce the current estimation: interaction is a crucial context in many social situations. Multimodal interactions allow understanding in a deeper way human behavior. In this particular setting, it is possible to understand a broader part of the multimodal natural language (see Figure 1). Studying the affective and social **phenomena like Opinions, Emotions, Empathy, Distress, Stances, Persuasiveness or speaker traits allows to greatly improves the response from the machine** (Pelachaud, Busso, & Heylen, 2021; Zhao, Sinha, Black, & Cassell, 2016), but this task is difficult even using multimodal data. My research focuses on designing and developing methods that integrate the multimodal context and how humans influence each other in discussion situations. The research goals of this project fall into this general research area: **how to use interactions and multimodality of non-verbal language to enhance social AI systems**.

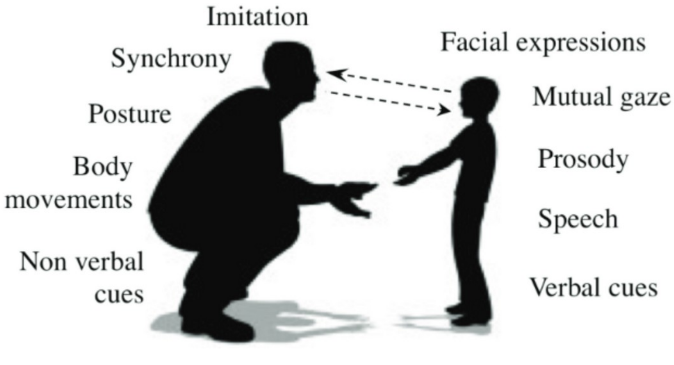


***Figure 1**: Examples of non-verbal language involved in a social interaction*

Multimodality:
Communication is not just limited to language, and it is essential to consider other modalities such as vision or audio when building natural language processing (NLP) systems (Baltrušaitis, Ahuja, & Morency, 2017; Liang, Zadeh, & Morency, 2022).
Incorporating multiple modalities, or multimodality, is critical in creating more human-like interactions between humans and machines. For instance, while language is the primary means of communication for humans, it is often supplemented by visual and auditory cues such as facial expressions, tone of voice, and gestures. Therefore, it is important building multimodal machine learning systems that can interpret and respond to these cues in a human-like manner.

According to (Fröhlich, Sievers, Townsend, Gruber, & van Schaik, 2019), both human and non-human primate communication is inherently multimodal. As an example, (Mehrabian, 1971) even states that 55% of the emotional content is in the visual signal (facial expressions and body language), 38% in the vocal signal (intonation and sound of the voice) and 7% in the verbal signal (through the meaning of the words and the arrangement of the sentence).

Interactions dynamics:
It is essential to consider the interactive nature of human communication and incorporate it into natural language processing (NLP) systems. By allowing the machine to understand the context and flow of the conversation, it can provide a more natural and seamless interaction with users (Sutskever, Vinyals, & Le, 2014). (Z. Li, Wallace, Shen, & Lin, 2020) suggested that these systems can provide tailored content and services based on the user's interests and preferences, leading to more engaging and personalized interactions with the user. **As humans, we are not learnig by looking at or enviroment, but by interacting with it and with our peers.** By considering the interactive nature of human communication and incorporating it into NLP systems, machines can learn to communicate in a way that is more similar to humans, making interactions more engaging and effective.

Proposed research project:
This research project aims at studying the complex phenomena characterizing social interactions between humans using different media, implying different modalities and data domains. My research objective is to design adaptive models that take as a starting point the specificities of the multimodal interaction: the media used to communicate, the interactants' social relationship, and the communication modalities used to transfer the information. **The general goals stand to: understand what the users are trying to achieve as a group, what is the output of this interaction, how a social agent helps reaching it**.

***Figure 2**: A user interacting with an Embodied Conversational Agent*

In particlar, this project aims to explore the dynamics of how a group of individuals with polarized opinions can reach a consensus. In this work, within groups of individuals debating hot societal topics and issues, the aim will be to automatically detect and retrieve stances and arguments towards the debate question and to ultimately moderate the debate using a human-computer interface that would be specific to such an interaction. To this aim, we think that an **Embodied Conversational Agent** (Cassell, 2001; Pelachaud, 2005) like the one illustrated in Figure 2, would be the most relevant. Indeed bodily representations structure the way humans perceive the world and the way they perceive other people. Cognitive sciences and social sciences altogether have stressed **the importance of embodiment in social interaction, highlighting how interacting with others influences how we behave, perceive and think** (Smith & Neff, 2018; Tieri, Morone, Paolucci, & Iosa, 2018), including our social behaviors with embodied intelligent agents such as virtual humans and robots (Holz, Dragone, & O'Hare, 2009).

Another goal is to explore the polarization of society's attitudes towards hot political topics and study the **difference in terms of the difficulty of finding a consensus** regarding the type of topics, and the human values involved in classical argumentation (Kiesel, Weimar, Handke, & Weimar, 2022; Mirzakhmedova et al., 2023). In today's society, the polarization of opinions on political topics is a common phenomenon that can be observed in many different areas. Debates about societal topics and issues can be especially polarizing and lead to a lack of understanding and cooperation between groups with different perspectives (Livingstone, Fernández Rodriguez, & Rothers, 2020). Therefore, **it is crucial to understand how individuals with polarized opinions can reach a consensus**, and this is the aim of this research project. To achieve it, this project plans to develop an automatic approach to detect and retrieve the stance and arguments of individuals involved in real-time multimodal debates about hot societal topics.
This research aims to delve into the complexities of group dynamics in polarized debates on societal issues. To achieve this, we will not only automatically **detect and retrieve stances and their arguments** toward the debate question, but also take into account the multimodal aspects of the debate, such as **body language, facial expressions and acoustics**, which are shown to be important for persuasion in a Vlog (Nojavanasghari, Gopinath, Koushik, Baltrušaitis, & Morency, 2016; S. Park, Shim, Chatterjee, Sagae, & Morency, 2014; Siddiquie, Chisholm, & Divakaran, 2015) or within a debate (Brilman & Scherer, 2015; Mestre et al., 2021). Real-time interaction within the group will be analyzed to understand **how individuals respond to each other and how the group as whole moves toward a consensus**.

B. Hypothesis or research questions and objectives

By studying the nuances of group dynamics in moderated polarized debates, this research aims to contribute to a better understanding of **how to promote cooperation between groups with different opinions** to foster productive and constructive discussions on complex societal issues.

In the end, the research can contribute to the development of more effective communication tools that can facilitate discussions and consensus-building on various societal issues. The insights gained from this research can also inform policymakers and other stakeholders about the importance of promoting constructive debates and finding common ground on polarizing topics.

Argumentation mining has been widely studied in the context of textual debates (Mochales & Moens, 2011; Palau & Moens, 2009), whether they are political or not (Barriere, Balahur, & Ravenet, 2022; Lai et al., 2020; Lai, Patti, Ruffo, & Rosso, 2018; Mohammad, Kiritchenko, Sobhani, Zhu, & Cherry, 2016). Moreover, participants can change their opinions and the intensity and valence of their stance during the debate (Sakketou, Lahnala, Vogel, & Flek, 2022), and the role of such social phenomena as persuasion have been found important (Walker, Anand, Abbott, & Grant, 2012a). Finally, **multimodality simply helps the argumentation mining tool to reach higher performances** (Mestre et al., 2021).

Furthermore, it has been found that a virtual agent that understands the arguments is helpful to find a consensus between different opposed parties (Argyle et al., 2023). For this reason, I plan to incorporate an Embodied Conversational Agent (ECA) as a human-computer interface to moderate debates between individuals with different opinions. ECA (Cassell, 2001; Pelachaud, 2005) have been used for several applications such as social coach for job interviews (Anderson et al., 2013; Gratch et al., 2014) or as a healthcare decision support tool (DeVault et al., 2014; Lucas, Gratch, King, & Morency, 2014). **ECA can affect user impression and engagement** (Biancardi, Dermouche, & Pelachaud, 2021), **and arguments** (Kantharaju et al., 2019). The ECA, which will have a greater influence on humans than a computer (André et al., 2011; Vossen, Ham, & Midden, 2009), will serve as an intermediary, guiding the debate in a constructive direction and **helping participants to understand each other's perspectives, by the mean of social skills like Empathy** (Parmar, Olafsson, Utami, Murali, & Bickmore, 2022; Reis, Lemay Jr, & Finkenauer, 2017).

Our project builds on the understanding that **debating is a persuasive behavior using verbal and non-verbal language** and aims to quantify this behavior using literature from psychology and lingiustics (Morio, Egawa, & Fujita, 2019). We have chosen debates as a context for our investigation because debates offer a unique opportunity to study how people use language and non-verbal cues to persuade and influence each other. Our work will leverage recent advances in multimodal signal processing and machine learning to analyze and model these interactions, as well as the impact of multimodal natural language on persuasive communication and empathy between participants. By focusing on non-verbal cues and the role of moderators, we aim to gain insights into how debates can be made more productive and how group dynamics can be improved. We expect the results of this project to have broad implications for the design of interactive systems that rely on natural language processing and multimodal signal processing.

For this, the debates have to follow a certain number of rules: (i) they must not be artificial, which means the participants should defend their own opinion to **allow opinion shifting** (Sakketou et al., 2022), (ii) they must contain **pure natural interactions** and not turns of monologues like in an Oxford-style debate in order to allow natural interruptions (Yang et al., 2022), (iii) they must be multi-party to **stimulate the exchange of ideas** from different speakers, and to use group effects and cohesion to reach a consensus (Kantharaju, Langlet, Barange, Clavel, & Pelachaud, 2020).

Hypotheses:

1. **Debating is persuading**. Humans are using verbal and non-verbal behavior that is related to persuasive behavior during debates. It is possible to quantify it using literature from psychology.

2. **The media used to communicate is important**. Humans are more prone to change their stance towards a specific topic if they are in front of other humans instead of in front of a computer.

3. **Listen to persuade**. A social agent can help to find a consensus between participants by asking for disambiguation and reformulating their arguments so that they fit on common ground.

Objectives:

1. **Advance research on the understanding of argumentation mining**, by adding the non-verbal dimension. We will enhance the classical systems by fusing efficiently different modalities such as acoustic, facial expressions, gestures, and gaze.

2. **Create a new dataset for research purposes** of citizens debating hot political issues. The dataset will be annotated for each debate in terms of stance and arguments. An ontology will be created to create a hierarchy of the arguments in the discussion, authenticating which one is answering to which one.

3. **Investigate empirically the use of an embodied virtual agent** to moderate a debate between humans in order to find a consensus. The study will be directed to analyze the impact of an ECA compared to a text-based agent like chatGPT.

This project relies heavily on **non-verbal language such as acoustic, facial expression and gesture analysis, but also on text-based argumentation mining in political debates, which makes it inter-disciplinary**. Because of the nature of the research questions, the results of this research project will have an impact on each of the entangled scientific areas implied in the project, and not only on one discipline. Each of the disciplines has a high impact on the proposed research and they are fundamentally linked as they influence one-another in the methodology and in the analysis of the results.

C. Scientific or technological novelty of your proposal

Multimodal argumentation mining is a field of research that aims to automatically analyze and understand arguments in multi-party discussions. As far as the author knows, there are some gaps in the literature. The proposed scientific work in this project is highly innovative and aims to address a gap in the field of multimodal argumentation mining. Firstly, while there has been some research on fine-grained argumentation mining and consensus finding, to the best of our knowledge, other approaches are restrained to bi-modal studies with text and audio (Lippi & Torroni, 2016; Mancini, Ruggeri, Galassi, & Torroni, 2022; Mestre et al., 2021). The few works using also visual non-verbal language do not tend to do a fine-grained analysis of the arguments (Brilman & Scherer, 2015). Furthermore, no work has attempted to integrate an Embodied Conversational Agent (ECA) as a moderator in group debates while also focusing on non-verbal-augmented systems. Research on persuasive communication and empathy has shown that non-verbal cues, such as facial expressions and body language, play a crucial role in shaping human communication (Mehrabian, 1971). However, much of the work in this area has focused on analyzing these cues in a passive manner.

Hence, current research in this area is still limited, and **there is a lack of studies on how to effectively mine multimodal argumentation and integrate ECA as moderators within a group debate**. In what follows, we describe the distinct ways by which non-verbal cues can be leveraged in debates, being use by a social agent and subsequently situate the uniqueness of our approach within this field. We propose three work packages to address this gap in the literature.

In Work Package 1[1], we will conduct a state-of-the-art review and collect a multimodal corpus of debates to be annotated by expert annotators. The corpus will include audio and video data, as well as transcripts of the debates. This corpus will serve as the basis for the subsequent work packages, providing a rich source of data for analysis.

Multimdodal datasets are numerous for the study of emotions or affects (Busso et al., 2017; Lotfian & Busso, 2018; McKeown, Valstar, Cowie, Pantic, & Schröder, 2012) but **not suitable for argumentation mining for political issues**. Debates datasets are then mainly public debates by experts, like politics in a dyad (Bousmalis, Morency, & Pantic, 2011; Mestre et al., 2021; Vinciarelli, Dielmann, Favre, & Salamin, 2009) or in a team in an Oxford-style debate (Brilman & Scherer, 2015). The latter are semi-structured debates, forbidding interruptions between participants, which are very important in natural multimodal communications (Yang et al., 2022). They can even takes the form of monologues (Sen et al., 2021). Another important wanted particularity is a dataset of debates with real opinions and not artificial ones in order to allow opinion shifts and consensus finding, removing many datasets where the participants are giving a side and need to win (Petukhova et al., 2018).

On the annotation side, there are datasets of multi-party interactions with non-experts having annotations of the speaker performance at the debate level (Shiota & Shimada, 2020) but **we want one at a more fine-grained scale**, allowing for more complex phenomena detection. We can take inspiration from several existing works for the annotation scheme: (Olshefski, Lugini, Singh, Litman, & Godley, 2020) are proposing an annotation scheme containing fine-grained collaborative-related behaviors annotation, making this work close to what we would like to achieve, but **they are working on English literature debates only, and not using video**. Also interesting is the work of (Nihei, Nakano, Hayashi, Huang, & Okada, 2014) that proposes to annotate the statements that will influence the discussion flow, which is very interesting for opinion shifting. Another interesting work regarding what we would like to achieve is the one of (Sakketou et al., 2022) which studies the dynamics evolution of the stance of the participants, finding moments of opinion shift inside a textual interaction.

Because multimodal argument mining can be limited in terms of complexity (Zhou et al., 2013) or in terms of modality (Bondarenko et al., 2023), the work we are proposing in WP1 is highly innovative.

In Work Package 2, we will focus on the modelization of the argumentation and consensus dynamics in a group, which will enable us to better understand the role that non-verbal cues play in shaping the outcomes of debates. Past works used machine learning algorithms to analyze the corpus data and develop models that can detect patterns in non-verbal data, but **they restrained to a more simple task like debate winner or agreement detection and not argumentation mining** (Bousmalis et al., 2011; Brilman & Scherer, 2015), some using natural interactions (Nihei et al., 2014; Shiota & Shimada, 2020) or structured interactions (Petukhova, Mayer, Malchanau, & Bunt, 2017; Petukhova, Raju, & Bunt, 2017) or even a structured debate without interaction (Hasan et al., 2021). Other works on audio and text focus on more granular annotations but no video (Haddadan, Cabrio, & Villata, 2020; Lippi & Torroni, 2016; Mancini et al., 2022; Mestre et al., 2021). In a general way, the major parts of the articles are focused on how to win a debate or how to convince the public. **They do not refer to how to convince the opposite debate partners, which will be a major point in our research**. Here we will be the first work proposing to tackle multimodal argument mining for consensus finding by focusing on the modeling of argumentation and consensus dynamics in a group. This will enable us to

[1] Work packages are described more in details in the next section

better understand the role that non-verbal cues play in shaping the outcomes of debates. For the subseqent reasons, the WP2 is highly innovative.

In Work Package 3, we will integrate an ECA as a moderator within the debates. The ECA will be designed to recognize and respond to non-verbal cues in real-time, while also monitoring the group's discussion and providing prompts to facilitate productive debate. The ECA will be implemented using state-of-the-art natural language processing and computer vision techniques and will be evaluated through a series of experiments using the multimodal corpus collected in Work Package 1. The works on the impact of an ECA over persuasion (Kantharaju, Pease, De Franco, & Pelachaud, 2018) or argumentation (Kantharaju et al., 2019) has been very limited. Some use a persuasive Agent that can debate using non-verbal language like characteristic gestures and mimic to convince their opponent (Rach et al., 2018), however they use the agent in a very constrained context, without communicating directly through natural language but in the form of a game using a Wizard-of-Oz setting. (Petukhova, Mayer, et al., 2017) used an ECA as a coach to train to debate better, using a set of prosodic, motion, linguistic and structural features. Then, (Petukhova, Raju, et al., 2017) integrated the non-verbal behaviors in an ECA, by analyzing the correlations between different non-verbal features and the debate-level performance using the Metalogue Data Trainee Corpus from (Petukhova et al., 2018). Nevertheless, they only extracted features by comparing the behaviors of non-experts and politicians, without focusing on a collaborative goal. (Weber, Rach, Minker, & André, 2020) created two virtual agents to present the pros and cons of a controversial topic, where the agents adapted their emotional reaction to the user's feedback to increase the perceived persuasiveness. As far as the author knows, **the successful use of a virtual agent to help the participants in a live debate in order to find a consensus has been limited to text-only (Argyle et al., 2023),** making the WP3 an innovative contribution.

By completing these three work packages, we aim to contribute to the development of more effective multimodal argumentation mining techniques and the integration of ECAs as moderators in group debates. Ultimately, this research has the potential to enhance our understanding of how to facilitate productive and inclusive discussions. The interdisciplinary nature of our proposal brings together expertise in NLP, computer science, psychology, and linguistics, and we anticipate that our results will have implications for improving communication and decision-making in various domains.

Academic impact: In terms of academic impact, our project has multiple objectives that go beyond the development of a new approach for multimodal argumentation mining. **We plan to contribute to the training of the next generation of NLP researchers** by supervising and supporting the development of at least two MSc and a PhD these.

In addition, we aim to disseminate our research through **the publication of at least 3 papers in high-impact NLP journals and conferences**. Our first journal paper will focus on a review of the state-of-the-art of argumentation mining and multimodal machine learning for social phenomena.

Our first conference paper will focus on the dataset of our novel approach to multimodal argumentation mining and will be submitted to a prestigious Linguistics Resources conference like LREC. This paper will present our experimental paradigm, data collection and annotation, as well as an analysis of it. Simple computational models that serve as a baseline for our approach. Our second journal paper will delve deeper into the NLP aspects of our project by exploring the language modeling techniques we used to extract persuasive language cues from the multimodal data. We aim to submit this paper to a top-tier NLP conference such as ACL or EMNLP. Our third publication will focus on the integration of an Embodied Conversational Agent as a moderator in group debates. This paper will explore the computational challenges we encountered in developing this component and present the results of our evaluations. This paper will be submitted to a **leading NLP journal such as Transactions of the Association for Computational Linguistics (TACL) or Transactions of Affective Computing (TAC)**.

Furthermore, we believe that the neural-driven learning method we will develop for our project will have broader applications in the field of NLP. We anticipate that this approach could be adapted to improve natural language understanding and generation in other domains. As such, we plan to publish another paper in rank A NLP conferences such as ACL or EMNLP (best venues to publish, journals included, in NLP area) to share our findings and contribute to the advancement of the field.

Collaborators: The extensive work of Dr. **Alexandra Balahur** (European Commission's JRC-Bruxelles) on methods and applications for automatic affect and stance detection from text, as well as applications of AI to societal challenges and the impact of AI on humans and society affords a good collaboration for this project.

Prof. **Roman Klinger** (University of Stuttgart) has world-level expertise in modeling psychological concepts in language which would significantly enhance the scope of the proposed project. Moreover, his work on sentiment and emotion analysis, as well as on generation of conditioned text based on emotions and styles could prove useful in enhancing the ECA verbal skills. Roman

A partnership with Dr. **Catherine Pelachaud** (ISIR, Sorbonnes Université) and Prof. **Brian Ravenet** (LISN, Université Paris Saclay) would be highly beneficial since both having vast expertise in the field. By leveraging their extensive knowledge of ECA, as well as psychologically-grounded multimodal social interactions the project will reach its fullest potential.

D. Methodology

In what follows, we outline the three Work Packages (WPs) that form the backbone of our research project, aimed at advancing the field of multimodal argumentation mining. Each WP is designed to address specific research questions and objectives, and their combination highlights the interdisciplinary nature of our project. WP1 involves conducting a state-of-the-art review, collecting multimodal debate data, and annotating it for various features related to argumentation. This WP is grounded in experimental psychology and cognitive neuroscience. WP2 focuses on developing a model for argumentation and consensus dynamics in groups, with a particular emphasis on non-verbal cues. This WP lies at the intersection of computational neuroscience and multimodal processing. Finally, WP3 centers around the integration of an Embodied Conversational Agent (ECA) as a moderator in group debates, leveraging AI and natural language processing. In addition, we have identified potential risks associated with each WP and developed mitigation plans accordingly. We also highlight national and international collaborations associated with each WP that will contribute to achieving our research goals.

Work Package 1: State-of-the-art Review, Multimodal Debates Collection and Annotation

The first phase of the project will be dedicated to a study of state-of-the-art on text-based argumentation mining and multimodal machine learning for social phenomena detection.

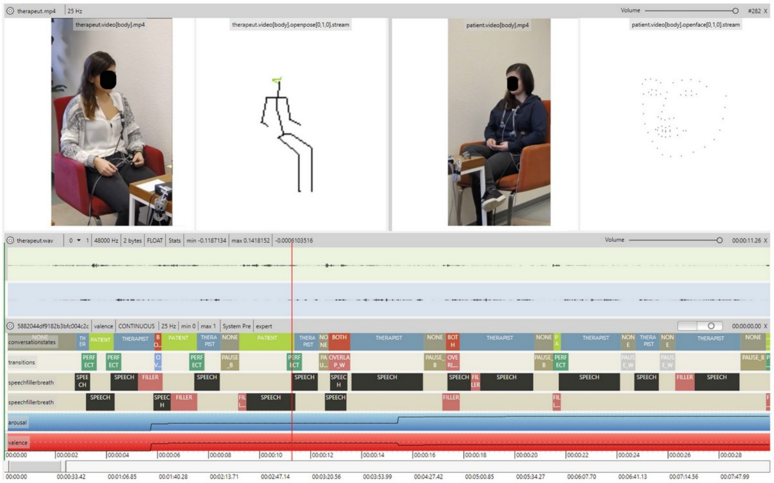
We will study the different models used for stance recognition, argument retrieval and argument hierarchization (Abbott et al., 2011; Barriere et al., 2022; Bondarenko et al., 2023; Hardalov, Arora, Nakov, & Augenstein, 2021; Swanson, Ecker, & Walker, 2015), trying to focus on the models extracting arguments from a discussion (Abbott et al., 2011; Kiesel, Spina, Wachsmuth, & Stein, 2021) how to find a deliberation (Argyle et al., 2023; Walker, Anand, Tree, Abbott, & King, 2012) and the links with other social phenomena like persuasion (Walker, Anand, Abbott, & Grant, 2012b).

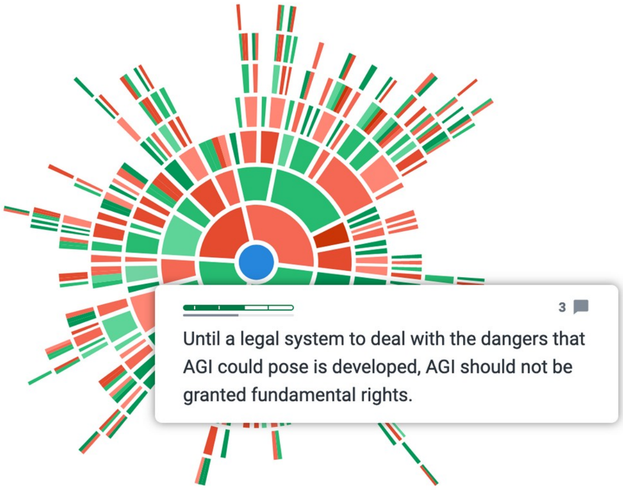
The study will also explore the different **models used to process multimodal natural language**, that takes the form of multimodal time-series comprising such modalities as verbal, acoustic, facial expressions or gestures for **affect recognition** (Lyu, Liang, Deng, Salakhutdinov, & Morency, 2022; Zadeh et al., 2018), humor (Hasan et al., 2019) or **social IQ** (Zadeh et al., 2019), **persuasiveness** (Nojavanasghari et al., 2016) and any **multimodal Natural Language Understanding** task (Liang, Lyu, et al., 2022).

Several past datasets using multimodal data for the study of different phenomena such as expertise and first impressions (Cafaro et al., 2017), emotions and collaborations (Nojavanasghar, Baltrušaitis, Hughes, & Morency, 2016; Ringeval, Sonderegger, Sauer, & Lalanne, 2013), persuasion (S. Park et al., 2014) will be reviewed. We will use the same process for annotation tools for multimodal data such as **NOVA** depicted in Fig. 3a (Baur et al., 2020; Heimerl, Baur, Florian Lingenfelser, Wagner, & André, 2019) or others (Biancardi, Ceccaldi, Clavel, Chollet, & Dinkar, 2021).

For simplicity reasons, the **dataset will be collected online using video-conference**, in order to get the same type of interactions as in RECOLA (Ringeval et al., 2013) . We will focus on several languages (English, Spanish and French) to promote diversity.

For the data collection, the participants will be recruited as volunteers, willing to spend 20mn maximum debating on one specific topic. **We will follow the protocol of (C. Y. Park et al., 2020)**, which create a dataset of dyadic debates in English on a social hot topic. We will send the participants an email with articles with different opinions about the topic of the debate to familiarize and see different arguments and points of view before the debate. Once they receive the emails, they will choose the topics they want to discuss within a list and give their apriori stance over the debate question. We will use debates from the Kialo website[2] that already contain a lot of debates and arguments (see an example Fig. 3b).





Until a legal system to deal with the dangers that AGI could pose is developed, AGI should not be granted fundamental rights.

Figure 3(a)*: Screenshot of the Nova annotation tool for social interacions, which allows to visualise various media and signal types and supports different annotation schemes. Full-body videos along with skeleton and face tracking, and audio streams*
Figure 3(b)*: Kialo's debate topology for "Should general AI have fundamental rights?". Rings are representing various levels of arguments surrounding, the color indicates the stance (pro or against). Each segment is an argument that can be hovered over to reveal the related premises and their respective impacts determined by the Kialo community through votes.*

We will create diverse groups that are representing the diversity of opinions on the whole topic to create a discussion group. They will converse with each other using a web platform for video-conference. The participants will be able to have the debate at home, or in the offices of the partner institutions.

For the data annotation, we will assess precisely the different annotation scheme present in the litterature in order to find the most adapted for consensus finding and collaborative argumentation. This will include several works already mentioned in this document (Nihei et al., 2014; Olshefski et al., 2020; Sakketou et al., 2022; Shiota & Shimada, 2020). We will choose **NOVA** as annotation tool (see Fig. 3a).

Work Package 2: Modelisation of the argumentation and consensus dynamics in a group, focusing on the training of non-verbal-augmented models.

The scope of this work package will be to create a model able to handle multimodal interaction data by detecting multimodal patterns that can be distributed over the different speakers, and that would be characteristic of argumentation, collaborative or not.

To begin, the project will draw on previous research, by analyzing the data in a classical way to find significant cues for the different types of behaviors related to argumentation. We will follow the process of (Brilman & Scherer, 2015) to identify key non-verbal cues that can influence argumentation and consensus in group settings.

We will use several state-of-the-art methods to model multimodal natural language. Those networks can model time series, by identifying the contributions of each modality, how they relate with each other, and how to compose them to create a multimodal prediction. We will start from a neural network able to represent the unimodal and multimodal information at the same time for complex temporal and multimodal pattern detection (Zadeh et al., 2018), like the method showed in Figure 4, to a Transformer-based network that integrates multimodal time-series (Rahman et al., 2020; Tsai et al., 2019). We plan to add more modalities, like gestures and gaze, that have been proven important for speech quality assessment or job interviews (Ohba, Mawalim, Katada, Kuroki, & Okada, 2022). **In a few words, joint modeling of multimodal data: cross-modality attention and multimodal fusion.**

[2] https://www.kialo.com/should-general-ai-have-fundamental-rights-6295

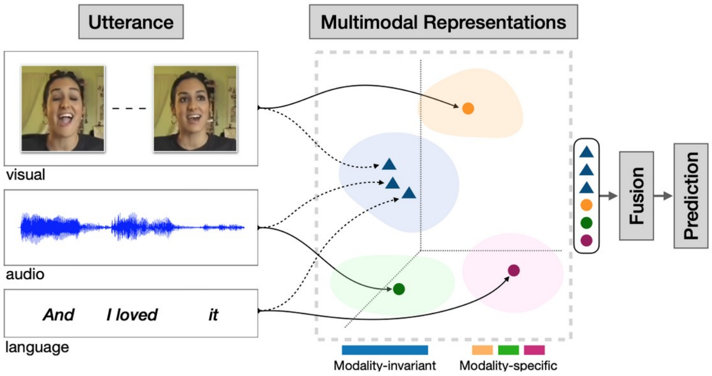


Figure 4: A modelization of multimodal time-series using uni- and multi-modal representations from CMU-MOSEI

The project will then **leverage multimodal datasets** without interactions, like the DBATES (Sen et al., 2021), CMU-MOSEI and CMU-MOSEAS (Zadeh et al., 2020) in order to pre-train the model on similar data and related tasks. These datasets contain videos and audio recordings of monologues that are annotated in debate performance, persuasion, emotion, or other social phenomena.

The mentioned networks are good at multimodal natural language understanding with time-series data from one individual (Liang, Lyu, et al., 2022), but they do not model the group interaction. Hence, **we will create a new type of network inspired by the DialogueRNN of (Majumder et al., 2019) to represent the different actors in a multi-party interaction** (see Figure 5). Coupling the multimodal encoder to an interaction-dedicated module will ensure to get a model able to detect complex patterns in multimodal interactional data. We plan to use semi-supervised training objectives to pre-train our network (Lian, Liu, & Tao, 2022).

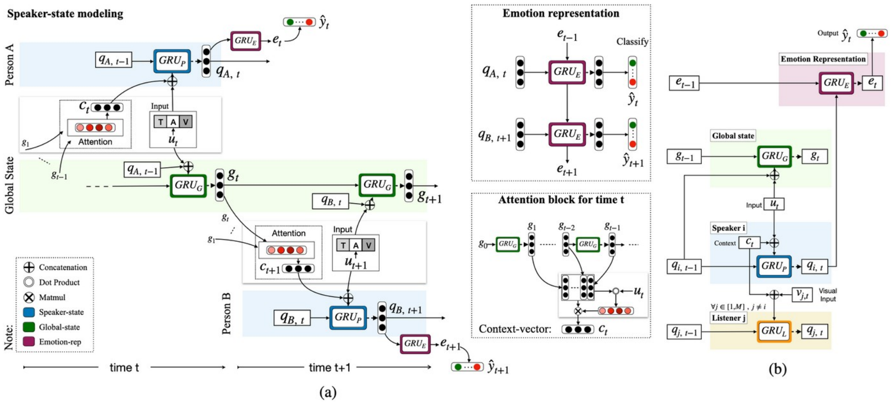


Figure 5: DialogueRNN for multi-party interactions with (a) architecture and (b) update schemes of the speaker and listeners

The last part of this WP will consist of an analysis of the deep learning model to understand its functioning, using several explainability methods. We will follow an in-depth interpretation of the predictions by an overall analysis of the features of social signals contained in important moments for the model prediction (Hemamou, Guillon, Martin, & Clavel, 2021). As this part of the project will be done in 2025, new multimodal medium language models will likely be able to process text and videos (Tang et al., 2023) like COSMO (A. J. Wang et al., 2024), VAST (Chen et al., 2023) and GPT-4. We plan to extract in a zero-shot way prompt-generated features representing lower-level phenomena and that can be important to recognize high-level complex human social phenomena.

Work Package 3: Integration of an Embodied Conversational Agent inside a group debate as a moderator

The last phase of the project consists of moderating debates with an ECA, which is a socially-aware virtual agent using verbal and non-verbal natural language to communicate (Zhao et al., 2016). Like the data collection of the WP1, the participants will have to inform themselves on the topic before the debate.

For this setting, we will organize real-life debates. The participants will be face-to-face, and filmed in order to extract their natural language. The agent will be present on the side of the debaters, between them. The audio of all the participants will be processed by a multimodal real-time off-the-shelf ASR system to get the transcripted text. The video, audio and transcript of each participant will be used as input to the multimodal argumentation mining model.

The **agent will be powered by a large language model, as we know this type of technology already gives impressive results as of early 2024**. This type of model can be prompted to condition their text generation using a character story (J. S. Park et al., 2023), or the history of the conversation for a chatbot. We will prompt the LLM using all types of information gathered by our multimodal model as a mean to integrate multimodality cues. **Outputs from the multimodal argumentation mining model will serve as context to help the LLM be far better at understanding the mental states of the debaters**, and ultimately better moderate the debate. The predictions of the models at the text level, but also non-verbal information about the debaters that would be extracted using explainable algorithms will both be used. This will enhance the model to understand the participants, and at the same time will be helpful for contextual generation of non-verbal behavior of the ECA by enabling the generatioin of very fine-grained mimicry behavior (Bilakhia, Petridis, Nijholt, & Pantic, 2015).

We will analyze different types of policies concerning the agent turn-taking, such as: the agent can stop the discussion when it gets very intense between the participants in order to help them to calm down, the agent can speak when a speech turn is over, the agent can interrupt the speakers,…

The participants will have to fill out a survey with several questions regarding their perception of the agent, their perception of the other participants, and their perception of the other participants arguments. Those answers, coupled with automatic analysis of each of the debates using the model from WP2 will allow quantifying the consensus finding. We will compare with test groups that do not sue agent and do not use an ECA but a text-based virtual agent through a computer. Ultimately we will **conclude of the effect of the embodied conversational agents using statistical tests on the results of the surveys**.

Finally, and even this is not defined enough, but could be a plus of this project, we woud **like to involve people from the public society**, by the mean of organization such as Debating Europe,[3] which aims to leverage information from debates between citizen in order to propose new ideas in the political scene. This collaboration could provide valuable real-world context to our findings and potentially lead to broader societal impact by bridging the gap between academic research and public discourse.

E. Workplan

	Year 1									Year 2									Year 3								
	Trimester 1			Trimester 2			Trimester 3			Trimester 1			Trimester 2			Trimester 3			Trimester 1			Trimester 2			Trimester 3		
WP 1	■	■	■	■	■	■	■	■	■	■	■																
1. Review of SOTA	■	■	■	■																							
2. Preparation of data collection				■	■	■																					
3. Data Collection and annotation						■	■	■	■																		
4. Publish Dataset (Journal)									■	■	■																
WP 2										■	■	■	■	■	■	■	■	■	■								
1. Dataset analysis										■	■	■															
2. Modelisation													■	■	■	■											
3. Model analysis																■	■	■									
4. Publish Results (A*-Conference)																	■	■	■								
WP 3																			■	■	■	■	■	■	■	■	■
1. Integration ECA																			■	■	■						
2. Data collection																						■	■	■			
3. Data analysis																								■	■	■	
4. Publish results (Conf + Journal)																										■	■

[3] https://debatingeurope.eu/

BIBLIOGRAPHIC REFERENCES:

Abbott, R., Walker, M., Anand, P., Fox Tree, J. E., Bowmani, R., & King, J. (2011). How can you say such things?!?: recognizing disagreement in informal political argument. *Proceedings of the Workshop on Languages in Social Media*, 2–11. Retrieved from http://dl.acm.org/citation.cfm?id=2021109.2021111

Anderson, K., André, E., Baur, T., Bernardini, S., Chollet, M., Chryssafidou, E., … Sabouret, N. (2013). The TARDIS framework: Intelligent virtual agents for social coaching in job interviews. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8253 LNCS*, 476–491. https://doi.org/10.1007/978-3-319-03161-3_35

André, E., Bevacqua, E., Heylen, D., Niewiadomski, R., Pelachaud, C., Peters, C., … Rehm, M. (2011). Non-verbal persuasion and communication in an affective agent. In *Cognitive Technologies* (pp. 585–608). https://doi.org/10.1007/978-3-642-15184-2_30

Argyle, L. P., Busby, E., Gubler, J., Bail, C., Howe, T., Rytting, C., & Wingate, D. (2023). AI Chat Assistants can Improve Conversations about Divisive Topics. *ArXiv*.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., … Kaplan, J. (2022). Constitutional AI: Harmlessness from AI Feedback. Retrieved from http://arxiv.org/abs/2212.08073

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2017). Multimodal Machine Learning: A Survey and Taxonomy, 1–20.

Barriere, V., Balahur, A., & Ravenet, B. (2022). Debating Europe : A Multilingual Multi-Target Stance Classification Dataset of Online Debates. In *Proceedings of the First Workshop on Natural Language Processing for Political Sciences (PoliticalNLP), LREC* (pp. 16–21). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2022.politicalnlp-1.3

Baur, T., Clausen, S., Heimerl, A., Lingenfelser, F., Lutz, W., & André, E. (2020). NOVA: A Tool for Explanatory Multimodal Behavior Analysis and Its Application to Psychotherapy. In *International Conference on Multimedia Modeling* (pp. 577–588).

Biancardi, B., Ceccaldi, E., Clavel, C., Chollet, M., & Dinkar, T. (2021). CATS2021: International Workshop on Corpora And Tools for Social skills annotation. In *ICMI '21: Proceedings of the 2021 International Conference on Multimodal Interaction*.

Biancardi, B., Dermouche, S., & Pelachaud, C. (2021). Adaptation Mechanisms in Human–Agent Interaction: Effects on User's Impressions and Engagement. *Frontiers in Computer Science*, *3*(August), 1–19. https://doi.org/10.3389/fcomp.2021.696682

Bilakhia, S., Petridis, S., Nijholt, A., & Pantic, M. (2015). The MAHNOB Mimicry Database: A database of naturalistic human interactions. *Pattern Recognition Letters*, *66*, 52–61. https://doi.org/10.1016/j.patrec.2015.03.005

Bondarenko, A., Fröbe, M., Kiesel, J., Schlatt, F., Barriere, V., Ravenet, B., … Hagen, M. (2023). Overview of Touché,2023: Argument and Causal Retrieval. In *ECIR*.

Bousmalis, K., Morency, L.-P., & Pantic, M. (2011). Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. In *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011*. https://doi.org/10.1109/FG.2011.5771341

Brilman, M., & Scherer, S. (2015). A Multimodal Predictive Model of Successful Debaters or How I Learned to Sway Votes. *Proceedings of the 23rd ACM International Conference on Multimedia*, 149–158. https://doi.org/10.1145/2733373.2806245

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., … Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Experiments with an early version of GPT-4*.

Busso, C., Parthasarathy, S., Burmania, A., Abdelwahab, M., Sadoughi, N., & Provost, E. M. (2017). MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception. *IEEE Transactions on Affective Computing*, *8*(1), 67–80. https://doi.org/10.1109/TAFFC.2016.2515617

Cafaro, A., Wagner, J., Baur, T., Dermouche, S., Torres Torres, M., Pelachaud, C., … Valstar, M. (2017). The NoXi database: multimodal recordings of mediated novice-expert interactions. In *ICMI*.

Cassell, J. (2001). Embodied conversational agents: Representation and intelligence in user interfaces. *AI Magazine*, *22*(4), 67–83.

Chen, S., Li, H., Wang, Q., Zhao, Z., Sun, M., Zhu, X., & Liu, J. (2023). VAST: A Vision-Audio-Subtitle-Text Omni-Modality Foundation Model and Dataset, (NeurIPS). Retrieved from http://arxiv.org/abs/2305.18500

Cho, J., Lu, J., Schwenk, D., Hajishirzi, H., & Kembhavi, A. (2020). X-LXMERT: Paint, Caption and Answer Questions with Multi-Modal Transformers. Retrieved from http://arxiv.org/abs/2009.11278

Chowdhery, A., Narang, S., Devlin, J., Reif, E., Barham, P., Bradbury, J., … Meier-hellstern, K. (n.d.). PaLM : Scaling Language Modeling with Pathways. *ArXiv*, 1–87.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., … Wei, J. (2022). Scaling Instruction-Finetuned Language Models.

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., … others. (2014). SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems* (pp. 1061–1068).

Fröhlich, M., Sievers, C., Townsend, S. W., Gruber, T., & van Schaik, C. P. (2019). Multimodal communication and language origins: integrating gestures and vocalizations. *Biological Reviews*, *94*(5), 1809–1829. https://doi.org/10.1111/brv.12535

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., … Morency, L.-P. (2014). The distress analysis interview corpus of human and computer interviews. *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, 3123–3128.

Haddadan, S., Cabrio, E., & Villata, S. (2020). Yes, we can! Mining arguments in 50 years of US presidential campaign debates. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 4684–4690. https://doi.org/10.18653/v1/p19-1463

Hardalov, M., Arora, A., Nakov, P., & Augenstein, I. (2021). Cross-Domain Label-Adaptive Stance Detection. In *EMNLP* (Vol. 19). Retrieved from http://arxiv.org/abs/2104.07467

Hasan, M. K., Rahman, W., Zadeh, A., Zhong, J., Tanveer, M. I., Morency, L.-P., … Hoque. (2019). UR-FUNNY: A Multimodal Language Dataset for Understanding Humor. Retrieved from http://arxiv.org/abs/1904.06618

Hasan, M. K., Spann, J., Hasan, M., Islam, M. S., Haut, K., Mihalcea, R., & Hoque, E. (2021). Hitting your MARQ: Multimodal ARgument Quality Assessment in Long Debate Video. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 6387–6397. https://doi.org/10.18653/v1/2021.emnlp-main.515

Heimerl, A., Baur, T., Florian Lingenfelser, Wagner, J., & André, E. (2019). NOVA - A tool for eXplainable Cooperative Machine Learning. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*.

Hemamou, L., Guillon, A., Martin, J.-C., & Clavel, C. (2021). Multimodal Hierarchical Attention Neural Network: Looking for Candidates Behaviour which Impact Recruiter's Decision. *IEEE Transactions on Affective Computing*.

Holz, T., Dragone, M., & O'Hare, G. M. P. (2009). Where robots and virtual agents meet: A survey of social interaction research across milgram's reality-virtuality continuum. *International Journal of Social Robotics*, *1*(1), 83–93. https://doi.org/10.1007/s12369-008-0002-2

Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., … Carreira, J. (2021). Perceiver IO: A General Architecture for Structured Inputs & Outputs. Retrieved from http://arxiv.org/abs/2107.14795

Kantharaju, R. B., Langlet, C., Barange, M., Clavel, C., & Pelachaud, C. (2020). Multimodal analysis of cohesion in multi-party interactions. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, (May), 498–507.

Kantharaju, R. B., Pease, A., De Franco, D., & Pelachaud, C. (2018). Is two beter than one? Efects of multiple agents on user persuasion. *Proceedings of the 18th International Conference on Intelligent Virtual Agents, IVA 2018*, 255–262. https://doi.org/10.1145/3267851.3267890

Kantharaju, R. B., Pease, A., Reidsma, D., Pelachaud, C., Snaith, M., Bruijnes, M., … Den Akker, H. O. (2019). Integrating argumentation with social conversation between multiple virtual coaches. *IVA 2019 - Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 203–205. https://doi.org/10.1145/3308532.3329450

Kiesel, J., Spina, D., Wachsmuth, H., & Stein, B. (2021). The Meant, the Said, and the Understood: Conversational Argument Search and Cognitive Biases. *ACM CUI'21*. https://doi.org/10.1145/3469595.3469615

Kiesel, J., Weimar, B., Handke, N., & Weimar, B. (2022). Identifying the Human Values behind Arguments. *Acl*.

Lai, M., Cignarella, A. T., Hernández Farías, D. I., Bosco, C., Patti, V., & Rosso, P. (2020). Multilingual stance detection in social media political debates. *Computer Speech and Language*, *63*. https://doi.org/10.1016/j.csl.2020.101075

Lai, M., Patti, V., Ruffo, G., & Rosso, P. (2018). Stance evolution and twitter interactions in an italian political debate. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10859 LNCS*(July), 15–27. https://doi.org/10.1007/978-3-319-91947-8_2

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, (2). Retrieved from http://arxiv.org/abs/2201.12086

Li, Z., Wallace, E., Shen, S., & Lin, K. (2020). Train Large , Then Compress : Rethinking Model Size for Efficient Training and Inference of Transformers.

Lian, Z., Liu, B., & Tao, J. (2022). Smin: Semi-supervised multi-modal interaction network for conversational emotion recognition. *IEEE Transactions on Affective Computing*.

Liang, P. P., Lyu, Y., Chhablani, G., Jain, N., Deng, Z., Wang, X., … Salakhutdinov, R. (2022). MultiViz: An Analysis Benchmark for Visualizing and Understanding Multimodal Models, (2). Retrieved from http://arxiv.org/abs/2207.00056

Liang, P. P., Zadeh, A., & Morency, L.-P. (2022). Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions, (1), 1–65. Retrieved from http://arxiv.org/abs/2209.03430

Lippi, M., & Torroni, P. (2016). Argument mining from speech: Detecting claims in political debates. *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, 2979–2985. https://doi.org/10.1609/aaai.v30i1.10384

Livingstone, A. G., Fernández Rodriguez, L., & Rothers, A. (2020). “They just don’t understand us”: The role of felt understanding in intergroup relations. *Journal of Personality and Social Psychology*, *119*(3), 633.

Lotfian, R., & Busso, C. (2018). Speech Corpus by Retrieving Emotional Speech From Existing Podcast Recordings. *Transactions on Affective Computing, IEEE*, *XX*(X), 1–14.

Lucas, G. M., Gratch, J., King, A., & Morency, L.-P. (2014). It’s only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, *37*, 94–100. https://doi.org/10.1016/j.chb.2014.04.043

Lyu, Y., Liang, P. P., Deng, Z., Salakhutdinov, R., & Morency, L.-P. (2022). *DIME: Fine-grained Interpretations of Multimodal Models via Disentangled Local Explanations*. *Proceedings of (Preprint)* (Vol. 1). Association for Computing Machinery. Retrieved from http://arxiv.org/abs/2203.02013

Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., & Cambria, E. (2019). DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In *AAAI*.

Mancini, E., Ruggeri, F., Galassi, A., & Torroni, P. (2022). Multimodal Argument Mining: A Case Study in Political Debates. *Proceedings of the 9th Workshop on Argument Mining*, 158–170. Retrieved from https://www.youtube.com/channel/

McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schröder, M. (2012). The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, *3*(1), 5–17. https://doi.org/10.1109/T-AFFC.2011.20

Mehrabian, A. (1971). *Silent messages*. *Silent Messages*. Retrieved from http://www.speakingaboutpresenting.com/presentation-myths/mehrabian-nonverbal-communication-research/

Mestre, R., Milicin, R., Middleton, S. E., Ryan, M., Zhu, J., & Norman, T. J. (2021). M-Arg: Multimodal Argument Mining Dataset for Political Debates with Audio and Transcripts. *8th Workshop on Argument Mining, ArgMining 2021 - Proceedings*, (2014), 78–88. https://doi.org/10.18653/v1/2021.argmining-1.8

Mirzakhmedova, N., Kiesel, J., Alshomary, M., Heinrich, M., Handke, N., Cai, X., … Stein, B. (2023). The Touché23-ValueEval Dataset for Identifying Human Values behind Arguments. In *arXiv (submitted to LREC 24)*. Retrieved from http://arxiv.org/abs/2301.13771

Mochales, R., & Moens, M. (2011). Argumentation mining. *Artificial Intelligence and Law*, *19*(1), 1–22. Retrieved from http://link.springer.com/article/10.1007/s10506-010-9104-x

Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). A Dataset for Detecting Stance in Tweets. https://doi.org/10.18720/MCE.75.10

Morio, G., Egawa, R., & Fujita, K. (2019). Revealing and predicting online persuasion strategy with elementary units. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 6274–6279. https://doi.org/10.18653/v1/d19-1653

Nihei, F., Nakano, Y. I., Hayashi, Y., Huang, H. H., & Okada, S. (2014). Predicting influential statements in group discussions using speech and head motion information. *ICMI 2014 - Proceedings of the 2014 International Conference on Multimodal Interaction*, 136–143. https://doi.org/10.1145/2663204.2663248

Nojavanasghar, B., Baltrušaitis, T., Hughes, C. E., & Morency, L.-P. (2016). EmoReact: A Multimodal Approach and Dataset for Recognizing Emotional Responses in Children. In *ICMI 2016 - Proceedings of the 2016 ACM International Conference on Multimodal Interaction*.

Nojavanasghari, B., Gopinath, D., Koushik, J., Baltrušaitis, T., & Morency, L.-P. (2016). Deep Multimodal Fusion for Persuasiveness Prediction. In *ICMI 2016 - Proceedings of the 2016 ACM International Conference on Multimodal Interaction* (pp. 1–5). https://doi.org/10.1145/2993148.2993176

Ohba, T., Mawalim, C. O., Katada, S., Kuroki, H., & Okada, S. (2022). Multimodal Analysis for Communication Skill and Self-Efficacy Level Estimation in Job Interview Scenario. *ACM International Conference Proceeding Series*, 110–120. https://doi.org/10.1145/3568444.3568461

Olshefski, C., Lugini, L., Singh, R., Litman, D., & Godley, A. (2020). The discussion tracker corpus of collaborative argumentation.

LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings, 1033–1043.

OpenAI. (2023). GPT-4 Technical Report, *4*, 1–100. Retrieved from http://arxiv.org/abs/2303.08774

Ouyang, L., Wu, J., Jiang, X., Ameida, D., Wainwright, C. L., Mishkin, P., … Lowe, R. (2022). Training language models to follow instructions with human feedback. *ArXiv*, *https://op*.

Palau, R. M., & Moens, M.-F. (2009). Argumentation Mining: The Detection, Classification and Structure of Arguments in Text. In *ICAIL* (pp. 98–107).

Park, C. Y., Cha, N., Kang, S., Kim, A., Khandoker, A. H., Hadjileontiadis, L., … Lee, U. (2020). K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data*, *7*(1), 1–16. https://doi.org/10.1038/s41597-020-00630-y

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). *Generative Agents: Interactive Simulacra of Human Behavior* (Vol. 1). Association for Computing Machinery. Retrieved from http://arxiv.org/abs/2304.03442

Park, S., Shim, H. S., Chatterjee, M., Sagae, K., & Morency, L.-P. (2014). Computational Analysis of Persuasiveness in Social Multimedia: A Novel Dataset and Multimodal Prediction Approach. *Proceedings of the 16th International Conference on Multimodal Interaction - ICMI '14*, 50–57. https://doi.org/10.1145/2663204.2663260

Parmar, D., Olafsson, S., Utami, D., Murali, P., & Bickmore, T. (2022). Designing empathic virtual agents: manipulating animation, voice, rendering, and empathy to create persuasive agents. *Autonomous Agents and Multi-Agent Systems*, *36*(1), 17.

Pelachaud, C. (2005). Multimodal expressive embodied conversational agents: Multimodal expressive ECAs. In *Proceedings of the 13th ACM International Conference on Multimedia, MM 2005* (pp. 683–689). https://doi.org/10.1145/1101149.1101301

Pelachaud, C., Busso, C., & Heylen, D. (2021). Multimodal Behavior Modeling for Socially Interactive Agents. *The Handbook on Socially Interactive Agents*, *1*, 259–310. https://doi.org/10.1145/3477322.3477331

Petukhova, V., Malchanau, A., Oualil, Y., Klakow, D., Luz, S., Haider, F., … Alexandersson, J. (2018). The metalogue debate trainee corpus: Data collection and annotations. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 749–755.

Petukhova, V., Mayer, T., Malchanau, A., & Bunt, H. (2017). Virtual debate coach design: Assessing multimodal argumentation performance. *ICMI 2017 - Proceedings of the 19th ACM International Conference on Multimodal Interaction*, *2017-Janua*(1), 41–50. https://doi.org/10.1145/3136755.3136775

Petukhova, V., Raju, M., & Bunt, H. (2017). Multimodal markers of persuasive speech: Designing a Virtual Debate Coach. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2017-Augus*(1), 142–146. https://doi.org/10.21437/Interspeech.2017-98

Rach, N., André, E., Weber, K., Minker, W., Pragst, L., & Ultes, S. (2018). EVA: A multimodal argumentative dialogue system. *ICMI 2018 - Proceedings of the 2018 International Conference on Multimodal Interaction*, (October), 551–552. https://doi.org/10.1145/3242969.3266292

Rahman, W., Hasan, M. K., Lee, S., Zadeh, A., Mao, C., Morency, L.-P., & Hoque, E. (2020). Integrating Multimodal Information in Large Pretrained Transformers, 2359–2369.

Reis, H. T., Lemay Jr, E. P., & Finkenauer, C. (2017). Toward understanding understanding: The importance of feeling understood in relationships. *Social and Personality Psychology Compass*, *11*(3), e12308.

Ringeval, F., Sonderegger, A., Sauer, J., & Lalanne, D. (2013). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*. https://doi.org/10.1109/FG.2013.6553805

Sakketou, F., Lahnala, A., Vogel, L., & Flek, L. (2022). Investigating User Radicalization: A Novel Dataset for Identifying Fine-Grained Temporal Shifts in Opinion. In *LREC* (pp. 3798–3808). Retrieved from http://arxiv.org/abs/2204.10190

Schick, T., Lomeli, M., Dwivedi-yu, J., & Dessì, R. (2022). Toolformer: Language Models Can Teach Themselves to Use Tools.

Schulman, J., Zoph, B., Kim, C., Hilton, J., Menick, J., Weng, J., … Hesse, C. (2022). ChatGPT: Optimizing Language Models for Dialogue.

Sen, T. K., Naven, G., Gerstner, L., Bagley, D., Baten, R. A., Rahman, W., … Hoque, E. (2021). DBATES: DataBase of Audio features, Text, and visual Expressions in competitive debate Speeches. *IEEE Transactions on Affective Computing*.

Shiota, T., & Shimada, K. (2020). The Discussion Corpus toward Argumentation Quality Assessment in Multi-Party Conversation. *Proceedings - 2020 9th International Congress on Advanced Applied Informatics, IIAI-AAI 2020*, 280–283. https://doi.org/10.1109/IIAI-AAI50415.2020.00062

Siddiquie, B., Chisholm, D., & Divakaran, A. (2015). Exploiting multimodal affect and semantics to identify politically persuasive web videos. In *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction* (pp. 203–210). https://doi.org/10.1145/2818346.2820732

Smith, H. J., & Neff, M. (2018). Communication behavior in Embodied virtual reality. In *Conference on Human Factors in Computing Systems - Proceedings* (Vol. 2018-April, pp. 1–12). https://doi.org/10.1145/3173574.3173863

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In *NIPS*.

Swanson, R., Ecker, B., & Walker, M. (2015). Argument mining: Extracting arguments from online dialogue. *SIGDIAL 2015 - 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference*, (September), 217–226. https://doi.org/10.18653/v1/w15-4631

Tang, Y., Bi, J., Xu, S., Song, L., Liang, S., Wang, T., … Xu, C. (2023). Video Understanding with Large Language Models: A Survey, 1–37. Retrieved from http://arxiv.org/abs/2312.17432

Tieri, G., Morone, G., Paolucci, S., & Iosa, M. (2018). Virtual reality in cognitive and motor rehabilitation: facts, fiction and fallacies. *Expert Review of Medical Devices*, *15*(2), 107–117.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., & Salakhutdinov, R. (2019). Multimodal Transformer for Unaligned Multimodal Language Sequences. In *ACL*. Retrieved from http://arxiv.org/abs/1906.00295

Vinciarelli, A., Dielmann, A., Favre, S., & Salamin, H. (2009). Canal9: A database of political debates for analysis of social interactions. In *Proceedings - 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009*. https://doi.org/10.1109/ACII.2009.5349466

Vossen, S., Ham, J., & Midden, C. (2009). Social influence of a persuasive agent: The role of agent embodiment and evaluative feedback. In *Persuasive '09: Proceedings of the 4th International Conference on Persuasive Technology* (Vol. 350). https://doi.org/10.1145/1541948.1542007

Walker, M. A., Anand, P., Abbott, R., & Grant, R. (2012a). Stance classification using dialogic properties of persuasion. *NAACL HLT 2012 - 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings*, 592–596.

Walker, M. A., Anand, P., Abbott, R., & Grant, R. (2012b). Stance Classification using Dialogic Properties of Persuasion. *Naacl*. https://doi.org/10.1142/7114

Walker, M. A., Anand, P., Tree, J. E. F., Abbott, R., & King, J. (2012). A corpus for research on deliberation and debate. *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, 812–817.

Wang, A. J., Li, L., Lin, K. Q., Wang, J., Lin, K., Yang, Z., … Shou, M. Z. (2024). COSMO: COntrastive Streamlined MultimOdal Model with Interleaved Pre-Training. Retrieved from http://arxiv.org/abs/2401.00849

Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., … Wang, L. (2022). GIT: A Generative Image-to-text Transformer for Vision and Language, *2*, 1–49. Retrieved from http://arxiv.org/abs/2205.14100

Weber, K., Rach, N., Minker, W., & André, E. (2020). How to Win Arguments: Empowering Virtual Agents to Improve Their Persuasiveness. *Datenbank-Spektrum*, *20*(2), 161–169. https://doi.org/10.1007/s13222-020-00345-9

Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., & Duan, N. (2023). Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models. Retrieved from http://arxiv.org/abs/2303.04671

Yang, L., Achard, C., Pelachaud, C., Yang, L., Achard, C., Pelachaud, C., … Pelachaud, C. (2022). Multimodal Analysis of Interruptions. In *HCII*.

Zadeh, A., Cao, Y. S., Hessner, S., Liang, P. P., Poria, S., & Morency, L. (2020). CMU-MOSEAS : A Multimodal Language Dataset for Spanish , Portuguese , German and French. In *EMNLP* (Vol. 1, pp. 1801–1812).

Zadeh, A., Chan, M., Liang, P. P., Tong, E., & Morency, L. (2019). Social-IQ : A Question Answering Benchmark for Artificial Social Intelligence. In *CVPR* (pp. 8807–8817).

Zadeh, A., Liang, P. P., Mazumder, N., Poria, S., Cambria, E., & Morency, L.-P. (2018). Memory Fusion Network for Multi-view Sequential Learning. In *AAAI*.

Zhao, R., Sinha, T., Black, A. W., & Cassell, J. (2016). Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In *Proceedings of Intelligent Virtual Agents (IVA 2016).* (Vol. 10011 LNAI). https://doi.org/10.1007/978-3-319-47665-0_20

Zhou, Y., Scherer, S., Devault, D., Gratch, J., Stratou, G., Morency, L., & Cassell, J. (2013). Multimodal Prediction of Psychological Disorders: Learning Verbal and Nonverbal Commonalities in Adjacency Pairs. *Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue*, 160–169.