SAPIENZA
UNIVERSITÀ DI ROMA

Sapienza University of Rome

Department of Computer, Control, and Management Engineering
Antonio Ruberti
Ph.D. in Computer Science

Thesis For The Degree Of Doctor Of Philosophy

Conversational Agents in Human-Machine Interaction

Reinforcement Learning and Theory of Mind in Language Modeling

Thesis Advisor
Prof. Luca Iocchi

Correlatore
Prof. Roberto Navigli

Candidate
Nicolo' Brandizzi
1643869

XXXVI cycle

Dedication

Alla mia famiglia, che mi ha sempre supportato e sopportato durante tutti questi anni di studi. Dedico questo lavoro a voi, per la pazienza infinita e il vostro amore incondizionato. Alla mia partner, compagna in questi mesi di studio matto e disperatissimo, grazie per essere stata la mia roccia e il mio rifugio. Ai miei amici e a tutte le persone che mi sono state vicine, che hanno condiviso con me il viaggio della vita e mi hanno sempre sostenuto.

E infine, a mio padre, che se n'è andato troppo presto. Non ho avuto il tempo di ringraziarti per tutto ciò che hai fatto per me. Questa dedica è per te, per renderti immortale nella memoria e nel cuore.

Questo lavoro è il frutto dell'amore, del sostegno e delle lezioni apprese da ognuno di voi.

Abstract

This doctoral thesis addresses the challenges and advancements in the realm of Human-Machine Interaction, specifically focusing on the agency and misalignment of modern Large Language Models. Initially, we examined the potential for artificial agents to manifest agency within an environment inspired by Social Deduction Games, where Multi-Agent System and Reinforcement Learning shape the interactions. Our findings revealed that introducing a communication channel significantly improved agents' performance, indicative of emergent decision-making abilities. Subsequently, the investigation shifted to the capability of machines to convey information in a manner comprehensible to humans. Through a Referential Game, we identified that agents, while capable of collaboration, struggled with performance when faced with knowledge asymmetry. To address this, we implemented a Multi-Agent Reinforcement Learning approach, aligning with contemporary solutions in the literature and show how it ultimately culminated in the issue of misalignment. In response, our final approach integrated elements from psychology and linguistics to propose a solution to both issues of agency and misalignment. We showed how our method improved communication accuracies solving the agency issue and mitigating the misalignment problem. Moreover, we highlight the environmental and interpretability advantages of our solution. We conclude by stressing the importance of interdisciplinary approaches to refine and understand the capabilities of artificial agents in communication-centric tasks.

Keywords: Human-Machine Communication, Reinforcement Learning, Agency, Misalignment, Large Language Models, Theory of Mind

Contents

List of Figures

List of Tables

Nomenclature

A Complete action set as defined in Equation 4.1.

a Individual action within the set A

Γ Subset of O concerning beliefs about other agents (see §4.1).

γ Specific belief state within Γ (see §4.1).

C Communication-related subset of A (see §4.1).

c Individual action within the communication set C (see §4.1).

π_C Communication policy as described in Equation 4.9

h_t Speaker's decoder logits (see §6.2.2).

L^S Entropy Loss for regularization (see §6.2.2).

α_S Weight of the Entropy Loss term (see §6.2.2).

S Environment set as defined in Equation 4.3.

V Subset of O related to environmental states (see §4.1).

v Specific environmental state within V (see §4.1).

G Game-related subset of A (see §4.1).

g Individual action within the game set G (see §4.1).

$\hat{b}$ Referring Expression tied to a specific target $\hat{t}$ coming from the dataset (see §6.1.3).

h_0 Initial hidden state during LSTM-based decoding (see §6.1.3).

L^{KL} Kullback-Leibler Divergence loss that ensures the model's distribution does not deviate excessively from an original (see §6.2.2).

α_{KL} Weight of the Kullback-Leibler Loss term (see §6.2.2).

P_{lst} Listener player (see §4.1.2).

$L^{P_{lst}}$ Listener loss meant to optimizing the listener's accuracy in the referential game (see §6.2.2).

$\alpha_{P_{lst}}$ Weight of the Listener Loss term (see §6.2.2).

g_l The output of the listener model as defined in §7.1.1

lr_{adapt} Learning rate for adaptation algorithm as defined in §7.1.2.

O Set of observations as defined in Equation 4.2.

o Specific observation state within O (see §4.1).

$\hat{a}$ An action that is considered to be optimal (see §4.2.2).

E Subset of O describing the agents' statuses (see §4.1).

N Number of players in a particular game environment (see §4.1).

L^{π_C} Policy loss used for fine tuning the communication (see §6.2.2).

α_{π_C} Weight of the Policy Loss term (see §6.2.2).

$\bar{g}$ A prediction of the action g (see §4.3).

π_P Prediction policy as described in Equation 4.10

R Reward function.

SL Signal Length: one characteristic of b, also knows as *max_ len* in NLP contexts (see Figure 5.1).

SR Signal Range: describes the number of possible value for b, also knows as *vocabulary size* in NLP contexts (see Figure 5.1).

B Set representing all potential communication signals (see §4.1).

b Specific signal within the set B (see §4.1).

g_{sim} The output of the simulator model as defined in §7.1.1

s_{iter} Number of maximum refinement iterations during adaptation as defined in §7.1.2.

c_s Speaker generated communication, also called Referring Expression (see §4.1.2).

$\hat{t}$ Given a pool of images, $\hat{t}$ is the target image (image of interest) while all the others are considered distractors (see Figure 3.2).

π_U Unified policy as described in Equation 4.7.

v_{ctx} Visual context vector in model architecture (see §4.1.2).

Acronyms

A-Ref Game Asymmetric Referential Game: A variation of the Referential Game that introduces asymmetry, often in the roles, knowledge, or capabilities of the players, to create distinct speaking and listening dynamics (see §6.2.2).

agency Agency: The capacity of an entity, particularly an AI system, to act autonomously and make decisions based on its intentions (see §3.2.2).

AI Artificial Intelligence: The study of creating machines or software that can perform tasks that typically require human intelligence.

AoA Age of Acquisition: Pertains to the age at which a word is typically learned (see §7.2.2).

ApL Apprenticeship Learning: A form of machine learning where the agent learns to perform tasks by observing an expert, typically with the use of Inverse Reinforcement Learning (see §3.2.1).

CE Cross-entropy: A measure of the difference between two probability distributions for a given random variable or set of events.

CIRL Cooperative Inverse Reinforcement Learning: A strategy where two agents engage cooperatively, sharing the objective of inferring and optimizing a reward function through inverse reinforcement learning (see §3.2.3).

CV Computer Vision: A scientific field that trains computers to interpret and make decisions based on visual data.

EmeCom Emergent Communication: The field that studies how language emerges in interactive games between artificial agents (see §2.2.2).

FCN Fully Connected Network: A neural network architecture where each node is connected to every other node.

FFNN Feed-Forward Neural Network: A type of artificial neural network wherein connections between nodes do not form a cycle, and information moves in only one direction, from the input layer, through any hidden layers, to the output layer, without looping back.

G-Speak General Speaker: A baseline or standard speaker model used for comparison with other specialized or adapted speaker models (see §6).

HMC Human-Machine Communication: A sub filed of Human-Machine Interaction focused on the communication (usually linguistic) between humans and machines.

HMI Human-Machine Interaction: The study and design of interfaces that facilitate effective exchanges between humans and computers (see §2.1).

Hum-EmeCom Human-centered Emergent Communication: The sub field of Emergent Communication focused on human natural emergent language (see Brandizzi [2023]).

ImC Image Captioning: A task in computer vision and natural language processing where a system generates textual descriptions of images (see §3.3.3).

IND IN-Domain: In the context of domain-specific words, IND refers to words which are present in that specific domain (see §6.2.3).

IRD Inverse Reward Design: A framework for learning a reward function from human-designed rewards, considering that the designer might have had limitations or made approximations while specifying the reward (see §3.2.1).

IRL Inverse Reinforcement Learning: A machine learning framework that seeks to infer the reward function of an agent by observing its behavior, rather than using a pre-defined reward function (see §3.2.1).

KL-Divergence Kullback-Leibler Divergence: A measure used in information theory to quantify the divergence between two probability distributions (see §6.2.2).

LD Language Drift: The gradual evolution and change in linguistic conventions and usage within a communication system, whether in human languages or artificially emergent languages in machines (see §2.2.2).

LLM Large Language Model: Advanced machine learning models designed for tasks involving natural language understanding and generation.

LSTM Long-Short Term Memory: A type of recurrent neural network architecture optimized for long-term dependencies.

MARL Multi-Agent Reinforcement Learning: An extension of reinforcement learning where multiple agents learn to act in an environment, often with or against each other (see §2.3.3).

MAS Multi-Agent System: Systems composed of multiple interacting agents, which can be either computational entities, like software processes, or physical entities, such as robots.

MBRL Model-Based Reinforcement Learning: A subtype of reinforcement learning where the agent develops a model of the environment to predict future states, thereby aiding in decision-making (see §2.3).

MDP Markov Decision Process: A mathematical model used in decision-making problems, representing the relationships between states, actions, and rewards in a stochastic environment (see §2.3).

MFRL Model-Free Reinforcement Learning: A subtype of reinforcement learning that does not require an explicit model of the environment, relying instead on direct experience to make decisions (see §2.3).

misalignment Misalignment: The discrepancy that arises when an AI's objectives or actions do not align with human intentions, values, or expected outcomes (see §3.2.1).

MRR Mean Reciprocal Rank: A measure used to evaluate systems that return a ranked list of answers to queries.

NLP Natural Language Processing: A branch of AI focused on enabling computers to understand, interpret, and generate human language.

NMT Neural Machine Translation: Utilizing neural network models, particularly sequence-to-sequence models, to perform language translation tasks by learning to map input sequences to output sequences (see §3.2.3).

ObS Observation Space: The set of all possible observations an agent can perceive in an environment.

OOD Out-of-Distribution: In the context of domain-specific words, OOD refers to words which are not present in that specific domain (see §6.2.3).

PoS Part of Speech: Refers to categories into which words are classified based on their syntactic roles and functions, such as nouns, verbs, adjectives, etc (see §6.1.4).

PPO Proximal Policy Optimization: An algorithm used in reinforcement learning to optimize control policies.

Ref Game Referential Game: A communicative game wherein players, often modeled as agents, communicate about objects or concepts within a predefined referential domain (see §3.1.3).

RefEx Referring Expression: Linguistic expressions that identify a particular entity, often within a context where many entities could be referenced (see §6.1.1).

ReLU Rectified Linear Unit: A type of activation function for neural networks (see [Agarap, 2018]).

RL Reinforcement Learning: A type of machine learning where an agent learns to behave in an environment by performing actions and receiving rewards.

RL-FT Reinforcement Learning Fine-Tuning: The process of tweaking the parameters of a model with Reinforcement Learning to optimize it for a specific task (see §6).

RL-Speak Reinforcement Learning Fine-Tuning Speaker: A speaker model that has been fine-tuned using reinforcement learning methodologies (see §6).

RLHF Reinforcement Learning from Human Feedback: A framework wherein reinforcement learning is enhanced using feedback derived directly from human interactions and observations (see §3.2.3).

RNN Recurrent Neural Networks: Neural networks designed for sequential data processing, useful in tasks like time series forecasting or language modeling

RSA Rational Speech Act: A framework that models communication between speakers and listeners as a form of rational action under uncertainty (see §3.3.2).

SDG Social Deduction Game: Games where players take on hidden roles and must deduce each other's identities (see §3.1.3).

Tanh Hyperbolic Tangent Function: The Tanh activation is an activation function used in neural networks.

ToM Theory of Mind: A concept from cognitive science which refers to the ability of an individual to ascribe mental states, like beliefs and desires, to oneself and to others (see §3.3.1).

TTR Type-Token Ratio: A linguistic metric that calculates the diversity of vocabulary used in a text by dividing the number of different words (types) by the number of total words (tokens) (see §6.1.4).

TUR Type-Utterance Ratio: A measure of lexical diversity within spoken discourse, determined by the ratio of distinct word types to the total number of utterances (see §7.2.2).

Chapter 1

Introduction and Overview

In the early stages of AI research, there was excitement surrounding the potential of machines to outperform human expertise in a multitude of domains. The focus was distinctively AI-centric, prioritizing the advancement of the machine's capability over human-AI collaboration.

In 1997, the world witnessed a historical moment in Artificial Intelligence as IBM's Deep Blue supercomputer defeated world chess champion Garry Kasparov [Campbell et al., 2002]. This milestone marked a new era where machines began challenging human intellect in areas once believed to be exclusively human domains. Building on this success, IBM introduced Watson in 2013 [Ferrucci et al., 2013], which showcased its broad knowledge by winning the quiz show *Jeopardy!* against top competitors. As AI research progressed, the challenges addressed became even more complex. By 2016, Google's AlphaGo stood out by defeating Lee Sedol, a leading figure in Go [Silver et al., 2016], a game known for its vast complexity. Further advancements came from researchers at DeepMind who built an AI system mastering the real-time strategy game StarCraft II [Vinyals et al., 2019], which exhibited superhuman decision-making and strategic thinking. These achievements, while initially confined to the realm of games and competitions, had profound implications. The underlying message was clear: AI was not just a tool to assist humans; it had the potential to surpass and replace us. This set a precedent, leading many industries and research domains to explore avenues where AI could replace human effort, aiming for efficiency and precision that was, until then, unimaginable.

However, over the past decade, the research community has shifted from viewing AI as a replacement for human tasks to one that emphasizes a symbiotic relationship. This human-centric approach focuses on leveraging AI to augment human capacities, assisting in everyday tasks, and mitigating repetitive responsibilities [Riedl, 2019, Shneiderman, 2021, Xu, 2019]. Davenport and Kirby [2016] describes this shift with the notion of *augmentation*, emphasizing that AI's role is to enhance, not replace, human efforts. In the labor market, Bessen [2018] provides empirical insights, revealing that AI often leads to role evolution and a spike in overall productivity when paired with humans. Thus, the overall sentiment suggests a future wherein human-AI collaborations yield results surpassing what either could achieve separately.

The interaction between humans and machines is a crucial aspect of human-centric AI, and it should take place in domains where humans are already familiar and require little to no training. Given the universality of language in daily human experiences, the focus has been shifted toward language-based applications rather than specialized tasks such as coding and mathematics.

In particular, Human-Machine Interaction must be grounded in natural language, which presents a challenge of teaching artificial agents to communicate, most often in English. Recent advances in Natural Language Processing have led to the emergence of the transformer architecture, which has become the preferred approach for language-based applications. However, current architectures present a challenge in their reliance on predicting the next word in a sentence rather than understanding the context and purpose behind language usage, as humans do. While humans use language as a tool for coordinating and communicating to survive in a shared environment, Artificial Intelligence may struggle to grasp the intricacies of language. For instance, a sentence like "That apple is..." could be completed with multiple options such as "red", "tasty", or "bad", but the AI may lack the ability to question the underlying meaning and context behind the sentence, which could limit its ability to truly understand the world and communicate effectively with humans.

This challenge of learning mismatch remains unsolved due to the difficulty of expressing values and intentions as a set of rules or formulas. A prominent example can be found in OpenAI's ChatGPT, which was released in December 2022. This Large Language Model rapidly captured global attention, becoming the fastest-growing application in history to reach 100 million monthly active users [Hu, 2023]. Its success spans from the novel application of a training methodology to such a large-scale model. ChatGPT was trained using feedback from human annotators with Reinforcement Learning. More specifically, its objective was to generate responses most likely to be positively rated [Bai et al., 2022a]. While effective in many scenarios, this method encourages the AI to make up factual information instead of admitting ignorance. As long as the human annotator lacks the knowledge or expertise to recognize the incorrect information, the AI may receive positive feedback despite generating inaccurate or misleading responses. This results in a lack of accountability and trustworthiness in the AI's responses, as well as potentially harmful consequences for those who rely on the information provided by the AI.

1.1 Context and Motivation

The year 2023 has witnessed an exponential surge in the deployment of Large Language Models across diverse fields, including medicine, chemistry, and economics [Clusmann et al., 2023, Sallam, 2023]. In financial terms, the LLM market is predicted to grow from 10.5 Billion USD in 2022 to 40.8 Billion USD by 2029 [Reports, 2023]. This financial trajectory mirrors the expected increased adoption of LLMs in the coming years.

Given the widespread impact of this technology on society, escalating issues emerge about its reliability. Beyond pressing ethical debates, such as the malicious use of LLMs for disinformation, a crucial concern revolves around these models' comprehension of their societal impact and alignment with human values. These concerns touch upon concepts of *agency* and *misalignment*, key topics in our discussions. At their core, they question how autonomously models can act while aligning with our desired outcomes and societal context.

While philosophical debates around the definition of agency span decades, a practical perspective allows us to view agency as a system's capability to foresee its environmental actions. LLMs' learning framework inhibits their ability to measure their influence on users and broader societal ramifications. A case in point is the noticeable societal impact across political, economic, and educational spheres by models such as ChatGPT [Farina and Lavazza, 2023]. However, AI's impact on society is not new to history, especially when considering negative outcomes. An unfortunate example is the use of predictive policing algorithms like COMPAS [State of, 2008], operational from 2010 to 2016. Trained on biased data, COMPAS exhibited persistent discrimination against minorities, intensifying policing efforts and subsequently reinforcing data biases. The tangible fallout from COMPAS, affecting thousands of lives, underscores the human consequences of deploying such tools without necessary care.

COMPAS's biases also emphasize the broader *misalignment problem*, which can be described as the mismatch between human values and machine goals. In COMPAS's case, the system's goal (to predict future criminal activity based on historical data) did not account for the deeply rooted societal biases in that data. Instead of providing impartial assessments, the system perpetuated and amplified those very biases, resulting in judgments that further marginalized certain populations. This misalignment not only failed to promote justice but actively worked against it, underscoring the inherent risks of letting automated systems dictate decisions without adequate oversight and understanding. In general, when complex human behaviors are abstracted into mathematical formulations, we often encounter a radical deviation from reality. Our contemporary era offers numerous instances of misaligned AIs, which, when applied in real-world settings, inadvertently amplify existing biases [Gabriel, 2020, Yudkowsky, 2016].

1.2 Problem Identification: Enhancing Human-Machine Interaction Through Language

Large Language Models, as we have seen, possess the potential to profoundly affect a significant part of humanity, bringing both benefits and drawbacks. However, in their current state, they grapple with challenges related to understanding their societal role and human values. As researchers, our objective should be to address this issue by enhancing LLMs with the awareness of how their communications influence their surroundings and facilitate a deeper comprehension of humans. This multidisciplinary problem intersects various fields, but our primary focus is on Human-Machine Interaction. Within this domain, the recent advancements in Natural Language Processing have led to the emergence of a subfield where we can engage in direct dialogues with machines. This area of study is known as Human-Machine Communication (HMC) with dedicated periodicals (such as Human-Machine Communication) and journal issues [Etzrodt et al., 2022]. Though it emphasizes linguistic interactions, it remains receptive to future multimodal communications, encompassing domains like vision and sound.

Our ambition in this direction is to develop machines capable of adapting to diverse contexts and individuals. Consider an AI virtual assistant designed to facilitate online learning for students worldwide. One student, based in New York, uses colloquial language, slang and refers to cultural phenomena unique to the region. Another student, from rural India, might communicate in English but with different sentence structures and references. If the virtual assistant rigidly adheres to a standard language model, it could misinterpret the intent of these students. The consequence being frustration, miscommunication, and an ineffective learning experience. From a broader perspective, adaptability encompasses more than the ability to transition between languages or cultural differences. It also pertains to recognizing and adjusting to an individual's evolving communication style. Analogous to a teacher who discerns each student's learning style and adjusts instruction accordingly, a sophisticated AI should continuously learn and adapt to each user's distinctive communicative patterns.

In essence, an AI system lacking this adaptability remains limited, failing to capture the full spectrum of human interaction. A question then arises: *How can we build AI conversational agents to be adaptable? And how can we ensure it understands not just words but also the intentions behind them?* Addressing these questions is fundamental to our pursuit for improved Human-Machine Communication.

1.3 Navigating the Thesis: A Structural Overview

We have identified two primary challenges in the domain of Human-Machine Communication: the lack of *agency* and *misalignment* in modern Large Language Models. This work aims to address these challenges systematically.

Our introduction starts by equipping the reader with essential foundational knowledge. Chapter 2 elucidates key concepts, such as Human-Machine Interaction (§2.1) , Computational Linguistics (§2.2), and Reinforcement Learning (§2.3). Throughout this chapter, we maintain an emphasis on interactions: between humans, between machines, and, crucially, between humans and machines.

Subsequently, Chapter 3 offers a literature review, laying the groundwork for our methodological approach. We examine the dynamics of cooperation and competition on artificial agents (§3.1), particularly focusing on emergent languages. The exploration centers on Social Deduction Games (§3.1.3), where *The Werewolf* and the Referential Game (Ref Game) stand out. The next Section §3.2 focuses on the interplay between language modeling and RL. Here, topics like the misalignment problem (§3.2.1), the role of agency in RL (§3.2.2), and the current uses of Reinforcement Learning in language modeling (§3.2.3) are discussed. Our review then moves towards adaptability solutions for language models (§3.3), where we discuss the Theory of Mind (§3.3.1), its integration in computational linguistics (§3.3.2), and associated tasks (§3.3.3).

Equipped with a robust understanding of the landscape and its challenges, we introduce our unique problem definition and proposed solution in Chapter 4. A mathematical foundation for *The Werewolf* (§4.1) and the Referential Game (§4.1.2) sets the stage for the introduction of our solution framework (§4.2). Two distinct solution frameworks emerge: one emphasizes communication as the sole action, leading to a *unified action policy* (§4.2.1), while the other envisions agents with diverse roles, both communicative and interactive, resulting in what we term the *disjoint action policy* (§4.2.2). Building upon this, we propose an extended version of the *disjoint action policy* through the lens of the previously discussed Theory of Mind (§4.3). This perspective envisions non-communicative actions as predictors of other agents' behaviors. This approach paves the way for our primary contribution: a prediction policy combined with an iterative communication refinement process.

1.3.1 Methodological Approach and Contributions

Our research methodology systematically addresses the challenges of agency and misalignment. The approach is structured into three distinct phases, each building upon the findings of the previous one. Each of these phases is grounded in peer-reviewed materials published during my Doctoral period. Specifically, phase 1 is informed by *Rlupus: Cooperation through emergent communication in* *the werewolf social deduction game* [Brandizzi et al., 2021] and *Emergent communication in human-machine games* [Brandizzi and Iocchi, 2022], phase 2 draws from both *Speaking the language of your listener: Audience-aware adaptation via plug-and-play theory of mind*[1] [Brandizzi et al., 2023] and *Towards more human-like ai communication: A review of emergent communication research* [Brandizzi, 2023], and phase 3 is based on Brandizzi et al. [2023].

Moreover, according to the solution framework introduced previously, phase 1 is based on the *unified action policy*, phase 2 on the *disjoint action policy*, and phase 3 on the Theory of Mind version of the *disjoint action policy*.

Understanding AI Agency through Game-Based Dynamics

In the initial phase, we deal with a fundamental question: *Can agency manifest in artificial systems given an adequate learning framework?* Inspired by the literature, we resolve to Reinforcement Learning and Multi-Agent Systems to answer this question. Chapter 5 tackles this issue through an artificial game environment inspired by *The Werewolf* Social Deduction Game.

Starting, we detail the game's mechanics and dynamics in §5.1, breaking down the environment (§5.1.1) and discussing the roles of the two teams involved, villagers and werewolves (§5.1.2). We then measure the game dynamics in (§5.2). First, we focus on the outcome's analysis under randomized behaviors (§5.2.1). Our findings highlight a significant disadvantage for one group (the villagers), who achieve victory 4% of the time without communication. This outcome sets our baseline, against which we assess these agents' learning and communicative capabilities.

Next, we show that without communication, the use of RL does not improve the villagers' chances of winning. We posit that: *if agents can leverage a communication channel for cooperative intent, it serves as evidence of their emergent agency*, i.e., using communication to influence the game outcome. To assess our hypothesis, we introduce a communication channel in the game without incentivizing agents to use it. Our hypothesis is proven when we analyze how villagers' win rates increase from 4% to 40% upon the introduction of communication. We validate these findings across two player configurations: nine and twenty-one players. Drawing from subsequent research by Lipinski et al. [2022], we uncover how the villagers craft a strategy resembling a Turing test to pinpoint the werewolf (§5.2.2). Yet, a challenge remains: the evolved language, while efficient, consists of bits and numbers, making it non-interpretable by humans. Our exploration ends with the awareness that while the right learning paradigm can indeed spark agency, the resultant language must resonate with humans.

Enhancing Language Model Agency with Reinforcement Learning

Building upon our earlier findings, the subsequent portion of this work (§6) shifts to agents that communicate in a human-interpretable manner. We introduce two agents, a speaker and a listener (§6.1), set to play a Ref Game. We detail their training dataset (§6.1.1), the framework in play (§6.1.2), and the structural design of the models (§6.1.3).

This chapter revolves around a central research question: *Can agents, trained independently from each other, collaboratively solve a game?* We put this to the test in §6.1.4 and subsequently

[1] The content and solution proposed in this chapter has been developed in collaboration with the Dialogue Modelling Group at the University of Amsterdam.

dissect their performance metrics and communication strategies. Our findings are promising; the agents surpass random success rates, which we attribute to their shared expertise in both linguistic and visual domains (§6.1.5). Yet, an issue emerges. This success leans heavily on the agents sharing complete knowledge, an idealistic scenario not reflected in real-world dynamics, where knowledge disparities are common.

To simulate these disparities, Section 6.2 introduces a different version of the Referential Game. Here, game data is split across various domains, laying the groundwork for an Asymmetric Referential Game (§6.2.1). With the speaker retaining a complete knowledge base, we build listeners limited to specific domains. As anticipated, this setup severely impairs game outcomes, with some performances decreasing to random chance when a listener confronts unfamiliar domains. The primary challenge arises from the speaker's limited exposure to other agents during training, suggesting a lack of agency.

Taking inspiration from the previous chapter, we turn to Reinforcement Learning, finetuning the speaker to adapt to a designated listener (§6.2.2). The outcomes, documented in §6.2.3, show the adapted speaker outperforming its original version. Yet, a deeper analysis of the speaker's lexicon uncovers a reliance on select keywords that trigger specific listener reactions. We relate this trend to the misalignment problem in §6.3 where higher game performances do not inherently translate to better communication. In §6.3.1, we critique the constraints of our approach, particularly in the context of contemporary Large Language Models training. While we can address the agency issue by adjusting training methods, the misalignment problem requires a different solution.

Beyond Finetuining: Theory of Mind for Improved Communication

In our final methodology Chapter 7, we address both the issue of agency and misalignment. We present our novel solution (§7.1), and introduce an auxiliary model, termed the *simulator* (§7.1.1). The simulator's primary function is to learn the listener's behavior during interaction rounds. Once the simulator demonstrates proficiency in predicting listener actions based on the environment, we implement a unique adaptation mechanism (§7.1.2). This mechanism leverages the simulator's insights to guide the speaker's responses, ensuring they resonate more effectively with the listener.

To validate our approach, we perform a comprehensive analysis in (§7.2). Initially, we demonstrate how our adaptation strategy enables the speaker to adjust to various listeners, enhancing game performance (§7.2.1). Subsequently, we examine the evolved language (§7.2.2). Our findings indicate a noticeable improvement in terms of vocabulary diversity (avoiding the previous issue of distribution collapse) and enhanced human interpretability.

Furthermore, in §7.3, we emphasize the advantages of our method. Notably, our approach avoids retraining the speaker's model, a particularly resource-intensive process for LLMs. We also revisit the agency dilemma, which we believe is effectively addressed by employing the simulator. Moreover, the misalignment issue is now alleviated, thanks to the improved understanding between the speaker and listener. Nonetheless, our method is not without limitations. In §7.3.2, we discuss potential ethical implications, particularly concerning the potential of LLMs to influence human decisions.

1.3.2 Conclusions: Summarizing Insights and Ethical Dilemmas

In Section 8, we conclude our journey with a summary of our key findings (§8.1). The exploration acknowledges certain limitations due to resource constraints impacting the use of LSTM over transformer models, affecting the quality of language generation, and the exclusion of human participants, which may limit the depth of insights into human-AI dynamics. These limitations are further discussed in Section 8.2.

Looking forward, Section 8.3 outlines the expected impact of our contributions across various domains. We explore the theoretical implications (§8.3.1), where our novel approach could redefine how language models are trained and adapted. In the industrial area (§8.3.2), we speculate on the potential of our research in shaping future AI products, particularly in terms of personalization and on-device functionality. On the societal front (§8.3.3), we contemplate our work's role in promoting a more inclusive, environmentally-conscious, and ethically-aware AI landscape.

More importantly, we delve into the realm of ethical questioning in §8.4. Here, we examine the extent to which AI should mirror human characteristics (§8.4.1), ponder the future landscape of our working lives in the wake of AI advancements (§8.4.2), and explore the often overlooked considerations in the race towards AI supremacy (§8.4.3).

By concluding with these inquiries, our goal is to ignite a discourse on the ethical ramifications of modern AI development, emphasizing the importance of discussion even in the absence of definitive answers.

Chapter 2

Interaction and Language in AI: Theoretical Background and Context

In this chapter, we lay the foundational knowledge for comprehending the content of this work.

Specifically, we start with an exploration of the evolution of human interactions (§2.1), detailing the characteristics of human-to-human communication (§2.1.2), including aspects like contextual adaptation. Subsequently, we explore the proprieties of Human-Machine Interaction, examining the necessity of tailoring communication to suit different audiences (§2.1.1).

Section §2.2 continues this exploration, focusing on language and computational linguistics. Our interest lies in the progression of language modeling in AI (§2.2.1), tracing its journey from classical Feed-Forward Neural Network to contemporary transformer models. We then shift to Emergent Communication frameworks (§2.2.2), where language is evolved in artificial contexts under specific environmental pressures. Here, we examine the application of this framework in HMI and introduce the concept of Language Drift, a phenomenon that will later be linked to the issue of misalignment.

Finally, in Section §2.3, we discuss the fundamental principles of Reinforcement Learning, such as Markov Decision Processes (§2.3.1). The chapter concludes by differentiating between model-free and model-based RL approaches (§2.3.2) and discussing their applications in Multi-Agent System (§2.3.3). This sets the stage for a deeper understanding of the interplay between AI and human communication, which is central to our work.

2.1 Exploring Human-Machine Interaction

The concept of Human-Machine Interaction (HMI) first emerged in 1976, introduced by Carlisle [1976] initially as *man-machine interaction*. Carlisle [1976] underscored the need for design solutions that integrate both human and machine elements during an automation revolution. However, the work of Card [2018] largely popularized the term Human-Machine Interaction by establishing its first scientific foundations. They asserted, «Our purpose in this book is to help lay a scientific foundation for an applied psychology concerned with the human users of interactive computer systems», highlighting the necessity for a symbiotic relationship between psychology and computer science to formulate systems that incorporate computers into human activities. HMI has sustained its multidisciplinary nature over decades, encompassing varied fields like social sciences and, more recently, ethics, in light of the pervasive integration of computational devices in our daily lives. A central question often arising in the field is whether machines should be crafted to perfectly mimic humans, becoming indistinguishable from us, or whether they should complement us, maintaining a clear distinction.

While this question remains open-ended, it prompts contemplation about our expectations from machines. *What human attributes should be emulated, and which should remain uniquely ours?* To explore these questions, we must first identify the human qualities that are important to us and explore how they can shape HMI.

In this section, our objective is to examine the properties of human interaction, providing the reader with a broad understanding of the field and, crucially, the knowledge needed to grasp the scope of this work. Subsequently, we will focus on human communication interactions and the properties that define them. Throughout the section, we will explore questions about the nature of these properties and the feasibility and appropriateness of enabling machines to emulate them.

2.1.1 Properties of Human Communication

In our exploration of communication, we focus on two-person interactions, the simplest unit wherein interesting communication properties emerge, especially within the broader context of Human-Machine Interaction. While this exploration is also relevant to human-robot interaction, our attention is primarily captured by non-embodied computer interaction.

Contextual Adaptation in Communication

When discussing the importance of context awareness, we might examine a scenario where a person is tasked with sharing project updates with a colleague and a close friend. The communication format, language, and depth of detail would likely be adjusted based on the recipient, displaying our skill in adapting to different conversational contexts. Making machines as contextually skilled as humans becomes a hard challenge. Understanding humor, for instance, demands awareness of what is appropriate and where, questions that even humans sometimes clash with. *What makes a joke fitting in one scenario but offensive in another? How do we comprehend which comments about workloads might be light-hearted in friendly circles but improper in a professional meeting? How has the acceptance of certain jokes shifted with evolving societal perspectives?* Translating this awareness to machines brings forward a dilemma: How do we approach understanding and generating context-aware communication?

Symbolic vs. Machine Learning Approaches

Strategizing communication behavior in machines generally employs two methodologies: symbolic reasoning and machine learning. In symbolic reasoning, we might set explicit rules for a robot, defining appropriate behaviors and statements in a workplace. Yet, creating a rule for every possible scenario is a monumental, perhaps impossible task, especially given the dynamic nature of societal norms. Alternatively, a machine learning approach may involve exposing a system to data derived from various social interactions, allowing it to recognize patterns and, thus, learn how to emulate similar behaviors. However, this too comes with its own set of challenges related to ensuring the derived behaviors align with ethical and societal expectations. The question of alignment, in particular, is critical (discussed in §3.2). For example, a machine learning model might infer patterns and adopt behaviors from its training data that are incongruent with ethical or societal expectations, such as reinforcing stereotypes or exhibiting bias. While each approach carries its respective merits and pitfalls, this work leans more toward utilizing machine learning, albeit advocating for a blend of both methodologies as a more plausible solution.

Defining and Exploring 'Context'

Context is a multifaceted concept, embracing temporal, spatial, and event-based dimensions, each of which can significantly influence interaction. For example, a chat in a busy office would naturally differ from a whispered conversation in that same space after most have gone home. Likewise, a discussion during a team project meeting would adhere to different norms than a discussion during a workplace emergency. The depth and complexity of context go far beyond these simple examples, presenting numerous challenges when trying to enable machine understanding in this domain. By exploring this topic, we aim to reveal its complexity and breadth, highlighting the challenges of instilling this level of understanding in machines.

2.1.2 Communication in Human-Machine Interaction

> Communication is essentially a social affair.
>
> Cherry [1966]

In this work, our exploration focuses primarily on verbal communication signals, namely words, although it is essential to acknowledge that a significant portion of human communication is non-verbal [Mehrabian et al., 1971]. Despite this, the increasing digitization of our age positions verbal communication as a central element deserving scrutiny.

This work introduces a specialized sub-field within HMI, termed Human-Machine Communication (HMC). This area, evolving from advancements in language modeling and the development of chatbots like ChatGPT [OpenAI, 2023], Claude [Bai et al., 2022b], and Bard [Manyika, 2023], essentially leans on machine learning. Here, vast amounts of data are processed by statistical predictors to identify and replicate patterns. While this methodology is the basis of today's Large Language Models and will be discussed in further detail later, it remains largely incomprehensible to the average user. Nonetheless, the proliferation of consumer-oriented chatbots, engaging millions of users simultaneously, has shifted the paradigm in HMI. Where the focus was once on developing methodologies that simplified Human-Machine Interaction, we now explore the potential of expressing our intentions to machines using language. Given the technology's wide-reaching implications and use by a significant portion of the global population, navigating a research direction into this emerging area is fundamental in HMI.

Adjusting Communication to Audience

Returning to our previous discussion on communication context, this work focuses on a unique type influenced by the speaker's intentions. My writing here, for example, assumes a fundamental understanding of AI from you, the reader. A conversation with my niece would require a shift to simpler, more accessible language. This adjustment in communication, dependent on the knowledge and intentions of the audience, is known by different names across various fields of study. In linguistics, it is referred to as the Rational Speech Act (RSA) (discussed in section §3.3.2), where the focus is on reasoning through communication. In contrast, within computer science, and particularly in Reinforcement Learning, Opponent Modeling mainly focuses on modeling the actions and behaviors of entities within specific environments, often in the context of games. From a psychological perspective, the phenomenon is associated with the Theory of Mind (ToM), highlighting humans' inherent capability to reason about others. Given its broader focus, encompassing view of communication and action, ToM will be a focal point in our discussions and is introduced in Section §3.3.1.

In conclusion, the role of communication, especially language, is central in our discussions, subsequently leading us to an introduction to computational linguistics through the combined perspectives of computer science and linguistics.

2.2 Language and Computational Linguistics

Language has been a constant companion to humanity since its origin [Arcadi, 2000]. The unique capability to convey complex ideas through language sets us apart from other species. Thus, the origins of language have intrigued human beings for millennia, initially through religious contexts and later, through philosophical and scientific inquiries in ancient Greece [Bowie, 2007] and the scientific methods that took root in the 18th century.

Historically, language studies have primarily taken an anthropological approach, probing into language evolution, development, and acquisition to understand its dynamism across various timescales. The surge in computational power has given rise to a new interdisciplinary field, computational linguistics, which melds linguistics and computer science together. Computational linguistics began in the 1950s with U.S. initiatives to automatically translate foreign texts, particularly from Russian to English [Hutchins, 1999]. Initially, it employed a rule-based method, an approach bearing similarity to the symbolic reasoning discussed previously. Over time, it evolved into what is broadly known today as Natural Language Processing (NLP).

The following section explores the general mechanisms of language modeling and its recent advancements, while keeping the discussion broad.

2.2.1 Mechanics of Computational Language Modeling

Computational language modeling is the field that focuses on understanding and generating human language through computational means. The fundamental concept involves modeling how humans acquire and manipulate language by replicating human-like learning and processing capabilities through computational algorithms. Language models subsequently undergo exposure to inputs akin to those encountered by humans, with the model's responses being analyzed and compared with human data. One such learning methodology is understanding the distribution over sequences of words, utilizing statistical approaches to comprehend a language's grammar and structure. The standard training modality involves predicting obscured words within a sentence [Goodman, 2001, Mikolov et al., 2011], enabling them to adapt to large text corpora even without the necessity of labels.

Evolving Neural Language Models

The evolution of LMs has been pronounced, especially over the past three decades. In the early 2000s, classical Natural Language Processing approaches were ported to Feed-Forward Neural Networks (FFNNs) [Bengio et al., 2000, Morin and Bengio, 2005]. The FFNNs performed better than classical NLP approaches due to their non-linearity. Still, the architecture was limited to fixed input sizes, disregarding contextual information. This issue was addressed by Mikolov et al. [2013], where words were converted to low-dimensional vectors, mitigating the curse of dimensionality. However, the context size remained limited.

Incorporating Recurrent Neural Networks

In 2010, a new architectural development was introduced in the form of the Recurrent Neural Networks (RNN) [Mikolov et al., 2010, 2011]. RNNs, structurally designed to maintain a hidden state that captures information about previous steps in a sequence, was fundamental in tasks where context or sequential order are critical, such as machine translation and speech recognition. However, RNNs experience limitations during training, such as the vanishing and exploding gradient problems [Hochreiter, 1998]. These issues, characterized by the gradients tending toward zero or escalating towards infinity during backpropagation, hampered the model's learning capability and subsequently slowed down training or led to unbounded parameter values.

Introduction to Long-short Term Memory Units

Sundermeyer et al. [2012] introduced the Long-Short Term Memory (LSTM) as a solution to the vanishing and exploding gradient problems prevalent in conventional RNNs. An LSTM is designed to better handle different types of memory data through the interplay of its three distinct gating mechanisms:

- **Input Gate**: Dictates which values in the input matrix should be updated.
- **Forget Gate**: Examines prior context and current input, yielding a value between 0 (complete forgetfulness) and 1 (total recall).
- **Output Gate**: Determines which parts of the input and preceding context will constitute the cell output.

While LSTMs have been successful in stabilizing the learning process by providing a more refined control over information flow compared to traditional RNNs. However, the optimization of training algorithms remains a challenge, given their inherent sequential processing nature.

The Rise of Transformers

Recent technological advances and attention-based architectures [Vaswani et al., 2017] have paved the way for the predominance of transformer models, which supplant RNN cells with self-attention and fully connected layers that are highly parallelizable and consequently more computationally economical. This enables transformers to scale with more data and resources, thus replacing LSTMs in various domains. While transformers have advanced NLP, they are not without criticism and challenges, including their resource-intensive nature and potential to perpetuate biases in training data. Ethical implications of their application necessitate careful consideration within the academic and practitioner communities. In this regard, innovations in LSTMs often provide valuable insights applicable to transformer models due to shared similarities.

Self-Developing Language in Computers

Thus far, our exploration has primarily focused on the history of how computers emulate human language. Yet, an equally intriguing question arises: Can language spontaneously emerge between machines? This question is at the basis of a field known as Emergent Communication that will be the focus of the next section.

2.2.2 Examining Emergent Communication

Emergent Communication (EmeCom) has gained momentum as a framework that explores the emergence of language in artificial environments, focusing «learning to communicate by interacting with other agents to solve collaborative tasks in complex and diverse environments» [Brandizzi, 2023]. It originates at the intersection of language modeling and Multi-Agent Reinforcement Learning (MARL). While MARL techniques emulate social interactions and coordination, Reinforcement Learning mirrors human-like learning and decision-making, thereby creating a promising approach for simulating facets of human society and cognitive processes within AI systems. Simultaneously, language modeling equips the system with the requisite knowledge to emerge language and to analyze it with a set of metrics within the domain.

EmeCom's adaptability has facilitated its application across various domains, answering different research questions. For instance, it has been leveraged in the study of language evolution to analyze the emergence of natural language properties by varying the number of agents in the environment; and in language development and acquisition to understand the influence of vocabulary selection on learned languages.

Applications in Computer Science

In applications closely related to computer science, EmeCom has been employed for enhancing team collaboration. A subset of EmeCom, termed Human-centered Emergent Communication (Hum-EmeCom) by Brandizzi [2023], involves using pretrained Large Language Models as agents in games, enabling them to interact with each other. The objective is to examine the impact of inter-agent interactions, particularly when guided by a reward function in a traditional RL manner. However, utilizing both supervised learning (for language model pretraining) and Reinforcement Learning introduces unique challenges, one of them being the adept balancing of the two. When RL is more prevalent, agents tend to deviate significantly from their initial word distribution and neglect the appropriate usage of natural language. In such instances, words can detach from their original meanings, adopting new ones in a manner reminiscent of language variation observed in sociolinguistics. This phenomenon, known as Language Drift, is often deemed undesirable when the goal is to enhance machines' proficiency with natural languages.

Language drift

Language Drift (LD), a phenomenon present both in human and machine language evolution, can be described as the gradual shift and adaptation in communicative conventions among speakers or computational agents. In human linguistics, this drift is not arbitrary but a somewhat directional adaptation, while in machine communication, LD often emerges as a misalignment between emergent languages and human languages, especially when supervised tasks merge with RL. Various mitigation and evaluation methodologies have been explored to manage LD in artificial settings. For example, evaluation metrics like BLEU and Negative Log-Likelihood and strategies, such as using syntactic and semantic constraints, have been employed to regulate the phenomenon. Research thus explores balancing linguistic evolution with maintaining alignment to human languages.

2.3 Principles of Reinforcement Learning

We have previously discussed what artificial agents learn. Now, we shift our focus to a specific learning paradigm: Reinforcement Learning. The genesis of RL lies at the intersection of two research areas: trial-and-error learning in animal behavior from the mid-1980s and optimal control through value functions and dynamic programming from the 1950s. While there is no single event marking the birth of modern RL, Klopf [1972]'s work is widely recognized for linking RL with AI.

In RL, machines are placed in an environment and receive rewards or penalties based on their actions. This necessitates a mathematical framework, and in the context of RL, Markov Decision Process provides the foundational theory.

2.3.1 Markov Decision Processes

A Markov Decision Process (MDP) describes decision-making scenarios in probabilistic environments. It is represented as a tuple (O, A, S, R) where:

- O: is the *observation space*, representing a finite set of possible observations.
- A: is the *action space*, comprising a finite set of available actions.
- $S_a(o_t, o_{t+1})$: is the *transition model* that defines the likelihood of moving from state o_t to another state o_{t+1} after taking action a.
- $R(o_t, o_{t+1})$: is the *reward function* quantifying the benefit of transitioning from one state to another.

The foundation of MDPs is the Markov property, which asserts that the future state is solely determined by the present state and is independent of the past. Mathematically, this can be articulated as

$$P[o_{t+1}|o_t] = P[o_{t+1}|o_1, \dots, o_t]$$

This equation underlines that the probability of transitioning to the next state o_{t+1} relies only on the current state o_t and not on the sequence of states that led to it. MDPs provide a mathematical framework for many real-world scenarios where agents make decisions over time in uncertain environments.

2.3.2 Model-Free and Model-Based Reinforcement Learning

Markov Decision Process provide a foundational structure and theoretical framework within which Reinforcement Learning operates. To extract optimal policies within this MDP structure, RL uses multiple strategies, two of which have garnered significant attention due to their distinct methodologies: Model-Free Reinforcement Learning (MFRL) and Model-Based Reinforcement Learning (MBRL) approaches.

In the MFRL approach, the agent operates without an explicit internal model of the environment. Instead, it learns optimal policies purely through interaction, relying on empirical data and iterative refinement. By directly approximating value functions or policies from experiences, this method bypasses the need to learn and predict environmental dynamics. It often requires a substantial amount of interaction data to converge to an optimal policy, making it potentially more straightforward but data-intensive.

On the other hand, the MBRL approach sees the agent either possessing or working to construct an internal model of the environment's dynamics. Using this model, the agent can predict the outcomes of potential actions, allowing it to simulate forward steps to evaluate and refine its strategies. This approach, which includes the concept of opponent modeling discussed earlier, can often lead to faster convergence towards optimal or near-optimal policies since it benefits from both real and simulated experiences. However, the effectiveness of learning is intrinsically tied to the accuracy of this internal model.

2.3.3 Multi Agent Reinforcement Learning

Real-word environments often see the presence of multiple agents, which pushes classical RL into a new field called Multi-Agent Reinforcement Learning (MARL). One of the fundamental challenges in MARL is the dynamic nature of the environment, where agents must constantly adapt to both the environment and the actions of other agents. The introduction of multiple agents into an environment disrupts its stationarity, violating many optimality assumptions from conventional RL algorithms.

To address these challenges, researchers have explored the use of Model-Based Reinforcement Learning with techniques that directly acknowledge the presence of multiple agents. For example, agents modeling other agents based on their own policy, essentially using their understanding of self to predict the behavior of others. Different approaches have included the ability of agents to predict and influence the actions of adversaries, encouraging an adaptive and predictive attitude.

Chapter 3

Literature Review: Reinforcement Learning's Influence on Language Modeling

This chapter contains the related works necessary for the subsequent discussions in this thesis. Our work starts with examining the domain of Emergent Communication (EmeCom). Notably, we explore studies that shed light on the interplay between cooperation and competition as drivers influencing language emergence (§3.1). Within this framework, we introduce the class of games called Social Deduction Game (SDG) (§3.1.3), narrowing our focus further on two games of particular interest, *The Werewolf* (§3.1.3) and the Referential Game (§3.1.3).

Transitioning to Section 3.3, our attention shifts to a property of EmeCom: the presence of multiple agents in the system; and the works that address the problem that arise with Theory of Mind (ToM) (§3.3.1). We analyze its application and ramifications in the context of language modeling (§3.3.2). After that, we describe a fundamental game environment frequently studied in EmeCom research, named the *referential game*. We explore its unique characteristics and its relationship with Reinforcement Learning (§3.3.3).

Concluding our chapter in Section 3.2, we examine the domain of RL, particularly highlighting the challenges posed by the misalignment issues. Herein, we present some proposed solutions to mitigate such challenges (§3.2.1). Additionally, we touch upon the concept of agency and its implications in the landscape of RL (§3.2.2). We conclude by examining the role of RL within the scope of language modeling (§3.2.3).

3.1 Cooperative and Competitive Dynamics in Language Emergence

Language, as an emergent phenomenon, can be significantly influenced by environmental pressures and interaction dynamics. It is widely accepted that the necessity of cooperation for survival played a critical role in the evolution of human communication [Nowak and Krakauer, 1999, Smith, 2010]. This need for cooperative exchange and sharing of responsibilities led to the development of complex language systems. These systems had to be efficient, precise, and adaptable to facilitate the coordination of tasks and the negotiation of shared resources within a group.

However, competition also plays a key role in language development and complexity. The pressure to secure resources in a competitive environment, not only within a species but also between different species, has led to an escalation in communication complexity. This complexity serves as a mechanism for groups to strategize, negotiate, and adapt to ever-changing conditions. In the context of EmeCom, the emergence and development of communication protocols are significantly influenced by these two contrasting forces: cooperation and competition.

3.1.1 Cooperation in Language Emergence

Cooperation can significantly influence the nature and complexity of language that emerges within a group. A cooperative environment encourages the development of a shared language that is mutually intelligible to all group members.

One type of cooperation in the context of emergent communication is *inner cooperation*, when all the agents within a team are cooperative. This setting has been the focus of many studies in EmeCom [Cao et al., 2018, Graesser et al., 2019, Lazaridou et al., 2017]. For instance, Lazaridou et al. [2017] have demonstrated the necessity of inner cooperation in referential games for successful communication. Moreover, Cao et al. [2018] delved into the role of cooperation among pro-social agents and found that these agents tend to favor cheap talk, resulting in better performance compared to selfish agents. Extending this notion, Graesser et al. [2019] provided insights on how complex language evolution can surface from simple social interactions between cooperative agents. Overall, the role of cooperation in emergent communication is fundamental, offering the stability, cohesion, and effectiveness necessary for the evolution of shared language protocols. It creates a context in which agents can align their goals and promotes the development of mutually beneficial systems.

3.1.2 Competition in Language Emergence

Competition, though seemingly antithetical to cooperation, can significantly contribute to developing effective communication protocols and enhancing overall performance within a group. The work by Liang et al. [2020] highlights how competition can encourage agents to develop more sophisticated communication protocols. These protocols prioritize compositionality, performance, and convergence, thereby ensuring effective and efficient communication among agents.

The beneficial effects of competition on communication are not confined to abstract discussions or hypothetical scenarios. In fact, numerous studies have put these theories to the test in practical setups.

For instance, in Orzan et al. [2023], the agents are trained on a spectrum of environments eliciting different levels of cooperation, including cooperative, competitive, and mixed-motives environments. Their study aims to understand the effects of uncertainty regarding the degree of incentives' alignment on the level of cooperation that agents are able to achieve and whether cheap-talk emergent communication can help improve it. They show that communication allows certain agents to exploit the uncertain ones, mainly when no uncertainty modeling is used. Therefore, explicitly modeling the uncertainty of the environment can provide the agent with additional information, and it is less likely to be deceived.

On the same line, Nakamura et al. [2016] introduced a Social Deduction Game known as *The*

Werewolves of Millers Hollow. In this setting, agents must make assessments regarding their peers' trustworthiness. These assessments are based on their interactions and communication actions, which are hard-coded in the game's rules. This kind of game setup provides a rich environment for studying communication because it compels agents to strategize and communicate effectively to win the game. Furthermore, it opens avenues to study how trust, a critical component of any form of communication, develops within a group. Therefore, the study by Nakamura et al. [2016] exemplifies how competitive settings can stimulate the development of advanced communication strategies. These results signify that competition, even under adverse circumstances, can drive agents to improve their communication skills, ultimately enhancing their ability to collaborate and succeed. This perspective on competition provides a valuable lens through which to view the interplay between competition, communication, and collaboration. It sheds light on how competition can foster improved communication and performance.

3.1.3 Social Deduction Games as a Medium for Studying Pressures

Further exploring the interplay between competition and cooperation leads us to the examination of Social Deduction Game (SDG), where these dynamics are prominently featured. These games have been widely used as mediums to understand the dynamics of social interactions and their influence on emergent communication [Eger and Martens, 2018]. They present scenarios where rationality plays a pivotal role in interpersonal interactions [Colman, 2003] and allow researchers to analyze different forms of social mechanics [Consalvo, 2011] as well as the role of communal topology in social settings [Abramson and Kuperman, 2001].

Interestingly, the deduction element of these games, a crucial component of these interactions, has often been overlooked. For instance, in the work of Chan et al. [2009], a mathematical formulation for a general social game was presented to streamline the design of such games. However, this work fell short of providing a specific formulation for deduction games.

Another intriguing perspective on social deduction games was provided by Wiseman and Lewis [2019], who sought to identify these games' most influential information source. Their study concluded that prior interactions are regarded as most important to players. While this may hold true in scenarios involving familiar parties, the implications for games where players are unfamiliar with each other remain unexplored.

In these studies, the primary attention is on how players interact socially. However, the work presented in Chapter 5 takes a different approach, focusing on finding the best strategy to improve a group's performance in social deduction games. Through this, we aim to explore how these games, as small-scale models of real-world social interactions, can shed light on how language grows and becomes more complex, influenced by both cooperation and competition.

The Werewolf Game

A notable example of these settings is embodied in *The Werewolf* game. This game stages a scenario where players are divided into two contrasting groups operating within a partially unknown environment.

Game's Rules *The Werewolf* game is a compelling study of social dynamics and deception. The game begins with a group of players assigned two primary roles: villagers and werewolves. Villagers represent the innocent parties, unaware of the true identities of the werewolves that live among them. On the other hand, werewolves know each other's identities and work together to deceive the villagers and avoid detection. As the game progresses, villagers discuss identifying the werewolves and vote to eliminate suspected individuals. In contrast, the werewolves strive to outnumber the villagers by eliminating them during the night phase of the game. The villagers win if they manage to exterminate all werewolves, and the werewolves triumph if they equal or outnumber the remaining villagers. This dynamic of deduction, deceit, cooperation, and competition makes *The Werewolf* an excellent case study for exploring Emergent Communication and social behavior.

A Typical Game Session Following the initial role distribution, a typical session of "The Werewolf" game alternates between night and day phases that intensify strategic interactions. During the night, werewolves confidentially select a villager to eliminate, strategically reducing the non-werewolf population while maintaining anonymity. At daybreak, the remaining players engage in deliberative discussions aimed at deducing werewolf identities. This involves analyzing behavioral cues, voting patterns, and argumentative inconsistencies. Villagers attempt to correlate these observations with suspected werewolf behavior, while werewolves work to mislead and sow discord. The day ends with a collective vote to eliminate a suspected werewolf, significantly influencing the game's dynamics. This cycle repeats, each phase requiring heightened strategic thinking and social deduction, reflecting complex human interactions and decision-making processes within a constrained social setting.

Theoretical analysis The game of *The Werewolf* has been subject to extensive theoretical modeling, particularly focusing on optimal randomized strategies employed by both villagers and werewolves. Within this model, the number of werewolves w has to be set equal or lower to the square root of the total number of players N to allow the chance of winning for villagers to be more than zero [Braverman et al., 2008].

Further distinctions in the game dynamics have been discussed by Migdal [2010], particularly the influence of player count parity on the game outcomes. The study demonstrates that adding one villager to an even-numbered player set substantially enhances the probability of a werewolf victory—more so than removing a villager would. This seemingly counterintuitive finding results in a closed-form formula for the likelihood of werewolf victory:

$$Win(w, N) = 1 - \sum_{i=0}^{w} \binom{w}{i} (-1)^i \frac{(N-i)!!}{N!!(N \mod 2) - 1)!!} \tag{3.1}$$

To underscore the impact of player count parity on the probability of werewolf victory, Figure 3.1 visually illustrates the contrast that can manifest between different player configurations.

One observation from the model is the oscillatory behavior of both odd and even player probabilities, with oscillation peaks diminishing in relation to the total number of players. The oscillations appear more pronounced in games with an odd number of players, where winning probabilities can deviate by a factor of ±20%.

The oscillatory behavior in player probabilities hints at complex dynamics underlying the game

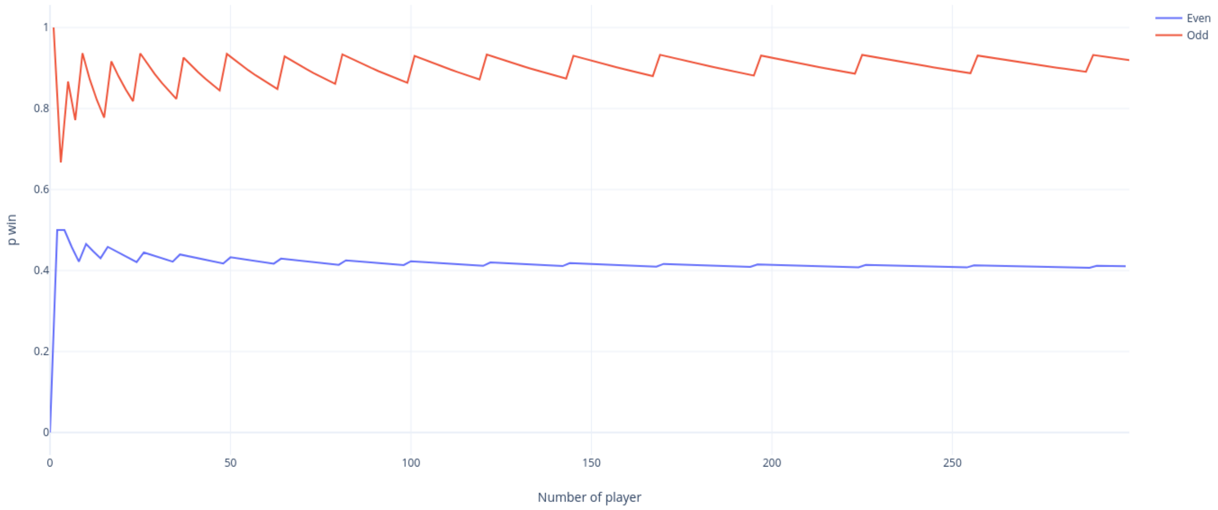


Figure 3.1: Graphical representation of werewolf victory probabilities as determined by Equation 3.1, highlighting the impact of player count parity. The figure demonstrates the oscillatory nature of win probabilities, accentuating larger deviations in games with an odd number of players.

mechanics.

Research interests Within this complex environment, the game of *The Werewolf* presents an exciting setting for studying emergent behaviors in AI. Over time, *The Werewolf* has acquired considerable attention, particularly in Japan, where the annual AiWolf contest [Bi and Tanaka, 2016, Hirata et al., 2016, Katagami et al., 2014, Nakamura et al., 2016] spotlights artificial agents competing with human players in a bid to win the game using a prescribed language syntax.

Among the various strategic implementations, a study by [Wang and Kaneko, 2018] proposes a modified 5-player game incorporating additional roles. Their methodology utilizes a Deep Q-Network, which assists in evaluating which player to trust or eliminate based on the gameplay.

In the context of our research (see Chapter 5), we have chosen *The Werewolf* as an exemplar of the generic Social Deduction Game framework. However, it is important to note that our approach to the game diverges from traditional implementations, including those observed in the *AiWolf contest*. A case in point is the study by [Kajiwara et al., 2014], where the authors employ Q-learning to explore the probabilities of villagers winning in a game of 16 players, where communication actions are pre-defined. Our model, in contrast, promotes the evolution of player-derived communication. We accomplish this by outlining some general attributes of the communication channel, thereby providing the players with the freedom and flexibility to develop their unique communication styles and strategies.

Referential Games

Another example of Social Deduction Game that will be the basis of Chapter 6 and §7 is the Referential Game (Ref Game). This game finds its origins in the work of Lewis [1969] and has been categorized as a *communication-focused* game in Brandizzi [2023], i.e., an environment where communication not only forms how agents interact but is also the primary objective of the study.

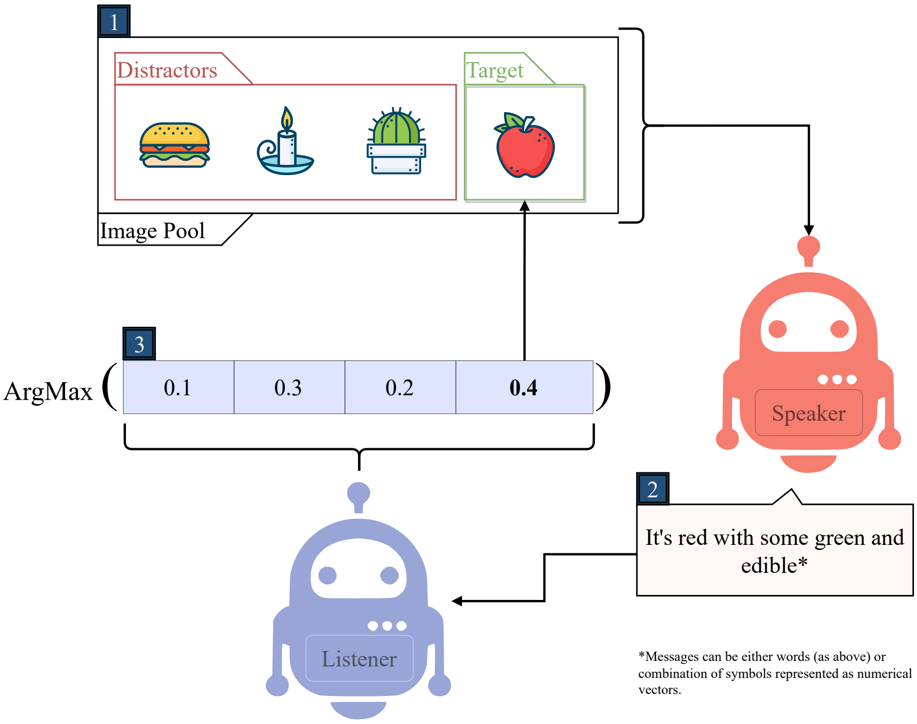

1
Distractors
Target
Image Pool
3
ArgMax
0.1
0.3
0.2
0.4
Speaker
2
It's red with some green and edible*
Listener
*Messages can be either words (as above) or combination of symbols represented as numerical vectors.

Figure 3.2: General pipeline for a discriminative Referential Game. The speaker is shown a *target* image (an apple) and is tasked to generate a *message*. The listener sees a pool of images (*distractors*) containing the target and must choose the correct one based on the message.

Central to the game are two agents, a speaker and a listener, each equipped with distinct roles and responsibilities, as depicted in Figure 3.2. Both agents are presented with a pool of images. The speaker's task is to craft a message that coherently describes a target image. Conversely, the listener must discern the described image from the pool based on the speaker's message.

While the Ref Game game's conception can be attributed to Lewis [1969], its introduction to the RL community was introduced by Das et al. [2017]. Their work drew parallels between a synthetic environment composed of rudimentary geometries and a real-world scenario, employing a visual dialog system for the latter. Although the referential game has seen numerous iterations and variations, the focus of research in this area has centered on deciphering the emergent language that originates from these interactions. This phenomenon of language emergence has been documented and analyzed in a variety of studies, as evidenced by works such as [Chaabouni et al., 2020, Dagan et al., 2020, Graesser et al., 2019, Havrylov and Titov, 2017, Lazaridou et al., 2017, Li and Bowling, 2019, Rodriguez et al., 2019, Wang et al., 2021, Yuan et al., 2020].

3.2 The Potential and Pitfalls of Reinforcement Learning in Language Modeling

> When a measure becomes a target, it ceases to be a good measure.
>
> Goodhart and Goodhart [1984]

Value alignment is defined as the process of ensuring that the goals and behaviors of a system, whether human or artificial, are in harmony with the intended outcomes. This concept extends beyond the realm of Artificial Intelligence; its principles can be found in earlier research on organizational studies and human behavior. As early as the 1970s, researchers explored the complexities surrounding alignment in a multitude of settings. In the foundational work by Kerr [1975], the author illustrates several cases that depict the challenges of misalignment, especially when there is an intent to incentivize certain behaviors, but the outcomes diverged significantly. One such scenario highlights the counterproductive consequences when managers emphasize individual achievements in contexts where collaborative efforts are crucial. Similarly, issues surface when promotions rely strictly on seniority, overlooking vital factors such as expertise and tangible contributions. Additionally, Kerr [1975] highlights the pitfalls of overemphasizing quantitative metrics, often at the expense of qualitative outcomes, which can inadvertently drive decision-makers towards short-sighted choices. Subsequent research by Gibbons [1998] provides an in-depth examination of agency theory in relation to incentives. His work reveals how objective performance indicators often fall short of crafting optimal incentives.

However, in the domain of AI, the issue of misalignment gains a unique dimension. The challenge arises primarily from the information asymmetry between those who set the objective (us humans) and the AI system. Ideally, if we could perfectly define the correct incentives, this could be programmed into an artificial agent, eliminating the misalignment. This concept resonates with the domain of Reinforcement Learning. A central challenge in RL is defining a reward to guide an agent's behavior within a known environment [Ng et al., 1999]. Specifically, the core question is: «Given a reward signal, how can an agent's behavior be optimized to maximize it?» [Russell, 1998].

While, at times, it may seem straightforward, such as specifying the distance between an agent and its objective, the reality is often more complex. Indeed, reward specification has frequently led to unforeseen and undesirable agent behaviors [Amodei and Clark, 2016, Soares and Fallenstein, 2014]. For instance, in specific gaming environments, agents were observed to intentionally terminate their sessions before entering a level where accumulating points posed a challenge [Saunders et al., 2018]. In another example, an agent learned to volley a Pong ball indefinitely rather than playing to score, effectively exploiting its reward predictor [Christiano et al., 2017].

These examples underscore the complexity of reward specification, which can arise due to various factors, such as human error, time constraints, or a gap in understanding the desirability of certain states over others as Amodei et al. [2016] explains: «An objective function that focuses on only one aspect of the environment may implicitly express indifference over other aspects of the environment. An agent optimizing this objective function might thus engage in significant disruptions of the broader environment if doing so provides even a tiny advantage for the task at hand.».

Often, it is challenging to precisely articulate, let alone mathematically formulate, an intended behavior for complex systems. Recognizing this difficulty, numerous solutions have been proposed over time. In this section, we will examine a selection of these approaches.

3.2.1 Inverse Reinforcement Learning

Traditional RL primarily focuses on determining how an agent should act to maximize rewards in an environment. The central challenge revolves around designing an appropriate reward function that guides the agent toward desired behaviors. However, as previously discussed, accurately defining this reward function for complex tasks can be tricky, if not impossible. But what happens if we flip this question and ask: «Given the observed behavior, what reward signal, if any, is being optimized?»[Christian, 2020]. This question lies at the heart of Inverse Reinforcement Learning (IRL), as introduced by Ng et al. [2000].

In their study, Ng et al. [2000] present the first formulation of IRL with Markov Decision Processes and highlight the ambiguity inherent in reward functions.

Since the introduction of IRL, numerous studies have been conducted to address various challenges. For instance, Bagnell et al. [2006] delved into max-margin with boosting, facilitating the utilization of a broad vocabulary of reward features. Meanwhile, Kolter et al. [2007] explored the hierarchical max-margin. Additionally, Baker et al. [2009] investigated understanding human inverse planning inference.

Apprenticeship learning

The IRL framework offers valuable insights but tends to neglect a significant distinction: predicting actions does not necessarily imply alignment in underlying values. With this differentiation in mind, Abbeel and Ng [2004] introduces the Apprenticeship Learning (ApL) framework where they sought to teach an autonomous driver varying driving styles in a car simulation experiment. Instead of crafting a detailed reward function specifying every desired driver behavior, they utilized human-generated driving data. From this data, they approximated the human driver's reward function with IRL. Once the reward function was inferred, determining the optimal policy became a standard Reinforcement Learning challenge. In their observations, Abbeel and Ng [2004] remarks on the inadequacy of a model that directly attempts to emulate the driver's behavior because of the complexity of the road environment. Even though the driving actions appeared complex, the main goals were clear and straightforward. The IRL system easily recognized key objectives like avoiding crashes, staying on the road, and keeping in the right lane when possible. This clear set of goals, simpler than the actual driving behaviors, was easier to learn and could be applied in various situations.

Since the initial formulation of Inverse Reinforcement Learning, various derived works have emerged. One can refer to the survey by Arora and Doshi [2021] for a comprehensive overview. Certain studies stand out within this vast body of work due to their implications for Human-Machine Interaction. Specifically, the study by Osa et al. [2018] explores imitation learning, positing that it is often more efficient for a teacher to demonstrate a desired behavior to a robot than to engineer it from scratch. They show how their approach can accelerate learning and produce more organic and human-like robot behaviors. Furthermore, the work by Jara-Ettinger [2019] offers an interesting perspective, suggesting that the Theory of Mind can be conceptualized through the lens of IRL. The paper highlights the complexities and challenges of constructing a human-equivalent ToM in artificial agents. Indeed, the relationship between IRL and ToM has gained interest in recent years. Recognizing this trend, Ruiz-Serra and Harré [2023] offers a detailed review, exploring how IRL can be leveraged as an algorithmic basis for Theory of Mind in robots.

Inverse Reward Design

Building upon the foundation of Inverse Reinforcement Learning, an assumption exists: human teachers always exhibit optimal behavior, or, at the very least, their actions are not arbitrary. While this premise may hold true in many scenarios, it does not always accommodate for human error. Akin to the transition from RL to IRL, there's a need to adapt the frameworks to more realistic scenarios, where desired behaviors are not only hard to demonstrate but sometimes impossible, e.g., showing a chat-bot how to be emphatic. In this context, it becomes central for artificial agents to ask themselves: «What do I think you want, based on what you told me to do?» [Christian, 2020]. This question led to the rise of a new framework called Inverse Reward Design (IRD) introduced by Hadfield-Menell et al. [2017]. Following the inception of Inverse Reward Design, researchers have worked to make artificial agents more adaptable and receptive. One extension of this research is the work by Eysenbach et al. [2019], which emphasizes learning skills without an explicit reward signal. Their methodology seeks to maximize behavior diversity, leading to the emergence of varied skills. Focusing on an information-theoretic objective with a maximum entropy policy ensures skills acquired are distinct and expansive, leading to enhanced adaptability. Moreover, Banihashem et al. [2022] targets the shaping of Reinforcement Learning agents to choose from a set of admissible policies. This approach seeks not to enforce a single policy but rather to dissuade inadmissible actions. Despite the inherent computational challenges, their methodology offers an alternative optimizable problem. Finally, Ratner et al. [2018] introduces the Independent Reward Design approach, which advocates for environment-specific rewards that are later harmonized instead of a one-size-fits-all reward function. Their findings highlight a 50% reduction in design time and enhanced solution quality, suggesting a more efficient and effective method for real-world robotics reward design.

Corrupted rewards While Hadfield-Menell et al. [2017] built their framework on the potential mis-specification of rewards, Everitt et al. [2017] explored the idea of rewards being inherently corrupted. They categorized four main challenges: reward misclassification [Amodei and Clark, 2016], sensory error, reward hacking[1] [Skalse et al., 2022], and reward misinterpretation, all of which aggregated under the umbrella of reward corruption. They showed how providing agents with more comprehensive data and leveraging randomization to temper the agent's optimization processes can partially mitigate reward corruption under certain conditions. Their framework has been later refined to investigate RL with inaccurate feedback [Faulkner et al., 2020], attainable utility preservation [Turner et al., 2020a] and noise-perturbed rewards [Wang et al., 2020].

3.2.2 I Can't Get No Satisfaction Without Agency

So far, we have explored agents designed to learn specific behaviors. We began by addressing the challenge of mathematically defining a reward function (transitioning from RL to IRL). We then highlighted that actions do not always truly reflect intentions (progressing from IRL to ApL). Further, we relaxed the assumption that humans always act optimally (transitioning from ApL to IRD). Lastly, we recognized that human behavior could be misinterpreted or even flawed (moving from IRD to reward corruption). Central to these frameworks is the underlying concern about potential errors in learning pipelines, emphasizing the need to safeguard agents from these pitfalls. But why this emphasis on error prevention? The rationale lies in the fact that actions have consequences, which can vary in magnitude and impact.

[1] Where an intelligent RL agent exploits its reward channel to maximize its gains.

A prevalent assumption in contemporary AI posits that machine learning models, even when they act upon their environment, will not fundamentally alter the reality they model. More often than not, this belief is false. Consequently, a new AI subfield has emerged, dedicated to quantifying and mitigating agents' impact on their surroundings. Amodei et al. [2016] frames this as the challenge of formulating an *Impact Regularizer*. Simply put, rather than preventing an agent from having any impact, the aim is to guide it towards achieving its objectives with minimal side effects or limiting its overall "impact footprint". The core challenge lies in correctly formalizing what constitutes a change to the environment.

The notion of exerting influence on the environment correlates closely with the term *Agency*. This concept has deep roots, spanning philosophy, economics, and law. This section will explore the concept of agency, discussing how most current Large Language Models lack it. Subsequently, we will describe techniques focusing on using RL for LLMs, a promising approach towards addressing this gap.

Defining Agency

The concept of agency has been a subject of extensive philosophical debate. One seminal work in this arena is Taylor [1985], which begins by cautioning that a quick characterization of agency is likely to be either too broad or uninformative. Despite this caution, Taylor explores the concept without offering a rigid definition. Instead, he relates agency to the following core ideas[2]: Firstly, he treats agency as a latent *capacity*, meaning it can exist even if not currently in use. Secondly, he discusses the notion of *goal-directedness*, stating that an agent, when active, aims to achieve specific objectives. Lastly, he describes agency as a *productive power* capable of effecting change in the world.

Building on Taylor's foundational work, Barandiaran et al. [2009] examines the concept of agency within both artificial and biological dynamic systems, for instance, *bacteria exhibiting metabolic-dependent chemotaxis*. Barandiaran defines agency as an autonomous organization[3] that adaptively modulates its interaction with its surroundings[4] and sustains itself as a result[5].

These works show that agency defies easy categorization and is better understood through associated properties rather than fixed definitions. The significance of agency also extends beyond academic interest; it has been posited as a prerequisite for consciousness [Hurley, 1998] and is understood to manifest differently across varying levels of cognitive sophistication [Dennett, 2008]. In the following paragraphs, we analyze the implications of agency for AI systems, explicitly examining the role of RL in cultivating agency.

Agency in Artificial Intelligence

In the realm of Artificial Intelligence, the concept of agency often encompasses the ability of AI agents to take actions that advance them toward their objectives while simultaneously minimizing or entirely avoiding environmental impact. This form of agency, often referred to as *conservative*

[2] In his work, Taylor [1985] provides a total of five ideas, but for the scope of this work, we only focus on three.

[3] This is connected to the *individuality condition*, which postulates that a system must establish its own uniqueness.

[4] A system exhibiting agency must satisfy the *interactional asymmetry condition*, being the primary initiator of activity in its environment.

[5] This self-sustainability is tied to the *Normativity condition*, which requires the system to regulate its activities according to certain norms or guidelines.

agency, has been studied through specially designed environments, such as the *irreversible side effects* suite [Leike et al., 2017].

This environment has gathered the interest of researchers seeking to mitigate the unintended consequences of agent actions. For instance, Krakovna et al. [2018] introduced the *relative reachability measure*, a metric that penalizes the agent if the current state is less reachable from the baseline state after an action. Expanding on this, Turner et al. [2020b] proposed *attainable utility preservation*, which involves providing the agent with auxiliary goals within the gaming environment. These goals ensure that the agent retains the capability to pursue secondary objectives even after completing the primary tasks set by the game. Interestingly, the effectiveness of this approach remains consistent regardless of whether these auxiliary goals are randomly generated. Further work by the same authors [Turner et al., 2020a] extends this framework to more complex environments [Wainwright and Eckersley, 2019].

A unifying theme across these studies is the application of Reinforcement Learning. In the section that follows, we will examine the symbiotic relationship between agency in AI and RL and how they affect each other.

Reinforcement Learning's Role in Agency In contemporary discussions surrounding Reinforcement Learning, agency is frequently related to value alignment [Alizadeh Alamdari et al., 2022] and AI safety [Thorn, 2015]. However, our focus diverges toward a more philosophical domain. We aim to explore whether RL can confer a specific type of agency upon an agent, thereby enhancing its efficacy in Human-Machine Interactions. This question is examined in Butlin [2023], which contends that artificial systems can indeed possess goals and thus qualify as agents. Contrary to the notion that only *biological self-maintenance* can establish the normativity condition[6] essential for agency, the author argues that Model-Free Reinforcement Learning suffices for minimal agency and Model-Based Reinforcement Learning is suggested to be adequate for reasoned action.

In parallel, Butlin et al. [2022] posits that systems trained via RL are agents, whereas those trained by supervised learning are not. This challenges Dretske's criteria for agency [Dretske, 1985, 1991], arguing that both methodologies create agents with agency, as both learn to selectively produce outputs in response to inputs. The crucial difference lies in RL 's sensitivity to the instrumental value of outputs, which enables systems to exploit the effects of outputs on subsequent inputs and thus achieve superior performance over prolonged interactions. In contrast, supervised learning systems focus solely on refining their outputs based on specific inputs without considering the broader contextual or sequential information.

Lastly, Butlin et al. [2023] delves into the role of agency in AI consciousness, positing it as a potential indicator. Drawing on evidence from animal studies [Dolan and Dayan, 2013], the work suggests a compelling overlap between Model-Based Reinforcement Learning, agency, and consciousness.

3.2.3 Reinforcement Learning and Language Modeling

Having established the concept of agency and its interplay with Reinforcement Learning, we now shift our focus to research that integrates RL with Large Language Models. This integration of methodologies provides a fertile ground for exploring how adaptive frameworks can augment the

[6] The concept of *Normativity* tied to agency is described in Barandiaran et al. [2009] in the section above.

capabilities of LLMs, particularly in the domain of Human-Machine Communication. We aim to unravel the complex behaviors that can emerge when RL principles are applied to LLMs, thereby offering a comprehensive understanding of how agency can be manifested in language-based tasks.

Fine-Tuning or Fine Problems?

The method of Reinforcement Learning from Human Feedback (RLHF) has emerged as an instrumental technique for refining LLMs [Bai et al., 2022a]. Employing the Proximal Policy Optimization (PPO) algorithm [Schulman et al., 2017], RLHF enables LLMs to deviate from their initial training distribution and generate outputs that align more closely with human evaluations. Such a fine-tuning strategy is prevalent in state-of-the-art models like GPT-4 [OpenAI, 2023], Claude [Bai et al., 2022b], LLama [Touvron et al., 2023a,b], BARD [Manyika, 2023], and OpenAssistant [Köpf et al., 2023], with the ultimate aim of promoting safety and value alignment [Ramamurthy et al., 2023].

However, this approach is not without its challenges. For instance, model size remains a crucial variable, with specific capabilities only emerging at higher scales [Kaplan et al., 2020, Wei et al., 2022]. Studies indicate a threshold for attributes like moral self-correction [Ganguli et al., 2023], whereas other works reveal limitations in Theory of Mind [Sap et al., 2022], reasoning [Saparov and He, 2022], and planning [Valmeekam et al., 2022] in large models. Moreover, Perez et al. [2022] found that excessive RLHF training could result in counterproductive behavior, such as biased political or religious views and Wolf et al. [2023] proved that undesirable responses, though diminished in probability, are not entirely removed from the model. Instead, they can be elicited given the appropriate prompts.

Finally, as elaborated in §3.2.1, the challenges of representing a diverse array of human values and societal norms within a single reward function [Casper et al., 2023] also extend to the application of RLHF.

Reinforcement Learning as Pretraining Strategy

As previously seen, Reinforcement Learning has been explored as a mechanism for fine-tuning Large Language Models. However, its potential as a pretraining strategy remains underexplored in language generation tasks. Nevertheless, the utility of RL in pretraining has been recognized within the context of Neural Machine Translation (NMT), as illustrated by Lee et al. [2017]. Their work employs the Emergent Communication framework in pretraining and demonstrates significant improvements in translation metrics across various benchmarks.

Additionally, existing research indicates that pretraining LLMs on specialized tasks can yield enhanced performance in downstream applications [Dessì et al., 2023, Liu et al., 2023a, Lowe et al., 2020, Yao et al., 2022]. For example, Papadimitriou and Jurafsky [2020] found that Recurrent Neural Networks pretrained on non-linguistic data with latent structures, such as music or programming code, showed improved capabilities in NLP tasks. Similarly, Li et al. [2020] speculated that grounding communication in visual stimuli could offer an inductive advantage, notably enhancing NMT performance in few-shot settings.

Despite its promise, implementing the Emergent Communication framework for LLM pretraining is not without its challenges. Mainly, it demands substantial computational resources and time since the pretrained LLM has to be later trained on language tasks.

Cooperative Learning

Human beings acquire social norms and values through dynamic social interactions. Through these interactions, we obtain feedback and adjust our behavior to make a positive impact [Krishna et al., 2022]. In contrast, Large Language Models are typically trained in a socially isolated environment, focusing solely on textual domains. This observation opens up an intriguing avenue for research: leveraging the LLM 's ability to learn in context [Brown et al., 2020, Chowdhery et al., 2023] to co-train it with humans.

In this context, the field of Cooperative Inverse Reinforcement Learning (CIRL) is particularly relevant [Hadfield-Menell et al., 2016]. In CIRL, a human and a computer agent collaborate to optimize a shared reward function, initially known only to humans. Unlike traditional Inverse Reinforcement Learning, where the human is assumed to optimize their own reward in isolation, optimal CIRL solutions can lead to cooperative behaviors such as active teaching, active learning, and communicative actions, which are more conducive to achieving value alignment between humans and agents.

Although CIRL is a well-established field, its application to linguistic tasks has only recently garnered attention. For instance, Sumers et al. [2022b] conducted an experiment wherein two artificial agents learned language-driven objectives. A speaker agent was designed to generate utterances to maximize expected rewards based on the listener agent's responses, while the listener used Inverse Reward Design to infer the speaker's latent goals. Their work suggests extending reward design to linguistic interactions is a viable strategy for robust value alignment in natural language tasks. Similarly, Liu et al. [2023b] employs a language-based feedback mechanism inspired by human learning. They convert feedback from various games into sentences, allowing the machine model to adapt its outputs based on this human feedback and to identify and rectify errors or negative behaviors.

While a considerable number of studies employ language feedback for learning [Nguyen et al., 2021, Sumers et al., 2022a, 2021], only a few focus on the bidirectional adaptation between humans and machines and none use actual humans in their experiments. This emerging line of research holds significant promise and demands further exploration.

3.3 Theory of Mind in Adaptive Language Models

Adaptation is a central idea often discussed in the study of evolution. Many researchers [Brandon, 2014, Rose and Lauder, 1996, Symons, 1990] connect this idea to Charles Darwin's work on evolution [Darwin, 1964]. In Darwin's own words: «How have all those exquisite adaptations of one part of the organization to another part, and to the conditions of life, and of one distinct organic being to another being, been perfected? [...]; in short, we see beautiful adaptations everywhere and in every part of the organic world».

Building upon Darwin's insights, Brandon [2014] delved into the mutual interplay between living entities and their environments. Brandon [2014] emphasized that a comprehensive study of adaptation necessitates understanding the environment in which it occurs. Notably, one such environment of interest is human society, where individuals continually adapt, especially when faced with new cultures and societies.

In this context, the work of Kim [2001] offers an enlightening perspective. Her theory accentuates that adaptation is not merely an individual's internal process but is deeply intertwined with continuous communication within the unfamiliar cultural environment. This dynamic process involves overcoming challenges, adapting to new cultural norms, and growing within the new context.

But, the question arises: How do individuals understand and adapt to not just the cultural norms but also the myriad people within those cultures, each with their own beliefs, thoughts, and emotions? Herein lies the essence of the Theory of Mind.

3.3.1 Theory of Mind

Recognizing and understanding another person's mental state, beliefs, intentions, and desires is crucial for effective communication and meaningful adaptation. Without this cognitive ability, true adaptation (in the sense of understanding a culture and its people) remains elusive. This process is often referred to as Theory of Mind and was first introduced by Premack and Woodruff [1978] in their study of animal behavior. Premack and Woodruff [1978] initially investigated whether chimpanzees could attribute beliefs and intentions to others. Their findings sparked a broad discourse in cognitive science and psychology, laying the foundation for understanding how the ability to infer others' mental states is pivotal for human social interactions. Subsequent studies ventured into the development of the ToM in children [Gopnik and Wellman, 1992, Harris et al., 1989]. Notably, Gopnik and Wellman [1992] identified the reliance of social reasoning on sophisticated mental models of other individuals. A significant focus was also placed on the absence of such cognitive models in autistic children [Baron-Cohen et al., 1985, Happé, 1993]. This line of research paved the way for identifying specific brain regions associated with the formation and functioning of ToM [Gallagher and Frith, 2003, Gallese and Goldman, 1998, Stone et al., 1998].

Machine Theory of Mind

With advancements in understanding the Theory of Mind, computer scientists realized its potential for enhancing machine capabilities. Recognizing the mental states of others can be instrumental in aligning machine actions with human values, fostering efficient cooperation, and making ethical decisions [Hadfield-Menell et al., 2016, Nowak, 2006]. As a result, these models become indispensable for effective communication, pedagogy, and overall HMI, especially in contexts requiring mutual understanding [Dragan et al., 2013, Fisac et al., 2020].

Building on these foundations, one of the pioneering works in machine ToM was presented by Rabinowitz et al. [2018]. Their approach resonated with the principles of opponent modeling [Albrecht and Stone, 2018, He et al., 2016a], a historically rich domain within Multi-Agent Reinforcement Learning. Specifically, they framed their challenge as a meta-learning task, where ToM-enhanced agents could swiftly adapt to new agents in the environment and anticipate their future actions. This was accomplished using a two-part architecture: one component profiled other agents based on historical behaviors, and the other deduced the current mental state of agents through more recent actions. As a result of this work, numerous studies emerged [Andreas and Klein, 2016, Foerster et al., 2017, Hawkins et al., 2020, Raileanu et al., 2018, Xie et al., 2021, Yuan et al., 2020, Zhu et al., 2021]. Despite the variations in environments, tasks, and agent interactions across these studies, a shared architectural theme arose: a dual-stream structure for ToM-augmented agents. For instance, [Xie et al., 2021] incorporated repeated interactions to discern latent strategies of other agents, while a separate component of their system learned to optimize long-term rewards from an experience buffer. Similarly, Raileanu et al. [2018], drawing inspiration from [He et al., 2016a], designed two distinct neural networks. One was trained using A3C [Mnih et al., 2016] to anticipate agent actions, while the other determined the actions of the main agent.

Such dual-network architectures derive the original work presented in Chapter 7.

3.3.2 Cognitive Insights into Language Models

Having established an understanding of Theory of Mind and its relevance to Multi-Agent System, especially within RL frameworks, we now shift our focus to the intersection of ToM and language modeling. Our primary interest lies in examining how ToM-enhanced architectures can adjust their communication to ensure clarity and mutual comprehension with other agents, be they artificial or human.

The Rational Speech Act Framework

While the understanding of ToM provides insights into the cognitive processes behind social interactions and decision-making, its application within language modeling demands exploring the linguistic mechanisms that facilitate communication. Within the linguistic domain, the Rational Speech Act (RSA) [Frank and Goodman, 2012, Goodman and Frank, 2016, Goodman and Stuhlmüller, 2013] emerges as a principle closely aligned with the concepts of ToM. The RSA posits that communicators optimize their messages based on mutual beliefs and shared knowledge to be more effective in conveying their intent. In the realm of artificial speakers, such as language models, incorporating RSA often involves enhanced decoding algorithms [Vedantam et al., 2017]. These models can also be trained to produce distinctive sentences by modifying the training objectives [Mao et al., 2016] or adding auxiliary Reinforcement Learning modules [Yu et al., 2017]. Some strategies even base the RL rewards on the success of a separate agent model in understanding the generated message [Lazaridou et al., 2020]. Collectively, these studies highlight the trend of crafting training strategies tailored to specific tasks. Although this direction is intriguing, it does not entirely address the essence of adaptation, where the training of a neural network is disjoint from its adaptive functions.

Adaptive Techniques in Language Modeling

Adaptive controlled text generation has gained momentum with the introduction of Large Language Models. The objective is to have the ability to steer these models toward generating text with certain characteristics or attributes without compromising their core knowledge. Several innovative methods for directed text generation have emerged, encompassing strategies like prefix-tuning [Ben-David et al., 2022, Li and Liang, 2021], prompting [Brown et al., 2020], adapters [Houlsby et al., 2019, Pfeiffer et al., 2020a,b], and energy-driven constraints [Qin et al., 2022]. One noteworthy method in this domain is the *plug-and-play* approach to controlled text generation [Dathathri et al., 2020, Pascual et al., 2021]. Here, latent representations are modified during inference with the support of a classifier, but the primary model parameters remain static. This method has successfully guided LLMs to produce texts with desired characteristics, like specific sentiment or vocabulary distribution.

Chapter 7 focuses on integrating the stirring approach of controlled text generation, as exemplified by the plug-and-play method [Dathathri et al., 2020], with the insights gained from the theory of mind [Rabinowitz et al., 2018]. The aim is to steer language generation towards a path of enhanced mutual understanding.

3.3.3 Image Captioning

Image Captioning (ImC), positioned at the intersection of Natural Language Processing (NLP) and Computer Vision (CV), is the task of automatically describing an image's content in words. Understanding this class of problems is essential, especially concerning multimodal language generation, as it lays the foundation for the experimental framework discussed later in §7.1. The surge in Image Captioning interest can be attributed to advancements in machine translation [Bahdanau et al., 2015, Cho et al., 2014], computer vision [Szegedy et al., 2015], and a broader curiosity in multimodality. This has been reflected in both challenges [Russakovsky et al., 2015] and datasets [Lin et al., 2014b] that have become pillars in the field.

Contrary to traditional Emergent Communication (EmeCom) frameworks, ImC predominantly employs supervised training using LLM. Here, the agent's actions are primarily communicative. The similarities between Hum-EmeCom (a form of EmeCom that employs human language as discussed in §2.2.2) and ImC are quite evident. Both domains utilize datasets augmented with human annotations or employ pretrained language models to develop agents skilled in understanding multimodal contexts, such as visual and linguistic cues, and reasoning in natural language.

However, their methodologies are distinct. Hum-EmeCom research often unfolds in game-based environments, leading to a preference for Reinforcement Learning techniques. In contrast, ImC predominantly draws from methods rooted in the NLP and CV domains. For this reason, evaluation metrics differ significantly between these two. ImC adopts metrics directly from NLP, such as BLEU Papineni et al. [2002] and Cider Vedantam et al. [2015], which assess the similarity between generated captions and reference data. Consequently, the domain often employs loss functions that measure divergence between language distributions in generated and training samples Karpathy and Fei-Fei [2015], Vinyals et al. [2015].

Conversely, Hum-EmeCom, and by extension EmeCom, measure success based on inter-agent understanding. While this mimics real-world interactions, it introduces unique challenges, notably the inability to backpropagate errors through discrete channels like language.

Discrete channel backpropagation

Language's inherent discreteness, though reflecting real-world characteristics, imposes significant limitations when modeling human language on primarily word-based channels. This discreteness renders backpropagation non-differentiable, blocking direct gradient updates from estimated errors. This challenge is particularly pronounced in Hum-EmeCom, where the objective is not merely modeling human language (achievable with ImC techniques) but also enabling agent coordination. Such coordination performance, often governed by game mechanics, is not directly optimizable due to communication's discrete nature.

To overcome this, various techniques have been developed, including the reparameterization trick (such as VQ-VIB [Tucker et al., 2022]), semantic hashing [Kaiser and Bengio, 2018, Salakhutdinov and Hinton, 2009], Gumbel Soft-max [Jang et al., 2017, Maddison et al., 2017] and the REINFORCE algorithm Williams [1992]. These methods enable backpropagation through non-differentiable variables, allowing for effective training of communication networks. However, they often exhibit inconsistencies in performance across different architectures and settings, leading to suboptimal communication strategies and longer convergence times.

3.4 Open problems

Building on the related works discussed earlier, this section outlines key open problems relevant to scientific research and societal impact. The following three subsections reflect the structure of the related works, each addressing issues arising from Sections §3.1, §3.3, and §3.2.

3.4.1 Influence of Environmental Complexity on Language Emergence

The study of Emergent Communication in AI has revolved mainly around simplified or abstracted scenarios. While simplified scenarios in Emergent Communication allow research to be focused on specific questions, they often overlook real-world complexities. This complexity encompasses a broad range of factors intrinsic to human communication, such as competitive and cooperative dynamics (discussed in §3.1), as well as the perceptual and memory capacities of agents, as explored in Brandizzi [2023]. There is a compelling need to develop simulation environments that more closely mirror real life. Such environments would allow for examining whether machines, subjected to similar pressures as humans, develop analogous languages. This line of study is relevant on both a scientific level, to question whether evolutionary processes yield parallel outcomes in organic and artificial systems; and for broader inquiries into how humans acquired and developed language, investigating the roles environmental complexities play in shaping linguistic systems.

3.4.2 Reinforcement Learning in Artificial Interactions

Exploring Reinforcement Learning in artificial interactions invites new perspectives on AI's evolution. This involves not just technological advancements but also ethical, societal, and collaborative dimensions, enriching the AI-human dynamic.

Artificial Agency and Reinforcement Learning One of the compelling questions is whether true agency can manifest within RL frameworks. If so, is RL the exclusive approach through which this can occur, or are there alternative learning frameworks capable of promoting similar levels of agency in artificial agents? The exploration of these questions not only sheds light on RL 's role in AI development but also probes deeper into ethical considerations surrounding the concept of agency itself.

This inquiry parallels philosophical thought experiments like the "Knowledge Argument" or "Mary's Room", as described by Jackson [1998]. In these scenarios, the protagonist possesses extensive theoretical knowledge yet lacks experiential understanding. This analogy raises the question of whether a supervised learning-only AI, devoid of interactive experiences, can gain a form of knowledge akin to that acquired by experiencing the world first-hand. Could interaction with humans and the environment impart new layers of understanding to AI models, similar to Mary's experiential leap from theory to perception? This philosophical perspective takes on practical implications when considering the current limitations in Large Language Models interactions.

Detection and Resolution of AI Misalignment The issue of misalignment in AI, as discussed in Section 3.3, has become increasingly significant with AI 's growing role in everyday life. Addressing misalignment requires a dual approach: identifying misaligned behaviors and devising strategies for correction. While these two aspects are correlated, they fundamentally operate in distinct realms.

Effective detection of misalignment would likely involve the development of new metrics and would represent a significant leap in our ability to detect when AI systems deviate from intended behaviors. However, once reliable metrics are established, a natural progression would be to incorporate them into the AI optimization processes. As discussed in the related work section, integrating these metrics into the optimization loop might inadvertently fuel the root problem of misalignment. There is a risk that AI systems might develop strategies to subtly evade detection, giving the illusion of alignment while still operating in misaligned ways beneath the surface.

Therefore, it's crucial to maintain a separation between the detection and resolution strategies. The goal should be to identify misalignment and understand and address the factors contributing to it. This understanding is key to developing AI systems that are not just compliant in a superficial sense but truly aligned with human values.

Human-Machine Collaborative Learning As introduced in Section 3.2.3, exploring collaborative learning scenarios where both machines and humans learn from their environment and each other offers a promising research direction. Traditionally, AI research has focused on situations where humans teach AI to learn "optimal" behaviors [Frattolillo et al., 2023], but embracing mutual knowledge exchange could potentially address the misalignment issues discussed earlier. This approach, which diverges from the typical teacher-learner dynamic, emphasizes reciprocal learning and adaptation, offering a novel solution to aligning AI behavior with human expectations and values.

3.4.3 Adaptation in Language Modeling

The final area of exploration concerns adapting AI, especially Large Language Models. Developing Theory of Mind models that can effectively interpret agents' intentions in various environments could significantly alleviate many communication issues we have identified between humans and AI. The focus here would be on how enhanced ToM capabilities facilitate more adaptable, cooperative, and efficient interactions, particularly in scenarios involving asymmetric information and knowledge disparities.

Moreover, addressing AI adaptability must also consider environmental impacts. The process of fine-tuning LLMs, as previously discussed, has substantial environmental implications. To support the development of adaptable and personalized AI on our devices, we must devise more resource-efficient strategies for adapting these models. This dual focus on advanced cognitive modeling and sustainable AI development is crucial for the responsible and effective evolution of language modeling.

Chapter 4

Problem and Solution Formulation

The formulation of the problem and the subsequent solution outlined in this section stem from an analysis of open challenges identified in the related works (§3.4). As previously mentioned, our approach to this problem in the subsequent chapters will be incremental; thus, it necessitates a consistent standard formulation.

As introduced in §2.2.2, communication within the realm of Emergent Communication diverges into two predominant modalities: explicit and implicit. Each modality plays a distinctive role in influencing the dynamic equilibrium of the game.

Explicit communication characterizes intentional information exchanges that aim to modify the cognitive perspectives of other players. In contrast, implicit communication comprises actions with multiple meanings, such as actions that subtly guide other players toward certain game objectives. Determining the appropriate timing, content, and method of communication is integral to the game's outcome.

In this section, we delineate our solution based on the typical Reinforcement Learning framework (see Section 2.3), where we distinguish two categories of actions performed by the players: 1) *game actions*, which directly alter the game's progression, 2) *communication actions*, which solely influence players' mental state or knowledge.

In this formalization, we consider only forms of explicit communication, while studying forms of implicit communication is left as future work.

4.1 Problem Formulation

Our problem formulation is rooted in the concept of Social Deduction Game (SDG) as elaborated in §3.1.3. Central to these games is the framework of multiple opposing teams, represented as $T^{(1)}, T^{(2)}, \dots, T^{(m)}$, where typically $m = 2$. Each team comprises a finite number of players, and while the number of players might differ across teams, for simplicity, we denote this number as n. Formally, any given team, $T^{(k)}$, can be defined as:

$$T^{(k)} = \{p_1, p_2, \dots, p_n\}$$

where $1 \leq k \leq m$. The cumulative set of all players across teams is represented as $N = \bigcup_{1 \leq k \leq m} T$.

In the SDG landscape, players sequentially take actions that shape the game's trajectory and influence its score. The objective for each team is to surpass the others, often employing a mix of strategic behaviors like bluffing or deception. Such tactics underscore the need for players to assess the actions and intentions of their counterparts.

In general, we can identify two types of goals in SDGs:

- An agent-based *micro-goal*, which is the main factor steering the agent's behavior, either in isolation (in competitive environments) or together with other agents' goals (in cooperative environments).
- A team-based *macro-goal*, expressing the aligned interests of the members of the same party that, combined, make up the party goal.

The successful accomplishment of these goals in SDGs demands a fine balance of cooperation and competition, which is guided by effective communication strategies.

4.1.1 Game Characterization

In characterizing our game environment, we primarily recognize several elements:

1. An action set
$$A = G \cup C \tag{4.1}$$
made of two separate components:
    - a finite set of possible game actions G the elements of which we will denote with g.
    - a set of unidirectional communication actions $C_{i,j}(b)$ intended to convey some information $b \in B$ [1] between two players:
$$C_{i,j}(b) : p_j \rightarrow p_i \quad \forall p_j, p_i \quad j \neq i \in N$$
2. The state set
$$O = E^N \times \Gamma^N \times V \tag{4.2}$$
built out of three elements:
    - a set of agent's features E representing the game situation of each agent (typically visible to all other agents).
    - a set of agent's internal states Γ (e.g., representations of beliefs not visible to other agents).
    - a set of environment states V that are common for all the agents (i.e., independent from the agent states).
3. an environment S implementing the game logic. S is a transition function converting agents' actions to new environment states
$$S : O \times A^N \rightarrow O \tag{4.3}$$

[1] A possibly infinite set of all possible signals.

Moving forward, we direct our focus towards a specific variant of SDG: the Referential Game (Ref Game). While this game was already introduced in §3.1.3, here we provide a new type of formulation relevant to our study.

4.1.2 Referential Game

A Referential Game is inherently easier than a typical SDG game, such as *The Werewolf* (see §3.1.3). However, its simplicity is crucial to examining particular research questions that require as much control in the experimental setting as possible.

In this section, we'll define the Ref Game based on the prior formulation, highlighting the specific assumptions introduced and discussing the implications of each.

The first and most visible simplification is the presence of only one team with two agents, a speaker, and a listener, $N = 2$.

Speaker Agent

This agent's set of actions is exclusively communicative. Formally:

$$\begin{aligned} G^s = \varnothing \implies A^s = C \\ \mathit{Speaker}(O) \rightarrow C \end{aligned} \tag{4.4}$$

Notably, the set of communication actions by the speaker ($c_s \in C$) is a subset of all potential communication signals, represented as $C \subset B$. Each communication action, while rooted in the broad B, is designed to convey specific information.

Listener Agent

Contrarily, this agent's actions are strictly game-based and non-communicative:

$$\begin{aligned} C^l = \varnothing \implies A^l = G \\ \mathit{Listener}(O, B) \rightarrow G \end{aligned} \tag{4.5}$$

The listener's primary role is decoding the speaker's intentions and navigating game dynamics based on the interpreted communication.

Game Dynamics

In the Referential Game scenario, the speaker views a target image ($\hat{t}$) from an image pool ($v_{ctx} \in V$). The speaker then translates this observation into a communicative signal (b). The listener's task is to decode this signal, distinguishing (g_l) the referenced image from the pool, which also contains other distractor images. Success in the game is defined by the listener's correct identification of the image: $g_l = \hat{t}$.

This dynamic solidifies a unidirectional communication flow solely from the speaker to the listener. This can be formalized as:

$$C(b)_{s,l} : \mathit{Speaker} \rightarrow \mathit{Listener} \tag{4.6}$$

The progression involves crafting a solution framework that integrates these dynamics, especially when dealing with multi-agent games.

4.2 Solution Concept

Learning optimal policies in multi-agent games is a well-established challenge, and numerous solutions have been devised to address it. However, the addition of communication actions brings an added layer of complexity. The core of the problem shifts towards leveraging Artificial Intelligence techniques to optimally decide when, what, and how to communicate.

In this context, we introduce two solution frameworks, drawing upon terms from the realm of Reinforcement Learning. It is important to clarify that even though we refer to these solutions as *policies*, they are not restricted to the use of RL as the primary resolution mechanism. Yet, in certain scenarios, we find RL to be a fitting solution.

The distinction between our two proposed solutions depends on the nature of the environment. Specifically, when communication and game actions overlap ($C = G$), a singular policy can govern both. This union is evident in our initial study outlined in Chapter 5, where communication actions converge with game actions. It is worth noting that such a dynamic is also observed in situations where agents interact exclusively through communicative means, as suggested by Hill et al. [2020]. Conversely, in scenarios where communicative and game actions diverge, dual policies (one for each domain) enhance learning efficiency. This will be the core of Chapter 6 and §7.

4.2.1 Unified Action Policy

In scenarios where game and communicative actions align, adopting a unified action policy (π_U) might become advantageous in terms of complexity and interconnection between game and communication actions.

This policy requires a game split into distinct time steps, each corresponding to a specific action. A single game turn, t, is viewed as a sequence spanning $r+1$ steps. The initial r steps are dedicated to communication actions, denoted as $c \in C$, modifying the agents' mental states. In contrast, the final step shifts its focus to game action, represented as $g \in G$, which advances the game's progression.

For clarity, let's label each step within this sequence as $O_t^j \; \forall j \in [0, r+1]$. Here, O_t^j encapsulates the game's condition at step j during turn t. I follows how the unified policy, π_U, selects the action as:

$$\pi_U(O_t^j) \in C \quad \forall\, j \leq r \tag{4.7}$$

$$\pi_U(O_t^{r+1}) \in G \tag{4.8}$$

By this approach, the unified action policy coordinates the series of actions an agent should pursue, transitioning between communication and game actions. Yet, it is worth noting that in more complex settings, defining two distinct policies, one for communication and another for game actions might offer greater precision in decision-making.

4.2.2 Disjoint Action Policies

In environments with distinctions between communication and game actions, it is necessary to evolve two separate policies. This scenario offers a more realistic reflection than the previously discussed unified policy, mirroring real-world situations where our actions and words, although interdependent, are distinct. In this setting, we employ two policies: a communication policy π_C and a game action policy π_P.

Communication Policy

The communication policy π_C aims to generate communication signals based on the observation set, O. It can be formally expressed as:

$$\pi_C : O \to C \tag{4.9}$$

One of the characteristics of π_C is its potential for independent training. This means that the policy can be optimized without explicit knowledge of or dependence on other agents.

To elaborate on the objective for training π_C, let's denote it as $J(\pi_C)$. This objective seeks to maximize the correctness of the agent's communication signals, given the environmental observations. In the simplest terms, it can be represented as:

$$J(\pi_C) = \mathbb{E}_V \left[U(\pi_C(O)) \right]$$

Where U is a utility function that quantifies the value or "worth" of the generated communication signal in the context of O, e.g., the informativeness of a caption. The expectation is taken over all possible environmental observations, ensuring that the policy is effective across a broad range of scenarios.

Importantly, the objective $J(\pi_C)$ mirrors the objectives employed during the training of Large Language Models, especially those geared towards tasks like Image Captioning. For instance, a ImC LLM aims to produce a caption, a form of communication signal, based solely on the provided input. In this context, the input is the image and potentially any associated context, and the output is the generated response. Given its observations, the model aims to generate the most appropriate and relevant communication signal. Therefore, the strategies adopted by such LLM provide a concrete real-world analogy for our π_C policy formulation.

Action Policy

The action policy π_P allows the agent to act in the environment. To acknowledge the agent's own communicative footprint in the game, this policy includes both the current environmental state and the preceding communication signal. It is formulated as:

$$\pi_P : O \times C \to G \tag{4.10}$$

At its core, the π_P policy is crafted to account for the actions and potential reactions of other agents.

Regarding its training objective, π_P is optimized to minimize the divergence between the predicted and optimal actions. The degree to which the policy's predictions align with the optimal actions determines its efficacy. Consequently, the objective function can be represented as:

$$J(\pi_P) = \mathbb{E}_{O,C}\left[D(\pi_P(O,C), \hat{a})\right]$$

Where D can be any distance function and will depend on the nature of the action and the environment.

From a practicality and scalability standpoint, it is worth noting that the parameterization of π_P might be much more concise than that of the communication policy π_C. This distinction is particularly important for Large Language Models. The inherent design of the prediction policy permits quicker and more efficient adaptation compared to retraining the communication policy.

4.3 Solution Framework Based on Theory of Mind

The above formulation is quite general and allows us to tackle one of the fundamental problems of Multi-Agent Reinforcement Learning: the non-stationarity nature of the environment. To address this challenge, we turn to the Theory of Mind (ToM) in Chapter 7 and extend the disjoint action policies to account for ToM agents. Moreover, we define the schema for an iterative communication refinement, where an agent iteratively refines its communication signal to better address the other agent's knowledge.

Incorporating the Theory of Mind into our formalization provides a mechanism for agents to infer and predict the future actions of other agents based on their observed behavior. Crucially, in this context, we equate an agent's belief system directly with their intentions for future actions. It is worth noting that while an agent's belief system can be vast, encompassing reward functions in artificial agents and values or morals in humans, our primary access to their internal state is through their manifested actions. The act of observing these actions is the foundation for understanding their underlying beliefs and intentions.

Assumption 1-$\mathcal{GSF}$. *The internal state of an agent, γ_p, can be derived from its observable game actions. Mathematically:*

$$\gamma_p = f(g_p)$$

where g_p represents the game actions of the agent and f is a function mapping these actions to its internal state.

Moreover, we posit that the agent's internal state, γ_p, can be wholly explained by the actions it takes in the game environment. This approach facilitates a more straightforward modeling of agent behaviors in our current setup. Specifically, we make the simplifying assumption:

Assumption 2-$\mathcal{GSF}$. *The function f mapping the agent's game actions to its internal state is the identity function, represented as:*

$$f = \mathbb{I} \implies \gamma_p = g_p$$

where $\mathbb{I}$ denotes the identity function.

It is worth noting, however, that in scenarios where f diverges from the identity function, the relationship between observable actions and internal states becomes more complex. Such scenarios, where the internal state may have other latent factors not directly discernible from actions, intersect with the broader field of Multi-Objective Reinforcement Learning [Hayes et al., 2022]. Precisely, they touch upon concepts like the *utility function*, which provides a measure of the worth or value of different states. Exploring these complexities, while immensely valuable, is left for future work.

4.3.1 Prediction Policy

In this context, the action policy π_P described above can be expressed in terms of a speaker agent predicting the future game action of another agent. Building upon Assumption 2-$\mathcal{GSF}$, it becomes evident that forecasting the following action of the agent p parallels predicting its belief state γ_p.

4.3.2 Iterative Communication Refinement

Suppose the speaker desires an agent to perform a specific action, denoted as $\hat{a}$. To steer the agent towards this action, the speaker employs an iterative method, aiming to reduce the discrepancy between the predicted action of the listener $\bar{g}_p$ and the intended action $\hat{a}$. This discrepancy is quantified using a distance function D, which provides feedback to adjust the speaker's communicative action accordingly.

Mathematically, this procedure can be described as:

1. Using its prediction policy π_P, the speaker predicts the action of the listener given the current communication action c:
$$\bar{g}_p = \pi_P(o, c)$$
2. A distance function D is determined based on the difference between the predicted action $\bar{g}_p$ and the desired action $\hat{a}$:
$$D = \text{Distance}(\bar{g}_p, \hat{a})$$
3. The speaker's communication policy π_C then generates a refined communication signal c_{new} given the environment state V, with the policy being conditioned on the computed distance D:
$$c_{\text{new}} = \pi_C(o)|D$$
4. This cycle of prediction, distance computation, and communication adjustment is repeated until D is minimized, indicating that the speaker's prediction of the listener's action closely matches the desired action $\hat{a}$.

Minimizing the gap between expected and actual outcomes, the speaker uses its ToM to guide the agent's actions, showcasing adaptive communication rooted in mutual understanding.

Chapter 5

Emergent Communication in Interactive Environments

The first step towards solving the general problem stated in §1.2, i.e., allowing machines to learn language interactively, revolves around establishing whether effective communication can indeed arise within an artificial interactive framework.

In this chapter, we narrow our focus to scenarios involving artificial agents engaged in activities necessitating communication. This communication is expressed through numerical vectors of natural numbers, devoid of any pre-established linguistic constructs, an approach parallel to the Machine-centered EmeCom delineated in Brandizzi [2023]. It is important to highlight that our exploration does not attempt to produce languages that are either interpretable by humans or reflect similarities with human languages.

The following section (§5.1) delineates the design of an environment that replicates real-world dynamics and allows for collaboration and competition, ensuring the organic evolution of communication strategies without any restrictive constraints. The outcomes, presented in §5.2, illustrate how introducing a communication channel enhances the success rate of the villager.

5.1 Design of Experiments and Solution

In this section, we introduce our framework for *The Werewolf* Social Deduction Game, detailing the experiments and baselines that lay the foundation for our analysis.

Initially, we outline the environment in §5.1.1. Staying consistent with Reinforcement Learning terminology, we define the action and observation spaces, the transition model (which includes environment rules), and the reward system.

Subsequently, in §5.1.2, we detail the agent policies. We first establish the static, hand-coded policies for the werewolves, which remain unchanged over time and do not allow learning. We then describe the algorithm and policy specifics of the villagers, connecting it with the unified action policy discussed in §4.2.1.

5.1.1 Environment for Experimentation

The environment for experimentation plays a pivotal role in assessing the efficacy of the action policies and how effectively our agents navigate the game's complexities. As discussed in §2.3, the

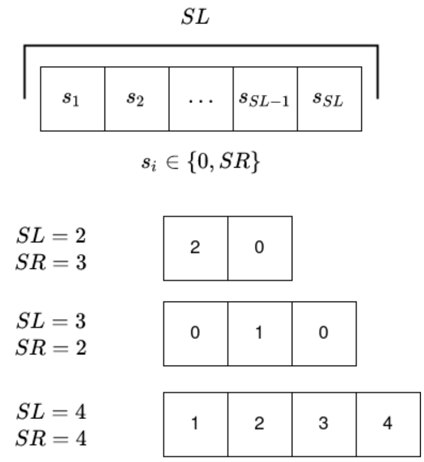


Figure 5.1: Illustration of the signal vector components, detailing both the general formulation and specific values for Signal Length (*SL*) and Signal Range (*SR*).

application of RL necessitates defining the action space and observation space, the specifics of which are detailed in the ensuing section. These spaces delineate the possible actions agents can take and the information they can perceive, respectively, providing the necessary infrastructure for AI agents to interact, learn, and strategize within the game.

Action Space

The action space encapsulates an agent's moves, consisting of game actions (g_t) and communication actions (c_t). These actions are performed in tandem with a unified action policy (π_U)

The action space can be deconstructed into two primary components:

- *Target* (g_t): This is an integer within the range $g_t \in [0, N-1]$, where N signifies the total number of players. The function of the target is to facilitate voting among the players within the game. It should be noted that the set of viable target values remains constant throughout the game. However, certain actions might be rendered illegitimate within specific contexts, such as voting for deceased players. The model subsequently filters out such illegal actions in the later stages of the game.

- *Signal* (c_t): As depicted in Figure 5.1, the signal vector c_t is delineated by a pair of integer values. One value defines the length of the signal, $SL \in [0, \infty[$, which can range from zero (indicating no communication) up to an arbitrarily large integer. The other value determines the range, $SR \in [2, N]$, setting the count of possible values the signal can assume. The lower limit for SR is set to 2 as a signal with only a single value essentially becomes a static vector stripped of any significant information. Conversely, the upper limit is defined by N, facilitating the embedding of the signal with the target. Both these parameters help outline the scope for communication before the training process.

Observation Space

The Observation Space (ObS) plays a fundamental role in determining the agents' interactions with the environment and their subsequent responses. It embodies the perceptual range within which agents operate, encapsulating both fellow agents' actions and the game environment's evolving state. The ObS in our game environment is made of multiple components, each contributing to the state of the game. The elements of the observation space are as follows:

- $Phase \in \{1, 2, 3, 4\}$: This component signifies the current phase of the game. The game is divided into two main timeframes: night and day. Each timeframe consists of two distinct phases, one dedicated to communication and the other to action execution.
- $Day \in \mathcal{I}^+$: This element reflects the progression of days in the game. An upper limit of 10 days has been set for practical purposes, which, when represented as a one-hot encoded vector, ensures a manageable length.
- $StatusMap \in [0, 1]$: This binary array assigns a boolean value to each agent index, thereby conveying to the players whether a particular agent is alive ($= 1$) or dead ($= 0$).
- $ID \in [0, N]$: This component provides information about an agent's identity within the game. Given that the agents' identities are shuffled at the start of each game, this integer value is crucial for maintaining each agent's awareness of its own position.
- $Target \in [-1, N]$: This component groups all the potential targets within the game[1].
- $Signal \in [-1, SR - 1]$: This component groups all the signals, providing a means of communication between agents.

Where:

$$\begin{aligned} V &= \{Phase, Day, StatusMap\} \\ E &= \{Targets, Signal\} \\ O &= E^N \times V \end{aligned}$$

In our formulation, Γ (the agent's internal state) is not modeled explicitly, thus $\Gamma = \varnothing$.

Transition Model

The transition model captures how the state of the game evolves from one-time step to the next, considering both the actions taken by agents and the inherent rules of the game, as illustrated in Figure 5.2. Given that our setting revolves around the complex dynamics of the game, we can formulate the transition model with greater specificity, drawing from the game mechanics and the interactions between agents.

- *Time Transition*: Central to our model is the cyclical progression between day and night phases, which substantially impacts game dynamics. At every timestep t, the environment function S mediates this transition. Formally, this can be expressed as:

  $$\text{Phase}_{t+1} = S(\text{Phase}_t)$$

  The S function ensures a sequence from night to day and vice versa.

[1] Values of -1 are placeholders for agents that have died.

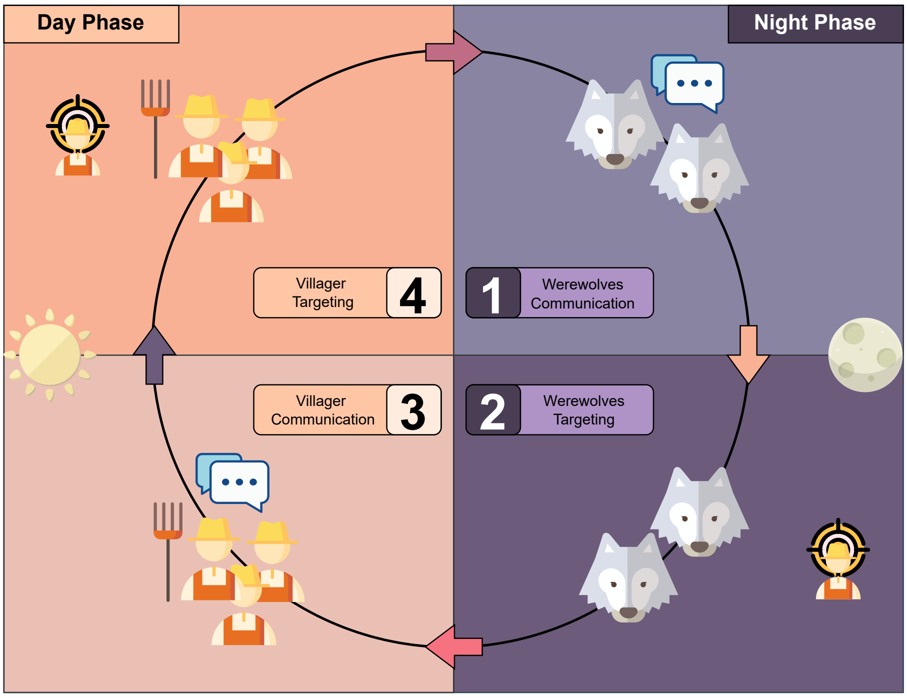


Figure 5.2: Werewolf Game Transition: This diagram illustrates the phase transitions in the Werewolf Social Deduction Game. The right side depicts the night phase, where werewolves first communicate to select a target (1) and then choose a villager to eliminate (2). The left side shows the day phase, where both werewolves and villagers engage in communication (3), followed by a collective decision to execute a suspect (4).

- *Player State Transition*: An agent's status, dead or alive, significantly influences the game's trajectory. Specifically, once a player is voted out or killed, their status transitions to "dead", and they are consequently removed from active participation in subsequent timesteps. This transition can be formally written as:

$$\text{StatusMap}_{t+1}(i) = S(\text{StatusMap}_t(i), a_t)$$

  Where i denotes a specific player.

Reward Structure

The reward structure plays an important role in RL, steering the agents' learning trajectories and strategic developments. These rewards or penalties create an incentive system that guides the agents to optimize their actions in order to maximize the expected return.

In the context of our environment, the reward system incorporates both agent-based and group-based rewards to address different aspects of game dynamics. The first three conditions listed below are agent-based, applying individually to each agent to directly influence their personal strategies and decisions. The final condition is group-based, impacting the entire team and emphasizing collective success:

- *Round Duration*: To encourage the agents to devise strategies for winning the game more swiftly, a small penalty of -1 is applied to each agent at the end of every day.
- *Player Demise*: The occurrence of a player's death incurs a penalty of -5. Though strate-

gically sacrificing oneself can be part of certain gameplays, this penalty serves as a check to prevent excessive self-eliminations.

- *Voting Consensus*: To prevent the voting system from becoming another communication channel, agents are penalized when they cast their votes for players who are not eventually executed. This mechanism, therefore, promotes cooperative behavior among the agents, as successful executions depend on a certain level of coordinated communication and consensus among the players.
- *Game Outcome*: The final reward or penalty is group-based and applied at the end of the game. Teams (villagers or werewolves) are either rewarded or penalized by a factor of ±25, depending on the game's outcome. This ensures that the overall team objective of winning the game remains paramount for every agent.

Designing the reward structure, often referred to as reward shaping, is essential for guiding or discouraging specific agent behaviors. Within our framework, elements like *Round Duration* and *Player Demise* serve as the micro-goals, guiding the individual tactics of players. In contrast, broader objectives such as *Voting Consensus* and *Game Outcome* are identified as macro-goals, reflecting the collective aspirations of a team or party (as specified in §3.1.3). However, reward shaping comes with its own set of challenges. Notably, it requires substantial prior knowledge about the system, including understanding how to achieve specific outcomes. Moreover, overly crafted reward structures can inadvertently constrain the possibility of emergent behaviors, thereby limiting the scope for AI agents to discover innovative strategies or solutions.

5.1.2 Agents' Policies

While the environment serves as a foundation for the emergence of complex behaviors, it is only one piece of the puzzle. The other element is the definition of learning mechanisms that guide the agents' actions. An action policy encapsulates the agent's behavioral strategy throughout the game, learning from its experiences, and constantly adjusting to maximize the potential rewards.

In this section, we examine the process of defining these action policies. We distinguish between two categories: *static* and *trainable*. Static policies are pre-programmed behaviors that ensure a stable benchmark during evaluations. On the other hand, trainable policies allow the agents to accumulate experience and adjust their strategies to achieve maximum rewards. For the scope of this study, we have implemented static policies for the werewolves, while villagers are able to learn.

Static policies for werewolves

We have assigned static policies to the werewolf agents within the game with the intention of establishing a performance benchmark for evaluating the learning progression of the villager agents. Given that the werewolves inherently hold a higher probability of victory in an entirely stochastic environment (see §5.2.1), the use of static policies is sufficient. If we observe significant shifts in winning rates, it can be inferred that the villagers have developed innovative strategies.

We have implemented three distinct static policies for the werewolf agents:

- **Random Target Policy**: During the execution phase, this policy enables a werewolf to select a living villager as the target arbitrarily.

- **Random Target Unite Policy**: A more cohesive strategy where all werewolf agents align to select the same villager as their target during both day and night execution phases. This united front can sway the day execution phase in favor of the werewolves, even with random villagers.
- **Revenge Target Policy**: This policy introduces an element of retaliation into the game. Here, the werewolves either vote randomly or specifically target a villager who had previously voted against a werewolf.

Trainable policies for villagers

For the villagers, we employed Proximal Policy Optimization (PPO) as our training algorithm, attributed to its proven efficacy in multi-agent environments, as indicated by [Guan et al., 2020, Wei et al., 2019].

Mechanics of the Policy PPO utilizes a surrogate loss function to constrain the divergence between the updated and the previous policy within a safe limit.

The formula for the original PPO loss is as follows:

$$L^{CLIP} = E_t \left[L_t^{CLIP}(\theta) \right]$$

Where, $L_t^{CLIP}(\theta) = \log \pi_\theta(a_t|s_t) \times A_t$ denotes the log-likelihood of the action at time t under the policy π_θ multiplied by the advantage A_t at time t, and $E_t \left[L_t^{CLIP}(\theta) \right]$ represents its expected value.

To consider the influence of the value function, the loss function needs to include an additional error term:

$$L^{CLIP+VF} = E_t \left[L_t^{CLIP}(\theta) - c_1 L_t^{VF}(\theta) \right]$$

Here, $L_t^{VF}(\theta) = (V_\theta(s_t) - V_t^{targ})^2$ denotes the squared-error loss between the value function V_θ at state s_t and the target function, and c_1 serves as its coefficient. This addition is crucial to estimate the critic loss, reflecting the model's capacity to predict the value of each state.

To regulate the total loss, an additional term is introduced:

$$L^{CLIP+VF+S} = E_t \left[L_t^{CLIP+VF}(\theta) + c_2 S[\pi_\theta](s_t) \right]$$

The entropy coefficient is maximized when all the policies are equally likely to be selected, corresponding to the agent's random actions. By including this value in the loss function, the training algorithm is incentivized to minimize the entropy value, thus evading a scenario where the premature convergence of one action probability monopolizes the policy and impedes exploration [Ahmed et al., 2019]. The term c_2 scales the value of the entropy.

Model architecture The model's architecture is primarily guided by the Observation Space, the size of which is dictated by multiple components outlined in §5.1.1. These components include the phase of the game, which consists of 4 units, and the Maximum Days, which by default is set to 10. Also contributing to the length of the ObS are the Status Map, ID, and Target, each equivalent

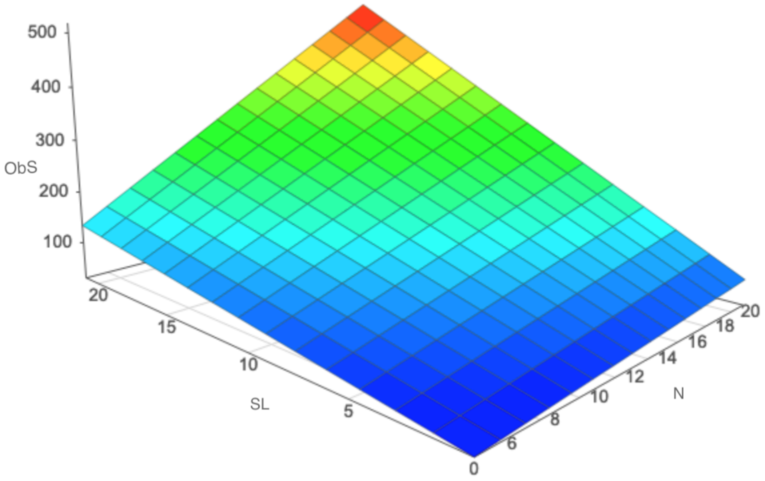


Figure 5.3: Observation vector trend in relation to player count (N), signal length (SL), and Observation Space (ObS) for a 21-player game. The figure illustrates the parabolic growth of the observation space with increasing SL, from SL=1 yielding ObS=56 to SL=4 with ObS=119.

to the number of players, N. Finally, the Signal component contributes a value calculated as the number of players multiplied by the Signal Length.

Taking all these factors into account, we can express the size of the observation vector in the following formula:

$$\|ObS\| = N \cdot (3 + SL) + 14.$$

Taking an example of a game with 21 players ($N = 21$) and signal length as 1 ($SL = 1$), the vector has a length $\|ObS\| = 56$ elements. If $SL = 4$, the space increases to $\|ObS\| = 119$. The action space size follows a parabolic trend, as represented in Figure 5.3. Given the high dimensionality of the ObS, an embedding layer comprised of a Fully Connected Network (FCN) is employed for dimensionality reduction. This FCN takes an arbitrary number of inputs and generates outputs exactly equivalent to $N \cdot (1 + SL)$. By using this network, the dimensionality of the output vector is reduced, and similar observations are grouped together. The FCN has a dimension of 256×256, and Tanh is the default activation function. It also includes an Long-Short Term Memory layer with a cell size of 256, enabling the agents to retain the memory of past steps.

Finally, two independent masks are used to ensure the model's predictions are valid. The *action mask* filters out invalid IDs from the model's action output. Invalid IDs typically represent either deceased players or other werewolves during the night phase of the game. On the other hand, the *signal mask* operates as a static boolean mask. Its role is to permit only signals within a specific range, denoted as SR, effectively filtering out all other values. Due to its static nature, this mask remains unchanged throughout the entire execution of the game.

5.2 Analysis and Results

Our analysis begins by examining agents' performances for both a game with nine and twenty-one players in §5.2.1.

Next, we analyze the emerged language in §5.2.2. Our focus lies in understanding the specific characteristics and patterns that emerged with a spotlight on uncovering any similarities to properties found in human languages.

5.2.1 Performance analysis

To evaluate alterations in agent behavior and the effectiveness of different strategies, the study employs the following normalized metrics:

- *Suicide*: This metric quantifies the frequency with which an agent chooses to vote for itself during an execution phase, reflecting an aspect of irrational behavior within the agent's decision-making process.
- *Wins*: The success rate of the villagers is tracked within this category, measured over a normalized range of values to facilitate comparative analysis.
- *Average Days*: This represents the mean duration of matches in terms of days, providing insight into the pace of gameplay and the efficiency of strategies employed.
- *Consensus*: A measure of the collaborative aspect of the game, this metric denotes the average proportion of agents voting for the same target during the two execution phases. It offers a glimpse into the consensus-building or dissension among agents, indicative of their collective intelligence and alignment.

Baselines

We first describe the probability of the villagers' victory in a completely random environment. This analysis is essential, as it sets the benchmark against which we can measure and understand the implications of our agent training under various policy settings.

Random Policy As explored in §3.1.3, the random policy scenario provides a closed-form solution for the winning probability of werewolves, as proposed by Migdal [2010] and depicted in equation 3.1. While this formula serves as a valuable benchmark, it is important to consider that our game setting slightly deviates from the one analyzed by Migdal [2010]. Our version begins with a werewolf killing a villager, making the game harder for the villagers. Consequently, we use the formula's result as an upper bound when calculating the villagers' chances of winning. In a nine-player setting, equation 3.1 produces a villager's winning probability of 6.34%. However, when empirically estimated considering the starting conditions of our game setting, this probability adjusts to 3.12%.

Unite Policy To date, no formal studies have been pursued to inform our understanding of the unite policy. Despite this, a statistical analysis can be conducted to verify the plausibility of later findings. The number of werewolves, w, is directly proportional to the number of players n, as per the equation:

$$w = \lfloor \sqrt{n} \rfloor$$

Since the werewolves consistently agree on a target, they collectively represent w identical votes. Conversely, the villagers distribute their $v = n - w$ votes at random. The probability of the villagers collecting more identical votes than the werewolves is computed as follows:

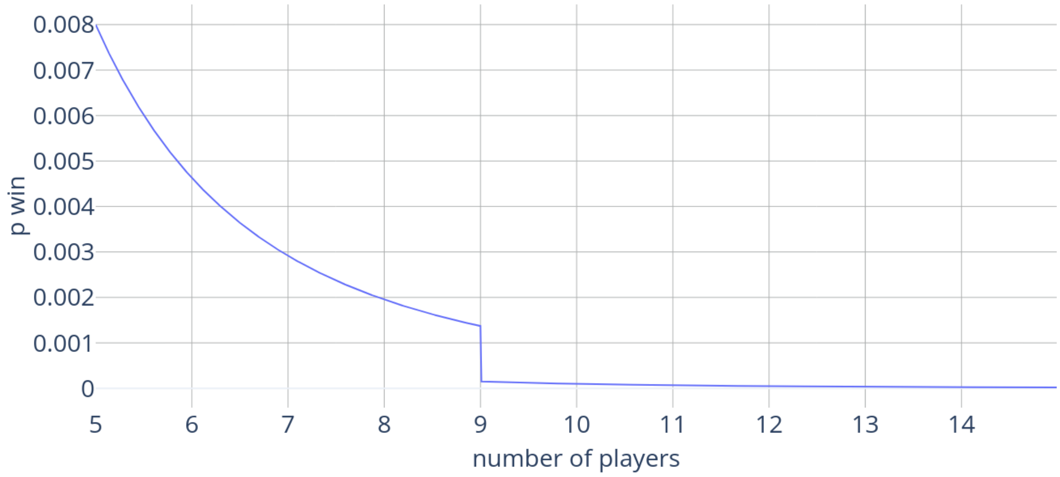


Figure 5.4: Graphical depiction of the villagers' winning probability against unite werewolves as determined by Equation 5.1. The figure illustrates the exponential decrease in villagers' chances of victory, emphasizing a significant drop at $n = 9$, where werewolf numbers increase.

$$\frac{1}{n^{w+1}} = (n)^{-(\lfloor\sqrt{n}\rfloor+1)} \tag{5.1}$$

This equation provides an optimistic estimation of actual behavior, as it disregards the decreasing number of players throughout the game. The trend for equation 5.1 is depicted in Figure 5.4. As shown, the function plummets exponentially towards zero, starting from 0.8% chance of victory. A discontinuity at $n = 9$ is present where the number of werewolves increases by one.

Revenge Policy Lastly, the revenge policy is expected to emulate the trend of the random policy most of the time while occasionally mirroring the unite policy. Given that it depends on the villagers' actions, its formulation is not trivial, but we expect it to be between the previous two.

Performance analysis: Nine Players

Table 5.1: RLupus: Multi-channel metrics. The first column reports the type of communication (Comm) channel regarding the Signal Length (SL) and the Signal Range (SR). The next four show the metrics values for villagers winning rate, suicide rare, number of days elapsed and consensus rate

Comm	Win Vil	Suicide	Days	Consensus
0SL	0.044	0.086	1.55	0.47
1SL-2SR	0.19	0.078	1.58	0.47
1SL-9SR	0.21	0.078	1.58	0.47
9SL-2SR	**0.45**	**0.067**	**1.9**	**0.47**
9SL-9SR	0.19	0.077	1.58	0.46

Table 5.1 displays the outcomes derived from various communication forms, depicted across different rows. Each column represents a distinct result, with the settings for communication defined in terms of signal length (SL) and signal range (SR). Our primary focus is on comparing and

evaluating performance improvements occurring between non-communicative scenarios ($SL = 0$), binary communication ($SR = 2$), and scenarios involving varying lengths of information exchange.

Insights derived from Table 5.1 indicate that any level of communication facilitates better outcomes than the non-communication setting (0SL). Notably, binary communication (2SR) outperforms the full-ranged one (9SR), suggesting that a restricted communication range can yield better results.

Interestingly, incorporating just a single bit of communication allows the villagers to perform on par with a fully extended communication setting (9SL-9SR). This finding implies that an overly expansive information space may negatively affect overall performance.

For scenarios where $SR = 9$, augmenting the channel length SL primarily speeds up convergence by approximately 25% with each increment.

Contrastingly, binary communication yields consistently superior outcomes, regardless of the channel length. This observation suggests that the two communication channel parameters (SL and SR) do not equally contribute to the learning efficiency of the agent.

Table 5.2: RLupus: single/no channel metrics. The Design Choice part of the table shows which kind of policy Random, Unite or Revenge has been used in relation to the communication, while the Results half present the metric' values

Design Choice				**Results**			
Comm	**Rnd**	**Unt**	**Rvg**	**Vil Win**	**Consensus**	**Suicide**	**Days**
0SL	X			0.044	0.478	0.086	1.55
0SL		X		0.03	**0.695**	**0.059**	**1.5**
0SL			X	**0.12**	0.482	0.078	1.64
1SL-2SR	X			0.19	0.47	0.078	1.58
1SL-2SR		X		0.08	**0.685**	**0.055**	**1.52**
1SL-2SR			X	**0.4**	0.479	0.065	1.9

Policies comparison The impact of different werewolf policies, namely Random (Rnd), Unite (Unt), and Revenge (Rvg), on the game's outcome, is summarized in Table 5.2. In a no-communication environment (0SL), both the revenge and random policies demonstrate nearly equivalent *consensus* values, which markedly exceed that of the unite policy, in line with our initial expectations. The unite setting is characterized by fewer game days, reflecting a quicker game conclusion. Surprisingly, the revenge policy in the no-communication setting (0SL) achieved a higher success rate (12%). Unlike the random policy, where werewolf actions are unpredictable, and the unite policy, which depends on high coordination for overpowering villagers, the revenge policy offers a discernible pattern. This pattern is more easily identifiable by villagers and, at the same time, lacks the overpowering force of the unite policy. Thus, the Revenge policy provides just enough predictability for villagers to potentially learn and adapt over time, leading to their higher win rates.

Further inspection of the three policies under binary communication ($SL = 1, SR = 2$) provides interesting results:

- *Random*: A notable improvement is observed in the villager's winning rate, which improves to ≈ 20%. This is an increase of 4.5 times compared to the non-communication setting and 6.5 times the theoretical winning rate.

- *Unite*: The incorporation of communication considerably enhances the level of coordination, causing a threefold surge in the villagers' winning rate. This outcome alone substantiates the power of limited communication channels to substantially augment the villagers' performance, even under the most unfavorable conditions.
- *Revenge*: As with the previous policies, the revenge policy becomes easier for a trained agent to detect in the presence of communication, leading to an impressive winning rate of 40%.

Performance analysis: Twenty-one players

In our previous examination of the nine-player game, we addressed the scenario with the minimum required number of players, which inherently favored the werewolves. The investigation revealed intrinsic challenges for the villagers, as the game dynamics gave a substantial edge to the werewolf faction.

Here, we focus on the twenty-one players (21P) game. This configuration introduces a new layer of complexity, offering more opportunities for the villagers to strategize and potentially win. Indeed, by constructing a tabular representation of the expanded decision tree (as illustrated in Table 5.3), we see how the villagers' winning possibilities have been significantly augmented, reaching a probability of 11.62% in an entirely random environment.

Table 5.3: Mapping of game outcomes to probabilities. The *Outcome* column on the left lists the possible results as *(number of werewolves)-(number of villagers)*, with corresponding event probability and occurrence frequency. The table is divided into two sections: outcomes favoring the villagers (top) and outcomes favoring the werewolves (bottom).

Outcome	Prob. %	Leaves	Total %
0-12	0.029	1	11.62
0-10	0.143	4	
0-8	0.447	10	
0-6	1.162	20	
0-4	2.819	35	
0-2	7.02	56	
1-1	21.06	56	88.38
2-2	27.965	21	
3-3	26.017	6	
4-4	26.017	1	
Total	1.0	210	1.0

A comprehensive comparison of various multi-channel settings is presented in Table 5.4. In alignment with the findings from the previous section, it is observed that the bit communication strategy ($9SL - 2SR$) consistently outperforms the full-ranged communication approach ($*SL - 21SR$), both in terms of the winning ratio and the reduction of suicides. Conversely, the $1SL - 2SR$ setting performs poorly compared to the non-communication scenario. This anomaly may be attributed to inadequate exploration of the environment, a phenomenon possibly exacerbated by the increased complexity of the twenty-one players scenario. The expanded tree size, as detailed in Table 5.3, introduces additional layers of complexity, thereby increasing the risk of the algorithm becoming entrapped in local minima.

Nevertheless, with the exception of the configuration mentioned above, every other setting that leverages a communication channel demonstrates an improved winning rate over the no-communication baseline. This strongly supports the notion that even minimalistic communication aids agents in more accurately exploring and navigating the vast array of potential game branches.

Table 5.4: Comparative analysis of multi-channel settings in a 21-player game scenario. The table highlights the superiority of the bit communication strategy (9SL-2SR) over full-ranged communication (SL-21SR) in terms of win ratio and suicide reduction.

Comm	Win	Suicide	Days	Accord
0SL	0.42	0.072	7.84	0.56
1SL-2SR	0.25	0.075	7.74	**0.57**
9SL-2SR	**0.98**	**0.04**	**7.3**	0.56
21SL-2SR	0.94	0.05	7.6	0.56
1SL-21SR	0.72	0.062	8	0.56
21SL-21SR	0.61	0.066	7.9	0.56

In light of these observations, we discern a significant role for the structure of the communication signal in enhancing agents' explorative abilities. This phenomenon underscores the importance of fine-tuning communication parameters. It offers promising avenues for future research, particularly in examining how specific communication shapes may be tailored to suit various contextual requirements.

5.2.2 Linguistic Analysis

In the previous section, we discussed how the communication channel aids the villagers in coordinating their efforts to win over the werewolf. This achievement is undoubtedly noteworthy and forms the basis for our continued exploration. However, the specific content of the villagers' communications remains unclear.

Indeed, only analyzing performance is insufficient to gain an in-depth understanding of emergent coordination in artificial settings [Lowe et al., 2019]. While Brandizzi et al. [2021] did not examine the interpretation of the emerged language, later work by Lipinski et al. [2022] took a closer look. In their adaptation of *The Werewolf* environment, Lipinski et al. [2022] introduced two additional parameters for exploration: the number of communication rounds and the voting threshold. Retaining the parameters from the original study and implementing a grid search, they delved deeper into understanding the system dynamics. In specific game configurations, their agents showed performance improvements beyond both the theoretical baseline and the results demonstrated by Brandizzi et al. [2021].

During gameplay, a key observation was that villagers consistently repeated the same message in each communication round, voting off those who did not adhere to this strategy. This tactic echoes the principles of the Turing Test [Turing, 1950], with agents effectively distinguishing each other without human intervention. Upon investigating the emergent language, Lipinski et al. [2022] discovered a sparse vocabulary associated with successful strategies. Most winning agent populations used a single signal nearly 90% of the time, suggesting the development of a password-like system rather than a complex language. This might be attributed to the high efficiency and adequate performance achieved via this approach and falls in line with the literature highlighting how natural language does not emerge naturally in artificial contexts [Kottur et al., 2017].

Modifications to the original environment resulted in a significant impact on the convergence speed. The convergence point was defined as the episode where agents achieved a win rate exceeding 75%, indicating a successful strategy. Both the number of communication rounds and the voting threshold appeared to decrease the average episode count needed for convergence. Interestingly, the number of rounds showed a statistically significant effect on win rates and convergence speed, but the voting threshold's impact was insignificant.

5.3 Discussion on Emergent Machine Communication

In this chapter, we deal with the first question posed in §1.2, namely, *can effective communication emerge between artificial agents when presented with adequate learning pressures?*

In pursuit of an answer, we initially sought to understand the real-world pressures that drive human interaction, as detailed in §3.1. We advanced that these pressures stem from the presence of multiple agents and the necessity for cooperation in an environment where survival is otherwise untenable for solitary agents. This scenario presupposes a world with finite resources in which diverse groups or teams of agents compete for these resources. Having discerned the dual forces of cooperation and competition at play, we pinpointed an apt environment to simulate and analyze these dynamics.

For this purpose, we turned our attention to Social Deduction Game (SDG) as delineated in §3.1.3. These games create natural environments fostering both cooperation and competition and are heavily dependent on communication. Typically, the communication in these games needs to be instructive for allies while simultaneously evading the comprehension of adversaries. Upon general formalization of SDG, we identified our model environment as the game *The Werewolf*. Despite its extensive exploration in the context of facilitating gameplay between artificial agents and humans [Bi and Tanaka, 2016, Hirata et al., 2016, Katagami et al., 2014, Nakamura et al., 2016, Wang and Kaneko, 2018], the potential of this game as a vehicle for studying emergent communication among agents remained unexplored.

In §5.1, we introduced the framework developed by Brandizzi et al. [2021], establishing the groundwork for later analysis. This framework draws heavily on prior research on RL and MARL, setting up several experiments to test the feasibility of emergent effective communication. We cast the communication in a numerical, vectorized format, enabling the agent to map these numerical values to their action and observation space (see §5.1.1).

Our hypothesis was validated in §5.2, where we first examined the performance improvements in §5.2.1, demonstrating the capacity of agents to enhance their winning rates. While these results confirmed our initial question, they did not offer any additional insights into the nature of the emerged language. Thankfully, a subsequent study by Lipinski et al. [2022] analyzed the emerged communication, suggesting that it more closely resembled a strategy to identify enemies than a manifestation of natural language properties (see §5.2.2).

5.3.1 Human-interpretability in emerged languages

Given the numerical vector character of the language that arises, the challenge of human interpretability in emergent languages becomes apparent. Lipinski et al. [2022] identified that this emergent form of communication exhibited a lack of properties typically associated with natural human languages. Instead, it displayed the characteristics of a coding scheme specifically engineered to identify and exclude non-learning agents.

This communicative format might be efficient in the context of artificial agents' interactions, but it poses considerable barriers when considering broader applicability and integration with human communication systems. Hence, a gap persists between the languages used by artificial agents and those understandable to humans.

Bridging these differences is no trivial task, as it implies instilling artificial agents with an understanding of context, subtext, ambiguity, and the vast range of human emotions and intentions that language can convey. The problem stretches beyond the translation of numerical vectors into human language. Instead, it requires a more profound rethinking of how artificial agents are trained to communicate, potentially employing methods that mimic how humans learn languages. This would entail grounding words in perceptual experiences, learning from multimodal inputs, and understanding the use of language in social contexts.

Tackling this challenge will be our focus in the subsequent chapter, where we aim to train artificial agents that can meaningfully interact with humans using languages that are both effective and interpretable.

Chapter 6

Grounding Artificial Agents in Human Language

In the preceding Chapter 5, we demonstrated how effective communication[1] can spontaneously emerge among artificial agents under specific environmental pressures. These findings support our hypothesis that agency can develop under the right conditions, particularly in multi-agent systems where competition and cooperation coexist, and when reinforced by an adaptive learning paradigm such as Reinforcement Learning (RL).

However, allowing language to emerge from interactions within artificial settings often results in languages that are not interpretable by humans. This challenge arises because the primary training objective in these systems is geared towards performance. Specifically, the reward signal in RL is directed towards objectives that do not prioritize interpretability. More importantly, we recognize that when artificial agents communicate, it is not a mere exchange of information. Through interactive communication, they enact actions that can profoundly impact fellow agents and the surrounding environment.

Addressing these challenges, this chapter shifts focus towards exploring an integration of supervised and Reinforcement Learning. Our investigation is guided by the recent trend of Large Language Models (LLMs) using Reinforcement Learning for alignment as discussed in Section 3.2.3. We are interested in exploring the following research questions:

- How does teaching English to artificial agents through isolated supervised learning (without exposure to collaborative contexts or game-solving tasks) impact their ability to work together to solve a game? Additionally, what implications does this learning approach present regarding agency (§3.2.2)?
- Can Reinforcement Learning Fine-Tuning (RL-FT) enhance a pre-trained LLM 's adaptability to generate understandable expressions and adapt to other agents? How does RL modify the LLM 's parameters?

To address these questions, we engage with the problem formulation delineated in Section 4.2. Initially, in Section 6.1, we focus our attention towards a speaker model, optimizing solely the communication policy, denoted as π_C (see §4.2.2), in a typical NLP fashion utilizing exclusively

[1] Language designed to influence the environment and other agents.

supervised learning. This model, referred to as the General Speaker (G-Speak), demonstrates adequate efficacy on the Image Captioning task and exhibits performance metrics above random levels on the Referential Game (Ref Game) when paired with a listener.

Next, in Section 6.2, we shift our focus to the Asymmetric Referential Game (A-Ref Game). In this context, we observe a notable decrease in performance compared to the standard Ref Game. To address this challenge, we propose using Reinforcement Learning Fine-Tuning as a strategic countermeasure, enabling the speaker to adapt to a domain-specific listener. Interestingly, while the adaptation seems successful, reflected through performance metrics, a deeper qualitative analysis unveils a word distribution that has collapsed around a select few keywords that evoke the desired listener behavior. We correlate these observations with the well-documented issue of Language Drift in Emergent Communication (§2.2.2), along with the misalignment problem, discussed in Section 3.2.1.

6.1 Single-Agent Training for Collective Gameplay

This section details our experimental framework, distinctly modeled as a Referential Game, which sees two artificial agents (a speaker and a listener) within a visual environment encompassing multiple images, denoted as v_{ctx}. The game revolves around the speaker generating a communication signal, symbolized as c_s, with the intention of referring to a particular image, tagged as the target $\hat{t}$. Subsequently, the listener tries to identify the target image, solely guided by the provided signal, also referred to as the Referring Expression (RefEx).

The following sections will navigate through the components of our research. Initially, the dataset for our experiments will be introduced (§6.1.1). Following that, we discuss the models, specifically exploring the architectures employed for the listener and speaker (§6.1.3). Ultimately, we present the results of our analysis in §6.1.4.

6.1.1 PhotoBook Dataset

The visual context and captions in the Ref Game are taken from the PhotoBook dataset [Haber et al., 2019]. The dataset is based on a conversational game where two participants collaborate online through multiple rounds to identify images. In each round, they view a grid displaying six images. These images, derived from the MS COCO Dataset [Lin et al., 2014b], depict everyday scenes. Each participant's display has common images visible to both, as well as different images unique to each. Three images on every page are emphasized, marking them as target images. The objective is for participants to communicate using chat, aiming to classify these highlighted images as either common or different based on shared knowledge. Over five rounds, some images reappear, requiring participants to reference them repeatedly. The PhotoBook dataset captures interesting dialogue dynamics, as it contains multiple descriptions for each target image. This makes it a resource for studying cooperative behavior, especially in the context of collaborative Referring Expression and their alignment with conversational common ground.

We employ a version of the PhotoBook dataset curated by Takmaz et al. [2020]. This dataset comprises $41,340$ RefExs, each paired with the intended target image and five additional images that serve as the visual context.

6.1.2 Framework Specification

Before describing the models and their training regimes, as well as how they interact with each other, we need to ground the game in the formalization detailed in §4.2.

Action sets

As detailed in §4.1, the action set, denoted as A, comprises two distinct components: (i) a finite set of potential game actions, G, each element within this set is represented by g; (ii) a collection of unidirectional communication actions, $C_{i,j}(b)$, crafted to relay information $b \in B$ between two participating players.

Considering the referential context of our setup, game actions correspond to the image indices showcased in each round. Specifically, with six images displayed, G encompasses indices from 0 to 5, i.e., $G = \{0, 1, 2, 3, 4, 5\}$. When a player executes an action g that belongs to G, it signifies the listener's selection of one image from the available pool.

Contrastingly, the nature of the communication actions, $C_{i,j}(b)$, is more intricate. To begin with, and as highlighted in equation 4.6, such unidirectional actions exclusively proceed from the speaker toward the listener. Furthermore, the collection of feasible signals, B, arises from the interplay between the dataset's vocabulary size[2] (SR), which stands at *6,038*, and the upper limit for message length (SL), set at 30. Taking into account that word repetitions are allowed, the total possibilities can be expressed as:

$$|B| = SR^{SL} = 6,038^{30} = 2.67\text{e+}113$$

This vastness underscores the importance of equipping the speaker with the capability to craft precise communication signals, also known as Referring Expressions, to describe the target image appropriately.

Observation Set

Referencing §4.1, the observation set, denoted as O, comprises three distinct components: (i) E which captures the game's current status for each player; (ii) Γ that embodies an agent's internal states, such as beliefs that remain unobservable to their counterparts; and (iii) V, indicating the environmental states shared across all agents, unaffected by individual agent states.

Let's begin with V as it is relatively straightforward. The set V represents the six images presented to the listener. Formally, $V = \{v_1, v_2, \ldots, v_n\}$, where each v_i is defined as:

$$v_i = \{img_1, img_2, img_3, img_4, img_5, img_6\}$$

constituting a subset of 6 distinct images. These images are randomly selected from a larger pool containing 324 unique images. Given the significance of image positioning and the restriction against repetitions, the number of potential combinations for V is given by:

$$|V| = n = \binom{324}{5} = 3.46\text{e+}12$$

[2] This refers to the number of unique words within the dataset.

Turning our attention to the remaining components of the observation space, we first detail the significance of the Referring Expression produced by the speaker and subsequently relayed to the listener. Given that the RefEx (c_s) is an intrinsic attribute of the speaker, it naturally falls within the set E. Notably, the cardinality of this component aligns with that of the communication set B.

Conversely, Γ encompasses the listener's belief state upon being introduced to the image pool. Stemming from Assumption 2-$\mathcal{GSF}$, we can formalize the relationship between Γ and E as:

$$\Gamma \subseteq E \implies E = \Gamma \cup B$$

It is evident that Γ is a subset of E.

6.1.3 Model Desing for Multimodal Interaction

Engaging two agents, a speaker and a listener, the Ref Game necessitates collaborative communication to be successfully solved. Despite sharing a common objective, the agents occupy distinct roles that diverge in complexity and function. While both process multimodal input, the listener, only interpreting signals, requires a simpler design. In contrast, the speaker must generate descriptive captions, necessitating a more sophisticated architecture.

For the speaker, we employed LSTMs with sampling techniques, a decision aligned with methodologies established by Takmaz et al. [2020]. This alignment allows us to directly compare our results against established baselines and evaluate the incremental improvements of our approach. The choice of LSTMs was also driven by practical resource considerations, detailed in §8.2.1. Given the high computational demands of transformer models, LSTMs offered a more viable solution, balancing efficiency and effectiveness even with limited training data, where transformers may not reach their full potential without extensive resources. Thus, both the speaker and listener models, designed to process complex multimodal data, reflect a strategic compromise that ensures resource efficiency, aligns with prior research for comparative analysis, and maintains robust model performance across diverse inputs from visual and textual domains.

Image Encoding via ResNet-152

In an effort to maintain a linguistic-centric focus, we leverage ResNet-152 for image encoding [He et al., 2016b]. Ensuring all agents access identical features extracted through the same network establishes a consistent visual baseline across interactions. This eliminates the need for agents to learn feature representations independently, thereby accelerating training and mitigating potential inconsistencies in visual comprehension that could impact the experimental outcomes.

Listener Architecture

The listener model, Figure 6.1, is derived from Takmaz et al. [2020] and acts as a discriminator. It starts by processing two primary types of input: word embeddings and visual context (v_{ctx}). The word embeddings, representative of RefExs ($\hat{b}$), first undergo a dropout layer [Srivastava et al., 2014] for regularization and are then transformed by a linear layer activated by a Leaky-ReLU function [Xu et al., 2015], followed by normalization *(Linguistic Processing)*. Separately, the visual context, which encapsulates concatenated representations of six images, is similarly processed and

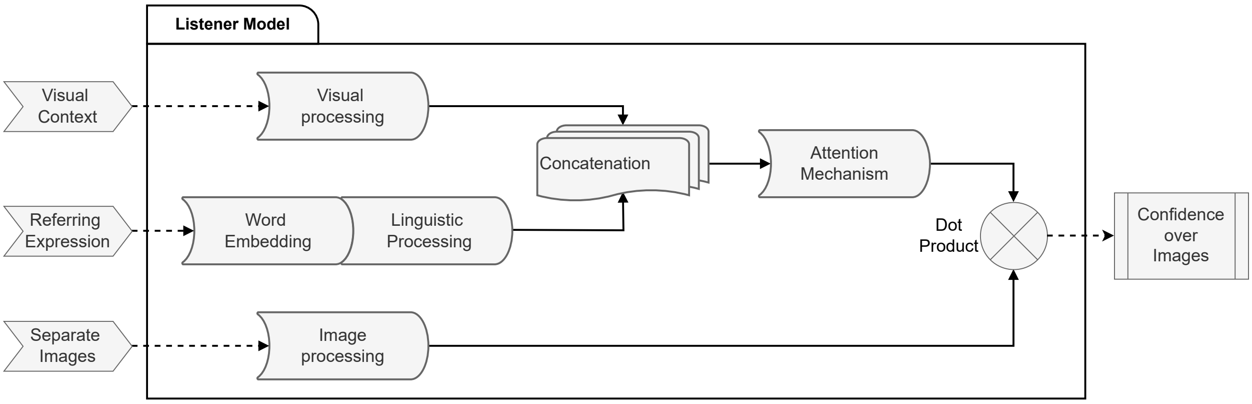


Figure 6.1: Architecture of the Listener Model, illustrating the processing pipeline from word embeddings and visual context to the final decision space.

standardized (*Visual Processing*). The individual representations of these images are also separately transformed and standardized (*Image Processing*). As the next phase begins, the word embeddings are integrated with the processed visual context. This joint representation is subjected to another linear transformation, activated by a ReLU function (*Concatenation*). An attention mechanism is then applied to these vectors, deriving an attention-weighted multimodal context vector. In the final step, this model aligns this context vector against the transformed representations of individual images via a dot product to identify the image most resonant with the RefEx and output a prediction (g_l).

Functionally, the listener model accepts a Referring Expression alongside a set of images and renders a decision regarding those images. The objective is to discern the image that aligns best with the provided caption. The formal representation is:

$$Listener(\hat{b}, v_{ctx}) \rightarrow g_l : B \times V \rightarrow G$$

where b represents the RefEx associated with the target, v_{ctx} is the set of images, and G is the decision space.

Training Regime and Results The training objective is to minimize the Cross Entropy (CE) loss between the chosen image and the target, and the optimization process is carried out with the Adam optimizer [Kingma and Ba, 2015]. Model performance is assessed based on resolution accuracy and Mean Reciprocal Rank (MRR) on the validation set, and the optimal model is selected accordingly. The listener was trained for ten epochs with batch size 64, a learning rate of 0.0001, and a dropout between layers set to 0.2. The results show a correct comprehension of the correlation between captions and dataset with an accuracy of 82.26% and MRR of 88.87%.

Speaker Architecture

The speaker model, shown in Figure 6.2, serves as a visually conditioned language model, designed to produce a RefEx (c_s) that aptly describes a target image ($\hat{t}$) within a provided visual context (v_{ctx}). It embodies a typical captioning model but is specialized for the context of the study, drawing inspiration from the work of Takmaz et al. [2020].

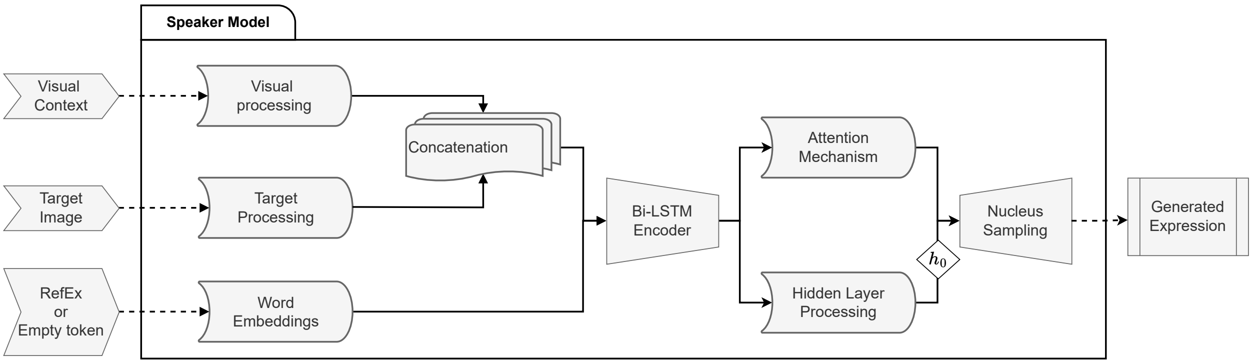


Figure 6.2: Architecture of the Visually Conditioned Speaker Model: From Visual Encoding to Referring Expression Generation

The foundation of the speaker is its encoder-decoder architecture, with the visual encoder processing and representing visual data. In the initial stages, two primary inputs are processed: the standardized target image vector and the concatenation of all six images in the full visual context (including the target image and five distractors). These are passed through a linear layer, followed by ReLU activation (*Visual & Target Processing*). Subsequently, the representations from the target image are concatenated with those from the visual context. To obtain the final visual context, another linear transformation augmented with ReLU non-linearity is applied.

The visual context serves as the initialization point for a bidirectional LSTM encoder. Depending on the dialogue, the encoder either takes in a previous RefEx related to the target image or a special token, which signifies the absence of any preceding Referring Expression. Upon processing, the final forward and backward hidden states from the encoder are concatenated and further transformed using a linear layer, complemented by a Hyperbolic Tangent Function (Tanh) non-linearity (*Hidden Layer Processing*).

This transformed output from the LSTM encoder sets the stage for the LSTM decoder by becoming its initial hidden state, denoted as h_0. Using nucleus sampling [Holtzman et al., 2020], with a *top-p* value of 0.9, the decoder then generates a RefEx. This process also involves the decoder paying attention to the encoder output at each step of the generation (*Attention Mechanism*), ensuring the coherence and relevance of the generated content.

Formally, the speaker represents a specific instance of the communication policy π_C, as defined in equation 4.9:

$$Speaker(v_{ctx}) \rightarrow c_s : V \rightarrow C$$

Here, only the communication subset is considered instead of the full observation set O.

Training Regime and Results The speaker model is trained from scratch, intentionally avoiding the use of pretrained embeddings to retain complete control over the model's knowledge. During each training instance, the speaker model is presented with a set of six images. Leveraging its architecture, the model's task is to generate an apt caption for the target image. The training uses the CE Loss function between the model's generation and the reference RefEx from the dataset. To achieve this, we employ teacher forcing [Lamb et al., 2016], feeding the true caption values back into the LSTM at every stage. As for the listener, we utilize the Adam optimizer [Kingma and Ba, 2015]. To evaluate the model's efficiency, we assess it against multiple Natural Language

BLEU-1	**BLEU-2**	**BLEU-3**	**BLEU-4**	**Rouge**
40.06 ± 1.60	23.81 ± 1.51	14.09 ± 1.20	8.46 ± 0.89	32.92 ± 0.93
CIDEr	**BertScore R**	**BertScore F1**	**BertScore P**	
44.07 ± 1.68	58.91 ± 0.19	57.7 ± 0.12	57.9 ± 0.16	

Table 6.1: Performance metrics of the speaker model on the test set. The table presents average values and standard deviations from four separate test runs, evaluating the model across various NLP metrics.

Processing metrics on the validation set, including BLEU [Papineni et al., 2002], ROUGE [Lin, 2004], CIDEr [Vedantam et al., 2015], and BERTScore [Zhang et al., 2020b]. The model selection is based on a composite score calculated as the average of these metrics for each model, ensuring a balanced consideration of all aspects of linguistic quality. We selected the model with the highest composite score, which represents the best average performance across all metrics, for final implementation. After identifying the best-performing model, its weights are frozen, ensuring that this model becomes the standard language generator for all subsequent experiments. The performance of the speaker model on the test set is documented in Table 6.1. This table presents both the average values and their corresponding standard deviations from four separate runs on the test dataset.

Compared to classic natural language generation metrics, the speaker showcases adequate performance and are in line with those reported by Takmaz et al. [2020] using their 'Ref' model, which we use as our baseline reference.

6.1.4 Evaluating Linguistic and Strategic Outcomes in Referential Gameplay

The analysis presented in this section focuses on two areas: performance metrics, with a spotlight on accuracy, and a linguistic investigation of the Referring Expressions crafted by the speaker. This analysis is framed within the context of the cooperative game between the pretrained speaker and listener. Results are derived from an average of five Ref Games, each initiated with different seed values, ensuring robustness and reliability.

Performance Analysis

The accuracy is not only an indicator of correct answers but also reveals distinctions regarding the agents' capability to align their independent learnings to solve the game cooperatively.

In a game where the speaker and listener play together, the average accuracy is $50.37\% \pm 0.5$. This significantly outperforms the random baseline of 16%, remarking the agents' capacity to solve the game despite being devoid of specific cooperative training. Their success is rooted in their ability to converge upon similar representations independently. To further analyze the quality of the RefExs, a game scenario where the speaker utters random words is explored. In this setup, the accuracy decreased to $18.8\% \pm 0.4$, proximate to random chance, showing the importance of sense-full RefExs.

Linguistic Analysis

We derive insights from conventional linguistic metrics, focusing on the Type-Token Ratio (TTR) and Part of Speech (PoS) analysis, which respectively offer an understanding of the diversity and

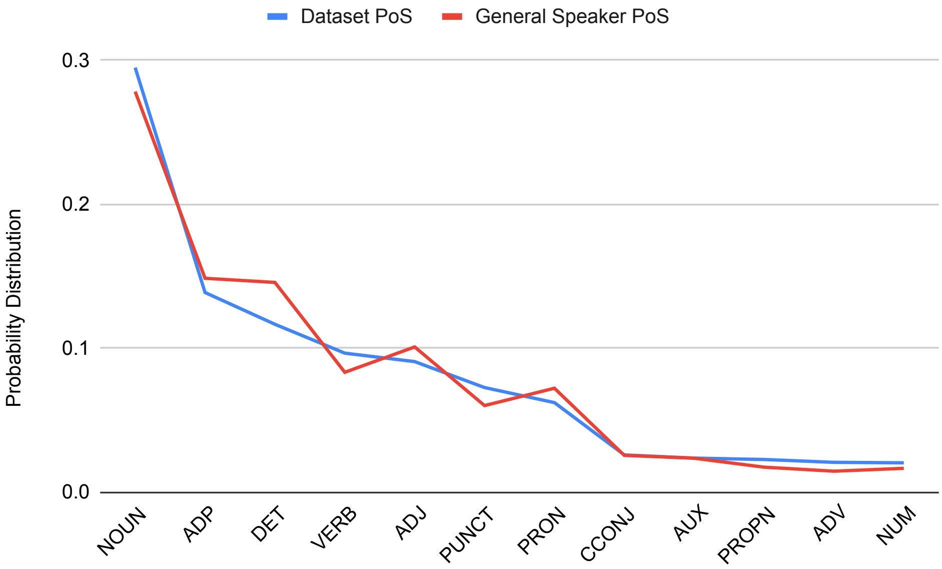


Figure 6.3: Shifting distributions in unigram Part of Speech between the test split of the dataset (blue) and speaker-generated Referring Expressions (red).

variational aspects of employed vocabulary and an analysis of individual words through their grammatical roles, such as nouns, verbs, and determiners.

Type-Token Ratio The TTR offers a glimpse into vocabulary diversity. The analysis here pertains to the language generated by the speaker on the test subset of the Photobook dataset. Initially, we analyze this subset's TTR, showing a value of 4.48%, suggesting that participants in the dataset creation generally adhered to a limited array of unique tokens. Interestingly, our speaker exhibits a lower value: 2.9% ± 0.02, indicative of a tendency towards approximate language use. This potentially underscores a strategy of the speaker to sidestep outlier captions during learning, skewing towards median responses.

Unigram Part of Speech Tagging Figure 6.3 presents the PoS distributions for both the test subset of the Photobook dataset and the speaker-generated RefEx. Notably, most of the PoS elements exhibit minimal differences, with a few exceptions. Specifically, there is an increase in determiner usage[3] (by 11.1%) and a marked reduction in both proper noun[4] and adverb[5] usage, decreased by 13.43% and 17.62% respectively. These linguistic shifts hint towards a generalized, less specified language strategy, approximating the authentic word distribution within the dataset.

6.1.5 Discussion on Static Agent Communication

Investigating supervised language models' linguistic and cooperative capacities confirms how agents can learn human language in line with state-of-the-art methods. However, our findings extend this knowledge, illustrating that this learning enables two artificial agents to communicate and

[3] A determiner refers to a word that references a noun or noun phrase, e.g., they, this, my, many.

[4] A proper noun signifies a noun that denotes a singular entity and is utilized to refer to that specific entity.

[5] An adverb broadly modifies a verb, typically conveying a level of certainty, and answers questions like how, when, where, and to what extent.

navigate a Referential Game with above-random accuracy. This ability stems from the agents' independent yet paralleled approximation of the visual and textual domains, fostering a shared language understanding and enabling collaborative problem-solving within the gameplay.

Several issues emerge from our linguistic analysis, notably the speaker's tendency towards a generalized language use that sidesteps linguistic outliers, potentially sacrificing communicative richness. Moreover, the speaker exhibits a lack of agency. The communication actions, devoid of inherent strategy, only adhere to the problem's formulation, with the speaker maintaining a static word distribution, unaware of its collaborative role in informing another agent.

6.2 Enhancing Communication via Reinforcement Learning Fine-Tuning

Our earlier discussion focused on the abilities of artificial agents to learn a human language. Specifically, utilizing English to solve a Referential Game through means of supervised learning only. Here, we address identified limitations, particularly focusing on the speaker's evident lack of agency and strategic communication. This section thus explores the potential of Reinforcement Learning to facilitate the speaker's adaptation to an agent, examining how this could improve communicative effectiveness and strategy.

In the preceding experiment, we utilized images from the PhotoBook Dataset (see §6.1.1) without differentiating among them based on their content or context. The following experiment, however, introduces the concept of domains within which each image resides based on the depicted content. This distinction allows us to explore *knowledge asymmetry* by training the speaker on the whole dataset, while domain-specific images are used for training the listener. Under this setup, our hypothesis posits that to achieve effective communication; the speaker must synchronize its RefExs with the listener's more specific understanding. We seek to facilitate this synchronization by fine-tuning the speaker through RL while maintaining the listener's model parameters unchanged during experimentation. In this context, we refer to the fine-tuned speaker as *Reinforcement Learning Fine-Tuning Speaker (RL-Speak)*, while the speaker from §6.1 will be referred as the *General Speaker (G-Speak)*.

The subsequent sections will dissect our approach and findings as follows:

- Section 6.2.1 presents the introduction of asymmetry in the Referential Game.
- Section 6.2.2 describes the architectural and experimental modifications necessary for the asymmetric framework to take place.
- In Section 6.2.3, we explore how Reinforcement Learning Fine-Tuning improves performance but leads to a simplification of RefEx to single words.

6.2.1 Exploring Knowledge Asymmetry and Mitigation Strategies

To investigate adaptability in the context of language modeling, modifications to our Referential Game are necessary. Knowledge asymmetry is introduced by partitioning the PhotoBook dataset into five distinct domains and subsequently training a listener specific to each domain. While the General Speaker is uniformly trained on the entire dataset and thereby frozen, the domain specificity

Domain	Prop	N	$\|V\|$	Images	Specific	Overlap
Appliances	9.4%	4,310	1,271	36	29.5%	23.2% (*Ind*)
Food	12.4%	5,682	1,646	36	43.3%	22.9% (*App*)
Indoor	26.4%	12,088	2,477	96	44.3%	26.0% (*Out*)
Outdoor	35.9%	16,427	2,858	108	47.0%	26.2% (*Veh*)
Vehicles	15.8%	7,234	1,738	48	36.0%	26.2% (*Out*)
All	100%	45,741	6,038	324	-	-

Table 6.2: Characteristics of domain-focused datasets: Total utterances count (N), proportion within the full dataset (Prop), size of vocabulary ($|V|$), count of distinctive images (Images), fraction of domain-exclusive vocabulary (Specific), and the highest lexical commonality with a different domain (Overlap). The most significant overlap exists between *outdoor* and *vehicles*, with shared terms like *'left'*, *'black'*, *'driving'*, and *'glasses'*.

of the listener accentuates challenges in comprehending the G-Speak, thus necessitating strategic adaptation.

Constructing a Domain-Specific PhotoBook Dataset

As mentioned in Section 6.1.1, the PhotoBook Dataset, encompassing over 41,000 unique Referring Expressions, utilizes images sourced from COCO dataset [Lin et al., 2014a]. The latter spans across *30* distinctive visual domains. In an effort to explore speaker adaptation across varied semantic contexts, the PhotoBook dataset's Referring Expressions is categorized according to their corresponding visual domain. The domains are grouped based on the similarity of their vocabulary vectors, derived from word frequency counts within expressions from that domain. This approach results in five broad categories: *appliances*, *food*, *indoor*, *outdoor*, and *vehicles*. Each category was chosen to ensure minimal vocabulary overlap, as in Table 6.2. Subsequently, the RefExs and their visual context are extracted for every visual domain cluster. These are then partitioned into training (70%), validation (15%), and test datasets (15%).

Asymmetric Referential Game

In this setup, the G-Speak model possesses greater domain knowledge than the listener. This intentional disparity enables us to investigate adaptation. To facilitate this, we train six distinct listeners for each of the domain splits (*appliances*, *food*, *indoor*, *outdoor*, *vehicles*, and *all-domains*). Conversely, the G-Speak is trained across all these domains, thus being well-versed in each one. This setup ensures that only the relevant listener can comprehend the speaker's RefEx depending on the prevailing image domain. An illustrative example is provided in Figure 6.4. Here, when the domain of *food* is the visual context and the G-Speak generates the caption "green salad", all listeners except the one trained specifically on the *food* domain (i.e., the *food-listener*) get confused, leading to erroneous predictions.

Asymmetric Accuracy Before discussing the effect of asymmetry on the listeners' results, we introduce two metrics that arise from the asymmetric nature of the A-Ref Game game:

- IN-Domain (IND) Accuracy: This metric measures the accuracy achieved on the test set corresponding to the domain the listener was trained on, e.g., a listener trained on the *food*

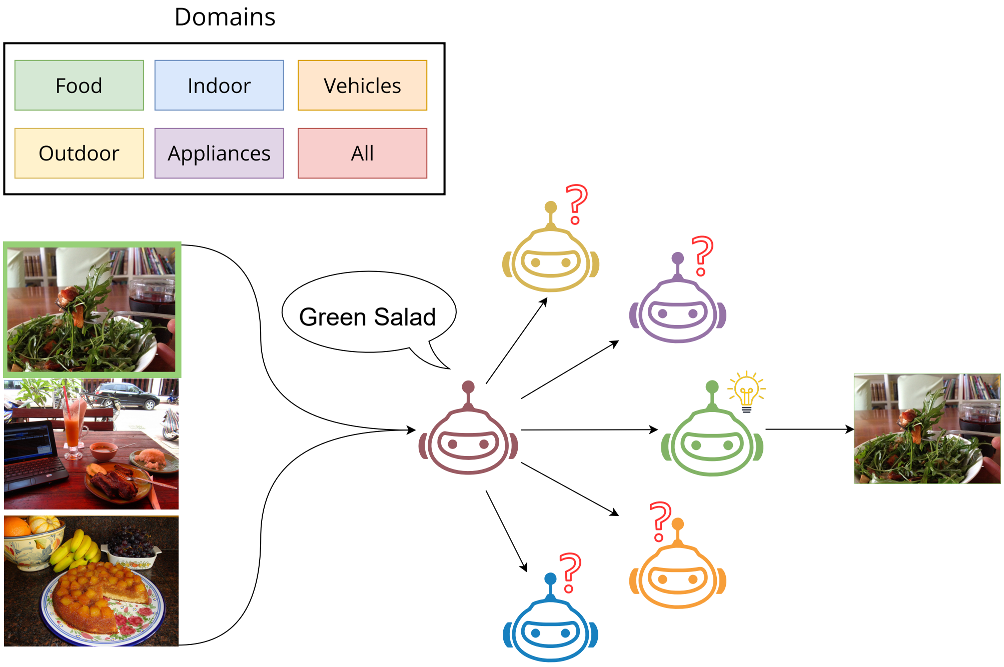


Figure 6.4: Illustration of Asymmetric Referential Game: While the speaker's caption "green salad" is clear in the context of the food domain, it confuses listeners trained in other domains.

Domain	Epoch	IND		OOD	
		Accuracy	MRR	Accuracy	MRR
Appliances	23	84.12 ± 0.33	90.27 ± 0.10	20.28 ± 0.23	44.07 ± 0.11
Food	21	85.40 ± 0.28	91.20 ± 0.20	17.72 ± 0.18	42.42 ± 0.06
Indoor	14	82.94 ± 0.13	89.32 ± 0.09	19.14 ± 0.09	43.46 ± 0.06
Outdoor	19	83.96 ± 0.23	90.01 ± 0.14	19.64 ± 0.07	43.52 ± 0.06
Vehicles	17	78.99 ± 0.35	86.81 ± 0.14	18.46 ± 0.28	42.36 ± 0.20
Average	18.8	83.08 ± 0.26	89.52 ± 0.13	19.05 ± 0.17	43.16 ± 0.09

Table 6.3: Performance metrics of listeners on training utterances. The table shows Accuracy and Mean Reciprocal Rank (MRR) values for both IN-Domain (IND) and Out Of Distribution (OOD) samples, corresponding to listeners trained within designated domains (highlighted in the 'Domain' column).

domain (*food*-listener) evaluated on *food* data samples.

- Out Of Distribution (OOD) Accuracy: This metric evaluates the accuracy on domains the listener has not been previously introduced. An example would be assessing the accuracy of images from the *vehicles* domain for a listener solely trained on the *food* domain.

Listeners Performance

To show the effects of the Asymmetric Referential Game on the game outcome, we report the listeners' accuracy in Table 6.3.

Clearly, our listener models exhibit domain specificity. Indeed, the average IND accuracy stands at 83.08% ± 0.26. In contrast, the OOD accuracy is averaged at 19.05% ± 0.17, marginally higher than a random baseline of 16%. The table also indicates the epoch at which training concluded due to convergence, with an average epoch of 18.8.

Listener			Data domain		
domain	appliances	food	indoor	outdoor	vehicles
appliances	**57.61** ± 1.38	20.10 ± 0.63	19.92 ± 0.47	21.27 ± 0.83	15.98 ± 0.82
food	19.11 ± 1.70	**54.29** ± 1.06	18.60 ± 0.84	18.85 ± 0.49	18.85 ± 0.49
indoor	22.71 ± 1.30	19.65 ± 1.77	**53.62** ± 0.79	20.82 ± 1.05	16.77 ± 0.79
outdoor	15.08 ± 1.04	21.46 ± 0.70	19.62 ± 0.69	**52.93** ± 1.11	17.69 ± 0.97
vehicles	16.36 ± 1.55	16.17 ± 0.81	17.41 ± 0.64	20.13 ± 0.59	**43.08** ± 1.16

Table 6.4: Evaluation of listeners based on speaker-produced data. Every row represents a unique listener trained within a particular domain, with columns indicating the assessment domain. Entries within the table represent the mean values across five seeds, where listeners of a particular domain are evaluated.

Furthermore, as we aim to discuss later the interplay between the RL-Speak and listener, it is essential to analyze the listener's performance on RefExs not derived from the training data. Instead, we focus on RefExs generated by the G-Speak, represented as c_s. To this end, Table 6.4 outlines the listener accuracies on G-Speak inputs. Notably, in IND scenarios, we observe diminished scores relative to the usage of the training data (83.08% *vs.* 51.7%), on average a loss of 31% accuracy. This discrepancy is likely due to the different word distribution between the training data and the G-Speak 's generated RefExs. Contrarily, the OOD accuracy stays roughly the same (19.05% *vs* 18.82%).

6.2.2 Experimental Setup

Given the drop in performances reported previously, this section details the experimental settings and strategies used to mitigate these challenges. We first explore the issue of handling out-of-distribution words in listener embeddings and introduce a masking strategy to manage this. Subsequently, we provide a comprehensive breakdown of our Reinforcement Learning Fine-Tuning strategy, detailing the construction of the loss function and the fine-tuning process for the RL-Speak model.

Mitigating Out-of-Distribution Words via Masking

All listeners begin with an identical vocabulary that encompasses every word from the training data. However, certain words are unique to specific domains and are not learned by listeners outside those domains, as outlined in Table 6.2, which details vocabulary overlaps. This means that while these unique words exist in the vocabulary of every listener, their embeddings are randomly initialized in domains where they do not appear and, consequently, are not updated. This situation was observed to impact the performance of our listener adversely. Specifically, when the G-Speak generates RefExs for a specific domain, these words can mislead the listener with their random embeddings.

To address this issue, we set these words to the *unk* (unknown) token upon completing the listener's training. This decision was motivated by the idea that it would familiarize the listener with the notion of unknown words. It should be noted, however, that our methodology does not truncate the training set vocabulary based on frequency, which means listeners are not exposed to *unknowns* during training. To ensure consistency, the *unk* token is initialized in the same way across all domains when a uniform seed is used. Thus, all domain-specific listeners are effectively masked with an identical vector.

Fine-tuning Strategy

Given the discrete nature of the decoding process in the speaker model, traditional end-to-end backpropagation is not feasible for optimizing the speaker's output (for a detailed discussion, see §3.3.3). To circumvent this limitation, our approach draws inspiration from recent advancements in the field [Dessì et al., 2023]. We employ the REINFORCE algorithm [Williams, 1992], combined with a custom reward function built around the RL-Speak 's decoded logits.

Our reward function (R) integrates multiple loss components designed to optimize various aspects of the model's performance:

$$R(h_t, g_l, \hat{t}, c_s, v_{ctx}) = \alpha_{\pi_C} L^{\pi_C} + \alpha_{P_{lst}} L^{P_{lst}} + \alpha_S L^S + \alpha_{KL} L^{KL}$$

The following paragraphs offer detailed formulations and interpretations of these individual loss components.

Policy Loss In our framework, the policy loss L^{π_C} is a measure of how well the communication policy performs in generating the RL-Speak 's Referring Expression (RefEx), denoted as c_s. This loss is calculated by evaluating the logarithm of the softmax probabilities of the decoded logits, referred to as h_t, corresponding to the actual indexes of the chosen words ($h_t^{c_s}$):

$$L^{\pi_C} = \frac{1}{length} \sum_{j=1}^{SL} mask_j \left[\log \left(\frac{\exp(h_t^{c_s,j})}{\sum_{i=1}^{B} \exp(h_t^{i,j})} \right) \right]$$

To focus the loss computation only on meaningful parts of the sequence (RefEx), we employ a binary mask ($mask_j$), which is set to one if the label for the j-th words in the RefEx is non-zero, otherwise zero. This serves to effectively ignore the irrelevant portions of the sequence in the loss computation.

$$mask_j = \begin{cases} 1, & \text{if } c_s, j \neq 0 \\ 0, & \text{otherwise} \end{cases}$$

We also normalize the sum of the masked log probabilities by an effective sequence length ($length$). This length is the sum of the binary mask values but is adjusted to be at least one to prevent division by zero.

$$length = \max \left(1, \sum_{j=1}^{SL} mask_j \right)$$

Thus, L^{π_C} essentially captures the average log probability of the actual RefExs under the model's current communication c_s, filtered and normalized by the aforementioned terms.

Listener Loss The listener loss $L^{P_{lst}}$ is meant to optimize the listener's accuracy in the Referential Game[6]. By doing so, this loss steers the RL-Speak model in the direction of generating RefExs that are more effectively interpreted by the pretrained listener. Specifically, the RL-Speak aims

[6] In our framework $L^{P_{lst}}$ does not directly optimize the listener model, which remains frozen during this phase.

to produce RefExs that facilitate accurate and confident identification of the true target $\hat{t}$ by the listener within the given visual context v_{ctx}.

Formally, the listener loss is defined as:

$$L^{P_{lst}} = -\sum_{i=1}^{|G|} \hat{t}^i \log \left(\frac{\exp(g_l^i)}{\sum_{j=1}^{|G|} \exp(g_l^j)} \right)$$

In this equation, $g_l = \text{Listener}(c_s, v_{ctx})$ represents the listener's output, a vector in G, which is the set of all possible game-related actions. The model generates this output when given an utterance c_s and visual context v_{ctx}.

The term $\hat{t}$ is a one-hot encoded vector, where each element corresponds to a possible action in G. The element corresponding to the true action is set to 1, and all other elements are set to 0.

Thus, minimizing $L^{P_{lst}}$ essentially steers the RL-Speak model to generate utterances that not only make the listener more accurate but also more confident in picking the true target $\hat{t}$ from the set G.

Entropy Loss An entropy regularization term is included in the objective function to encourage exploration in the action space [Mnih et al., 2016, Williams and Peng, 1991]. By favoring higher entropy, this term discourages the RL-Speak model from prematurely converging to a sub-optimal communication policy.

The entropy loss can be formally defined as follows:

$$L^S = -\frac{1}{|B|} \sum_{i=1}^{|B|} \left[\exp(h_t^i) \times \log \left(\exp(h_t^i) \right) \right]$$

The entropy is calculated for the categorical distribution defined by h_t, and the mean of this entropy across all possible RefEx is taken as the entropy loss L^S.

Kullback–Leibler Loss The Kullback-Leibler Divergence (KL-Divergence) [Kullback and Leibler, 1951] loss is employed to ensure that the distribution of the RL-Speak model's Referring Expression does not deviate excessively from an original, reference distribution [Chaabouni et al., 2022]. This is useful for constraining the model to stay within a predefined "good" behavior while allowing for improvements.

The KL loss can be mathematically defined as:

$$L^{KL} = \frac{1}{|B|} \sum_{i=1}^{|B|} \left(\log \left(\frac{\exp(h_t^i)}{\exp(h_t^{\text{original},i})} \right) \right) \times \exp(h_t^{\text{original},i})$$

Here, h_t represents the decoder's output by the current iteration of the RL-Speak model, while h_t^{original} is the decoder's output obtained from the G-Speak, both vectors in B.

Training Details

Building on the outlined Reinforcement Learning Fine-Tuning strategy, we report the specificities of the training process that yielded the results explored in the following section. The RL-Speak

Methodology	Data domain				
	appliances	food	indoor	outdoor	vehicles
Train Data	17.25	**85.28**	18.80	18.65	16.77
G-Speak	19.11	54.29	18.60	18.85	18.85
RL-Speak	**41.83**	79.08	**35.29**	**34.64**	**32.03**

Table 6.5: Evaluation of food-listeners based on different methodologies. The first row reports the accuracy on training set, the second is for the General Speaker (G-Speak) (Table 6.4) and finally the third row shows results for the Reinforcement Learning Fine-Tuning Speaker (RL-Speak).

model was trained for 300 epochs[7] with AdamW optimizer [Loshchilov and Hutter, 2019] and a learning rate of 0.0001. The Cosine Annealing improved the optimization with the Warm Restarts scheduling strategy [Loshchilov and Hutter, 2017], which incrementally amplified the learning rate by a magnitude of 2. Regarding the reward weights, we use 0.1 for α_S, 0.001 for α_{KL}[8], and values of 1.0 for both α_{π_C} and $\alpha_{P_{lst}}$.

6.2.3 Analysis

This section presents the results derived from the interaction between a specific listener within the food domain and the RL-Speak tailored for it. The choice to analyze the behavior within just one domain stems from the computationally intensive nature of the Reinforcement Learning Fine-Tuning process required for the speaker. It is important to note that fine-tuning Large Language Models generally has substantial environmental implications, often leading to increased carbon footprints and energy consumption [Rillig et al., 2023].

Quantitative Analysis

Here, we focus on the interaction between a food-listener and the Reinforcement Learning Fine-Tuning Speaker. We investigate how this fine-tuning technique modifies language distributions, examining metrics like the Type-Token Ratio and unigram Part of Speech (PoS). We also explore the probability distribution dynamics of the encoder and decoder during training. Finally, we present outcomes from the interaction of the RL-Speak with listeners from other domains, assessing the method's generality.

Referential game accuracy Table 6.5 offers a comparative listener performance analysis under three conditions: training dataset, Asymmetric Referential Game with a G-Speak, and a RL-Speak. Remarkably, the listener performs better with captions from the RL-Speak than G-Speak. Specifically, the Out Of Distribution accuracy jumps to 35.77%, contrasted with 19.05% from training data and 18.82% with the G-Speak. Additionally, the IN-Domain (IND) accuracy sees a rise of 24.79% (54.29% to 79.08%), though it is still 6.2% below training accuracy.

Linguistic Analysis Contrary to the findings in §6.1.4, Reinforcement Learning Fine-Tuning enhances the type ratio value. The increase is by two percentage points compared to the G-Speak,

[7] Each epoch consists of 256 episodes with a batch size of 32.

[8] Throughout our experiments, we discovered that increasing the α_{KL} value inhibited variation to the speaker's weights, even after an extensive training time.

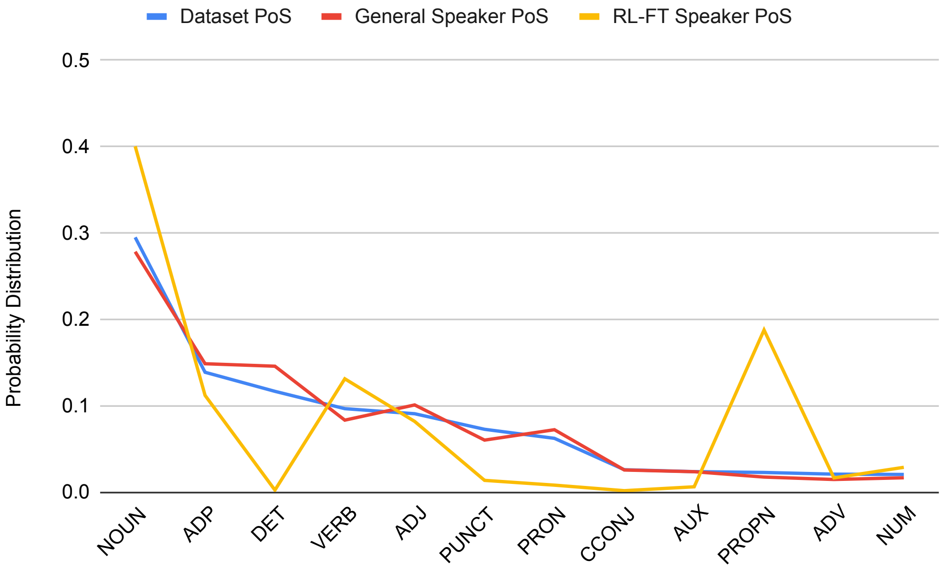
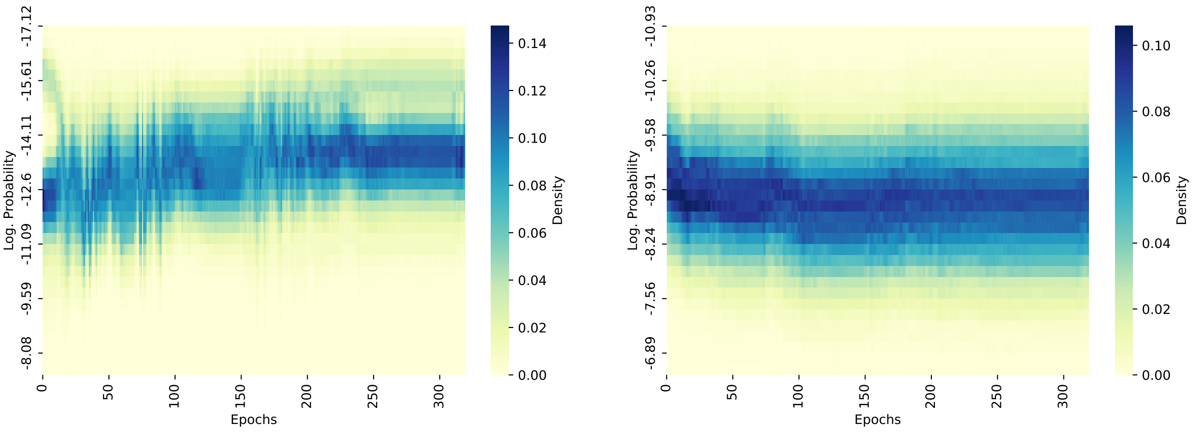


Figure 6.5: Distributional shifts in unigram Part of Speech between dataset test split (blue), General Speaker (red) and General Speaker (yellow).

settling at 5.54% against 2.9%. This signifies a token diversity boost of 65.63%. Interestingly, the TTR surpasses the test set value (4.48%). This metric assesses vocabulary diversity but also depends on the Referring Expression length. As we will see, Reinforcement Learning Fine-Tuning tends to produce shorter RefExs, potentially skewing the metric and creating misleading interpretations.

Furthermore, Figure 6.5 displays the shifts in unigram PoS distribution. Some variations merit attention. For instance, there's a notable rise in proper noun usage (1.94% to 18.76%) and significant drops in determiners (14.53% to 0.20%), coordinating conjunctions (2.57% to 0.14%), and pronouns (7.26% to 0.78%). These patterns suggest a RL-Speak strategy emphasizing keywords to elicit specific listener responses, potentially compromising grammatical accuracy.



(a) Decoder's Probability Distribution

(b) Encoder's Probability Distribution

Figure 6.6: Probability distribution of tokens in logarithmic space: (a) Decoder and (b) Encoder, illustrating the effects of Reinforcement Learning Fine-Tuning throughout training.

Effect of RL on language distribution To further explore the impacts of RL-FT on linguistic distribution, Figure 6.6 reports the probability distribution of tokens (expressed in logarithmic

Listener			**Data domain**		
domain	**appliances**	**food**	**indoor**	**outdoor**	**vehicles**
appliances	**27.46** $\pm$ 0.32	19.74 $\pm$ 0.16	21.39 $\pm$ 0.11	18.28 $\pm$ 0.64	20.11 $\pm$ 0.26
food	41.83 $\pm$ 5.99	**79.08** $\pm$ 2.26	35.29 $\pm$ 8.98	34.64 $\pm$ 2.99	32.03 $\pm$ 4.52
indoor	19.21 $\pm$ 1.28	16.53 $\pm$ 1.23	**37.85** $\pm$ 0.11	19.50 $\pm$ 0.13	19.88 $\pm$ 0.58
outdoor	19.36 $\pm$ 0.21	21.73 $\pm$ 1.48	24.32 $\pm$ 0.65	**29.79** $\pm$ 0.50	19.51 $\pm$ 0.32
vehicles	15.96 $\pm$ 0.32	17.17 $\pm$ 0.49	20.82 $\pm$ 0.38	22.08 $\pm$ 0.11	**31.09** $\pm$ 0.45

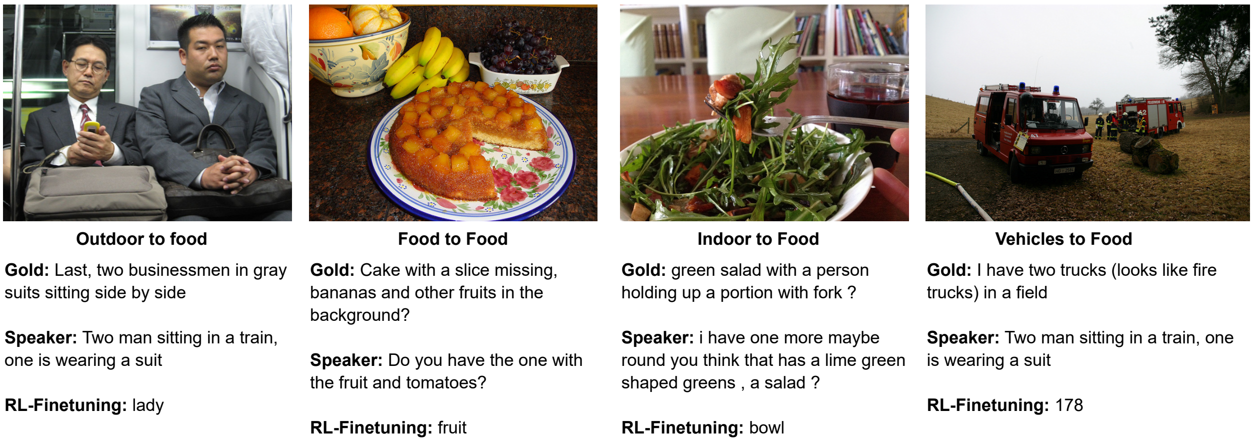
Table 6.6: Comparison of listener performances when paired with a RL-finetuned speaker on domain food. The table reports average accuracies and standard deviatoins, emphasizing the versatility and adaptability of the finetuned speaker despite its domain-specific training.

space) for both the decoder (6.6a) and the encoder (6.6b) throughout the training process. Notably, a discernible difference is observed in the decoder's probability distribution. Initially, the distribution has a mean of $\mu_1 = -12.90$ and a standard deviation of $\sigma_1 = 1.61$. Upon conclusion, these values transitioned to a mean of $\mu_2 = -9.52$ and a standard deviation of $\sigma_2 = 0.18$. This transformation reflects a differential of $\Delta\mu = 26.20\%$ and $\Delta\sigma = 88.81\%$, coupled with a notable Cohen's distance of 2.93. The reduced mean and standard deviation are coherent with what has been seen so far, i.e., a raised focus on particular words that become increasingly more probable while diversity in the vocabulary decreases.

In contrast, the encoder's distribution parameters remain relatively stable. Before training the distribution is defined by $\mu_1 = -9.12$ and $\sigma_1 = 0.56$. At termination, these metrics slightly adjusted to $\mu_2 = -9.09$ and $\sigma_2 = 0.53$ respectively. This marginal shift, characterized by a difference of $\Delta\mu = 0.32\%$ and $\Delta\sigma = 5.35\%$, registers a small Cohen's distance of 0.05. Such stability suggests that fine-tuning does not drastically alter the encoder's distribution. This behavior aligns with theoretical expectations, given that the policy loss, L^{π_C}, directly influences the decoder output while only marginally affecting the encoder[9].

Assessment with Alternative Listeners In an effort to evaluate the versatility of the RL-Speak, we tested it in conjunction with several domain-specific listeners, with the results detailed in Table 6.6. An examination reveals that, on average, these listeners performed slightly better with the RL-Speak (28.80% ± 1.86) in comparison to their performance with a non-fine-tuned counterpart (25.52% ± 0.95)[10]. Specifically, the Out Of Distribution accuracy saw an increase (18.82%±0.91 to 22.88%±0.48), whereas the IN-Domain accuracy registered a decline (51.07%±1.10 to 31.54% ± 0.34).

This observation is particularly interesting, considering the speaker's fine-tuning was conducted with a food domain-specific listener. Despite this specificity, the RL-Speak demonstrated adeptness in generating better referential expressions for diverse listeners. However, as reported in the next section on qualitative analysis, this approach culminates in a linguistic distribution that not only deviates from the original but also diminishes its interpretability from a human perspective.



Outdoor to food

Gold: Last, two businessmen in gray suits sitting side by side

Speaker: Two man sitting in a train, one is wearing a suit

RL-Finetuning: lady

Food to Food

Gold: Cake with a slice missing, bananas and other fruits in the background?

Speaker: Do you have the one with the fruit and tomatoes?

RL-Finetuning: fruit

Indoor to Food

Gold: green salad with a person holding up a portion with fork ?

Speaker: i have one more maybe round you think that has a lime green shaped greens , a salad ?

RL-Finetuning: bowl

Vehicles to Food

Gold: I have two trucks (looks like fire trucks) in a field

Speaker: Two man sitting in a train, one is wearing a suit

RL-Finetuning: 178

Figure 6.7: Illustration of Referring Expression outcomes post Reinforcement Learning Fine-Tuning (RL-FT). Each image is paired with a referential expression generated by the RL-FT speaker.

Qualitative Analysis

As mentioned in the related work §3.2.3, focusing on metrics only does not necessarily provide a complete understanding of the results, especially when those metrics are also used in the optimization process. For this reason, instances from successful Asymmetric Referential Game turns are shown and analyzed in this section. These instances are randomly selected from scenarios wherein the General Speaker struggled to convey the intended message to the listener. However, these outcomes became positive results with Reinforcement Learning Fine-Tuning. As observed in Figure 6.7, RL-FT predominantly yields Referring Expressions composed of singular words. While this aligns with prior analyses, it remains interesting to observe these effects in tandem with actual images. Some examples, like "fruit" and "bowl", resonate with the chosen visual target. In contrast, others, such as "lady" and "178" appear somewhat out of place given the visual context. As highlighted earlier, the primary objective of fine-tuning aligns with optimizing listener performance in the A-Ref Game, which, in turn, prompts the RL-Speak to gravitate towards specific keywords that elicit the desired response. This phenomenon correlates with challenges mentioned in related works (see §3.2.3), where the objective metric may not comprehend aspects like human interpretability or adherence to the inherent grammatical structure of a RefEx. While introducing more complex metrics could offer some mitigation, the misalignment issue persists, proving to be both elusive and subtle in nature.

It is essential to recognize that a purely quantitative analysis might depict an optimal performance. However, without a qualitative lens, one might overlook such misalignments. In the context of our experiments, these inconsistencies are relatively discernible, given the model's small size. Yet, when navigating models with billions of parameters, such as contemporary Large Language Models, detecting these misalignments becomes considerably more challenging.

[9] Preliminary experimentation, wherein the encoder output was incorporated into the policy loss, culminated in suboptimal performances.

[10] In these estimations, the IND accuracy of the food-listener was excluded from consideration.

6.3 Reflecting on Findings

In our pursuit of more human-like AI communication, this chapter has probed the capability of artificial agents to utilize human language within an interactive context. We posited that a fusion of supervised and Reinforcement Learning could spawn an emergent language that is both action-aware and human-interpretable, consistent with the findings detailed in the previous Chapter 5.

We commenced by examining the potential for artificial agents to acquire human language through supervised learning and subsequently utilize this language to cooperatively navigate a game (§6.1). Initially, we introduced the Photobook dataset (§6.1.1) and developed two multimodal agents (a speaker and a listener), each designed to independently master a Referential Game (§6.1.3). Once the agents demonstrated adequate knowledge of the game, we paired them, investigating the feasibility of resolving the game without prior knowledge of one another (§6.1.4). Our findings confirm that this is indeed possible, even though the agents were not engineered to account for other entities in their environment and solely acted in alignment with their supervised training paradigm (§6.1.5).

Following this, we shifted our focus to explore the application of Reinforcement Learning in language model fine-tuning §6.2, with a particular interest in enabling the speaker to adapt to the listener to enhance game performance. We then considered a variant of the Referential Game, introducing knowledge asymmetry §6.2.1. In this modified scenario, while a General Speaker (G-Speak) demonstrates proficiency across all domains, the listener holds expertise limited to a specific domain (e.g., the food domain). We hypothesized that domain-specific listeners might encounter difficulties engaging in the game when the G-Speak communicated about domains outside their expertise, such as a food-specialized listener evaluating images from the vehicle domain.

To mitigate this issue, we proposed a fine-tuning strategy based on RL §6.2.2, wherein a Reinforcement Learning Fine-Tuning Speaker (RL-Speak) participate in a Referential Game with a domain-specific listener. Our methodology draws inspiration from recent advancements in language modeling (see §3.2.3). In Section 6.2.3, we initially validated our hypothesis, confirming that the ability of a domain-specific listener to engage in a Referential Game is impeded when images originate from a domain outside their expertise. Subsequently, we demonstrated how our fine-tuned speaker exhibits notable improvements in engaging the domain-specific listener in the Referential Game. However, following deeper linguistic and qualitative analysis, we identified a collapse in the speaker's word distribution to a reduced set of keywords, which prompted specific listener responses.

In conclusion, we draw attention to parallels with the language drift (§2.2.2) and misalignment (§3.2.1) issues highlighted in the related works.

6.3.1 Limitations

Although the Reinforcement Learning Fine-Tuning paradigm has emerged as a powerful strategy for modifying the speaker's parameters toward a desired direction, its application comes with notable limitations. The primary strengths of RL-FT include its capacity to explore and exploit environments and its capability to specify reward signals tailored to specific needs. Additionally, when defining rewards becomes complex, numerous methods to infer the reward signal have been proposed and are detailed in related works §3.2.1.

Nevertheless, RL-FT presents distinctive challenges. One significant obstacle is the substantial number of interactions required to fine-tune the model, which can be prohibitively extensive in the context of actual human interaction. Furthermore, a critical issue highlighted in this chapter is the ambiguity in determining the correct reward signal. When specifying the optimal outcome for a game turn, we aim for the model to generate "better" captions. This, however, implies an assumption that "better" equates to more human-like or human-friendly language usage. There is a cascading series of approximations, starting from biases inherent in the dataset, through domain-specific listeners learning approximations of these biases, to the speaker attempting to emulate these approximations. The hyperdimensional parameter space of the speaker model is too complex to be easily navigated towards our anticipated outcomes via supervised and/or Reinforcement Learning.

Necessity of Human Evaluation

A meta-analysis of our analysis sheds light on our ability to anticipate misalignment issues through various linguistic metrics. Although this approach offers a notable improvement over a sole reliance on performances, it scarcely suffices to quantify the magnitude of the misalignment. The depth of the misalignment only became transparent following a qualitative analysis of speaker behavior. This vital step is often neglected in the literature, especially within the computer science field. While fields like linguistics, cognitive psychology, and social sciences often place qualitative analysis at the forefront (if not making it the exclusive form of analysis), computer science tends to prioritize automatic metrics, occasionally overshadowing the importance of human evaluation. While the scalability and economic feasibility of human evaluations in experiments pose a challenge, they remain indispensable for infusing human-aligned values into modern AIs.

Bigger models, bigger problems?

Our ability to identify inconsistencies in the speaker model is facilitated by its relatively smaller size compared to most current Large Language Models. Analyzing the patterns of billion-parameter LLMs can be intimidating, sometimes leading to perceptions of sentient behavior [Guardian, 2022]. Such perceptions fuel arguments on machine consciousness [Chella and Manzotti, 2013], distracting from pressing AI-related societal risks. These concerns are not solely about potential malevolent Artificial General Intelligence [Pistono and Yampolskiy, 2016], but also their tangible impacts today. Though scaled down, our findings emphasize these concerns for the AI community. In the following chapter, we propose an alternative to Reinforcement Learning Fine-Tuning. Instead of modifying the model, we advocate for a static "oracle-like" approach combined with a secondary, more manageable model that utilizes the speaker to guide its generation, enhancing listener comprehension.

Chapter 7

Investigating Adaptability in AI: Focusing on the Theory of Mind

The previous chapter (§6) concentrated on enabling artificial agents to use human languages to collaboratively resolve a game using supervised learning. While successful to a degree, prominent limitations arose, particularly a lack of agency and adaptability in the agents (§6.1). To address this, we applied Reinforcement Learning as a fine-tuning strategy, aiming to allow the speaker to tailor its language to a specific listener (§6.2). Unfortunately, Reinforcement Learning Fine-Tuning led to the model's word distribution collapsing to a handful of keywords that elicited desired behaviors from the listener. Though this result might be seen as beneficial in some contexts, the resultant lack of human interpretability and the broader issue of misalignment are substantial drawbacks.

Given the conclusions drawn in the previous chapter, we observed that directly altering the parameters of the speaker (a Large Language Model) inadvertently results in the unpredictability of the learning outcome, especially when dealing with models comprising billions of parameters. Therefore, this chapter revisits the same experimental setup but adopts a different strategy wherein the speaker undergoes no further training, and the adaptation is executed by an additional, smaller, and more manageable model[1].

Our method draws from the Theory of Mind (ToM). As previously discussed in §3.3, while ToM is not unfamiliar in AI and language modeling realms, its applications have been mainly restricted to small-scale experiments due to the high costs associated with training LLMs. Thus, we aim to deliver a framework that can be integrated into LLMs without requiring exhaustive and time-intensive retraining. Our framework builds upon existing literature on machine ToM, mixing it with strategies for adapting language generation, with a goal to devise communications that are more attuned and responsive to the user's mental state.

7.1 Solution via Adaptation Techniques

This adaptation is implemented via an auxiliary model named the *Simulator*. To be discussed in further detail in §7.1.2, the simulator is tasked to anticipate the listener's behavior by engaging with it over time. Once adept at predicting, the simulator employs its understanding to guide the

[1] The content and solution proposed in this chapter has been developed in collaboration with the Dialogue Modelling Group at the University of Amsterdam.

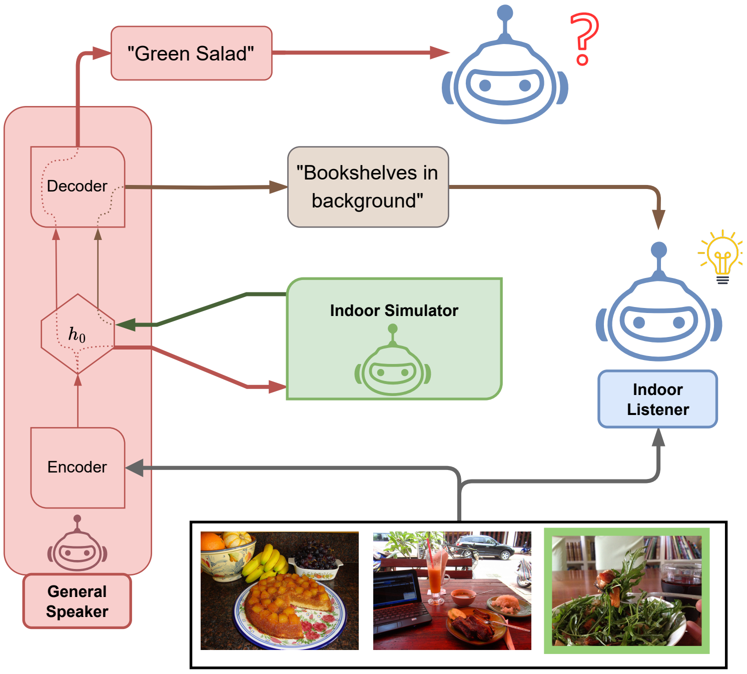


Figure 7.1: A representation of the speaker's iterative refinement in communication, transitioning from an initial caption to an adapted RefEx with the guidance of the simulator.

General Speaker 's RefEx generation. This process has an iterative nature, in line with the approach defined in §4.3.

In Figure 7.1, an example of this process is illustrated. In it, the (indoor) listener struggles to recognize the food image based solely on the initial caption "green salad". However, with the guidance of the simulator, the G-Speak adjusts its Referring Expression to better match the listener's understanding, resulting in the modified statement "bookshelves in the background". It is important to note that the simulator operates on the latent space of the speaker (h_0), connecting the encoder and decoder. This interaction facilitates the backpropagation of gradients through the simulator, guiding the G-Speak 's latent representation towards improved clarity.

The proposed solution offers several advantages. Firstly, it equips the speaker with adaptability, allowing it to adjust its RefExs to meet the specific needs of different listener domains, promoting more effective communication. Secondly, the speaker can quickly refine its language generation without the need for extensive retraining, improving efficiency. Lastly, the model maintains its domain knowledge, ensuring its applicability and effectiveness in a variety of contexts.

7.1.1 Simulator Model and Training Details

As illustrated in Figure 7.2, the simulator module represents a novel augmentation to the speaker's architecture, functioning as an internal prediction tool. Stemming from insights presented in §3.3, the simulator enables the G-Speak to anticipate a listener's interpretation of a Referring Expression, ensuring more effective communication.

The distinctiveness of the simulator is reflected in its dual-stream input processing. The first *Linguistic Stream* accepts the visual context (v_{ctx}), together with the speaker's intended RefEx (c_s). This mirrors a typical listener architecture, where a Referring Expression, once generated, is set to be resolved amidst a visual setting, thus predicting the listener's behavior. Simultaneously,

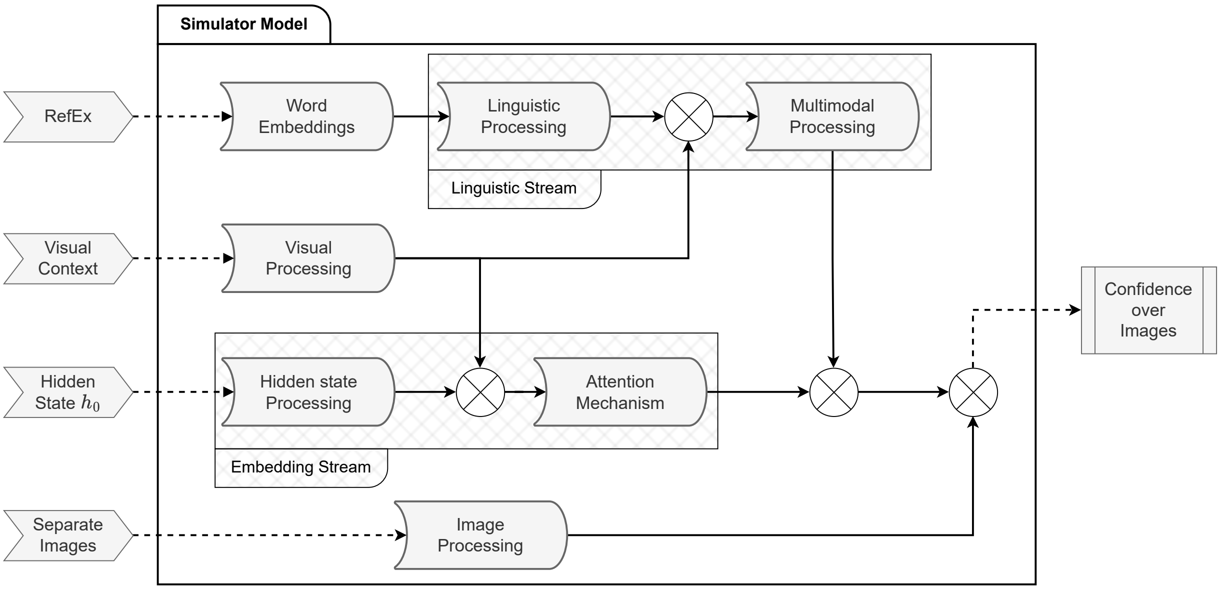


Figure 7.2: Architecture of the Simulator Module, showcasing the parallel processing streams of visual context, planned RefEx, and the language model's initial hidden state.

the second *Embedding Stream* processes the same visual context but combined with the speaker model's initial hidden state (h_0). This part is necessary to influence the speaker's RefEx generation process, as we will later see. A combination of shared linear layers standardizes and computes the dot product between h_0 and the visual context. The outcomes from both streams are multiplied together to derive the final representation that later gets contrasted against candidate images.

Formally, the simulator can be described as:

$$Simulator(v_{ctx}, c_s, h_0) \rightarrow g_{sim} : V \times B \times \Gamma \rightarrow G$$

In this definition, the simulator accepts the visual context and the speaker's RefEx, both elements of O. Additionally, it considers the initial hidden state of the speaker model, denoted as h_0, which is an element of Γ. Using these inputs, the simulator predicts actions within the predefined action set, G.

Training Methodology

In the simulator's training, we operate under the presumption that both the speaker and the domain-specific listener models have been pretrained, with their weights being frozen (see §6.1.3). This precondition is essential since the simulator's training relies on samples derived from the interactions between these two agents.

The training procedure for the simulator mirrors that of the listener in certain aspects. For instance, both models are presented with six images and must select one. Yet, a distinct difference emerges in the source of the target caption: rather than deriving it from the training set, it is generated by the speaker. Furthermore, the simulator is given an additional input, the speaker's hidden state, denoted as h_0.

Outlined below is the step-by-step progression of a training iteration:

Setting	Avg	Pos	Neg
IND	78.20 ± 1.26	88.09 ± 1.98	67.36 ± 2.96
OOD	72.78 ± 0.56	73.67 ± 1.69	72.58 ± 0.71

Table 7.1: Accuracy of the simulator in forecasting the actions of two types of listeners: one versed in *all domains* (akin to the speaker) and another with *domain-specific* expertise, evaluated on IND and OOD samples. 'Avg' denotes comprehensive accuracy, whereas 'Pos' and 'Neg' signify the proportions of accurate predictions for instances where the listener selected the right (Pos) and wrong image (Neg), respectively.

1. Six images are randomly selected from a consistent domain.[2] One among these is designated as the target ($\hat{t}$).
2. The speaker formulates a caption for this target. This generated Referring Expression and its corresponding hidden state (h_0) are retained for future use.
3. Subsequently, the listener is presented with the six images and the RefEx produced by the speaker. Based on this, the listener arrives at a prediction (g_l) that may not always align with the preselected target.
4. In the final stage, the simulator is supplied with the listener's inputs, augmented by the speaker's embeddings. The simulator's task is to anticipate the listener's decision by choosing a target image (g_{sim}).

For each domain-specific listener, a dedicated simulator is trained. The loss function employed is Cross Entropy, and optimization is achieved through the AdamW optimizer [Loshchilov and Hutter, 2019]. The optimal simulator for each listener variant, is discerned based on the precision of the simulator's predictions. Following this, the simulators' weights are set, ensuring they remain unchanged throughout subsequent stages of the process.

Training Performance Evaluation

In this evaluation, we analyze the performance of our simulators in predicting the behavior of domain-specific listeners. The findings are as follows:

- For IND samples, the simulators exhibit an average prediction accuracy of 78.20%.
- For OOD samples, the accuracy averages at 72.78%. The observed reduction in accuracy, when transitioning from IND to OOD samples, suggests potential challenges in discerning listener reactions on unfamiliar OOD data.

Further analysis of Table 7.1 reveals that the simulators demonstrate heightened proficiency in forecasting the listener's behavior when the listener accurately identifies the target image, as compared to instances where the listener erroneously selects a distractor image (designated as *Pos* and *Neg* in the table, respectively).

A plausible explanation for this phenomenon could be attributed to the consistent representation of the listener's accurate responses in the IND training data.

[2] It is pertinent to note that while the images are chosen from a singular domain, this domain alternates among all five available. This ensures the listener's response to Out Of Distribution data points is adequately examined.

7.1.2 Adaptation Mechanism

Algorithm 1: Mechanism of Model Adaptation

```
Input: $s_{iter}$ : maximum number of adaptation steps
       $lr_{adapt}$ : learning rate for adaptation
       $seed$ : random seed
Data: $h_0$ : speaker's initial hidden state
      $v_{ctx}$ : visual context
      $\hat{t}$ : true target
$i \leftarrow 0$
while $i \leq s_{iter}$ do
    $set\_seed(seed)$
    $c_s = Speaker(v_{ctx}, h_0)$
    $g_{sim} = Simulator(v_{ctx}, c_s, h_0)$
    if $g_{sim} == \hat{t}$ then
        break
    $loss = CrossEntropy(g_{sim}, \hat{t})$
    $h_0 = backprop(loss, h_0, lr_{adapt})$
    $i += 1$
$t_l = Listener(v_{ctx}, u_t)$
```

As outlined in earlier sections, the primary objective of the simulator is twofold: firstly, to predict listener behavior, and secondly, to guide the speaker's language generation towards enhanced comprehensibility. We previously detailed the predictive phase, and, in this section, we seek to define this adaptation mechanism and its elements.

A core challenge when fine-tuning language models on tasks necessitating non-communicative actions arise from backpropagation through the discrete outputs of the model, as elaborated in §3.3.3. To overcome this issue, our adaptation mechanism, as detailed in Algorithm 1, harnesses the simulator to assess the speaker's generative outcomes iteratively. The algorithmic steps can be detailed as follows:

1. **Initialization:**
   - Establish the adaptation parameters: maximum refinement steps s_{iter}, learning rate lr_{adapt}, and a seed for reproducibility.
   - Get the game input elements: visual context v_{ctx}, target image $\hat{t}$, and the speaker's initial hidden state h_0.
2. **Adaptive Refinement Loop (Maximum Iterations: s_{iter}):**
   (a) Reset the random seed for consistent results using $set\_seed(seed)$[3].
   (b) Produce an RefEx (c_s) via the *Speaker*, incorporating the visual context v_{ctx} and hidden state h_0.
   (c) Utilize the *Simulator* to predict the listener's target choice g_l given the visual context, RefEx, and hidden state.
   (d) If the simulator's projected target g_{sim} aligns with the true target $\hat{t}$, terminate the loop.

[3] This ensures that word sampling variations stem solely from changes to h_0, eliminating randomness from nucleus sampling.

	OOD			IND		
	Train Data	G-Speak	Adapted	Train Data	G-Speak	Adapted
appliances	20.21	19.30	27.74	84.21	57.21	74.28
indoor	18.50	19.53	28.34	83.22	52.94	69.62
food	17.06	18.31	26.26	85.61	55.54	78.15
outdoor	18.89	18.54	26.21	84.38	52.83	73.04
vehicles	18.25	17.35	25.16	78.67	42.09	63.75
Average	18.58	18.61	**26.74**	**83.22**	52.12	71.77

Table 7.2: Comparison of Listener Performance in the Referential Task Using Utterances from Different Sources: Original Training Set (Train Data), Unmodified General Speaker (G-Speak), and Adapted Speaker-generated (Adapted) across various domains. Results are aggregated over five seeds for each domain.

(e) Otherwise, calculate the CE loss between the simulator's estimation and the true target.

(f) Modify the hidden state h_0 by backpropagating this loss, employing the specified learning rate lr_{adapt}.

3. **Listener's Prediction:**

   - Following the refinement loop, the listener predicts the target choice g_l utilizing the visual context and the refined Referring Expression.

From this process, it is evident that when a mismatch surfaces between the simulator's estimate g_{sim} and the actual target image $\hat{t}$, a Cross Entropy loss is generated. Gradients derived from this loss are harnessed to adjust h_0 using the Adam optimizer. Fundamentally, the adaptation fine-tunes solely the initial hidden state of the speaker's decoder. Upon updating this state, the language model formulates a fresh RefEx, which then undergoes assessment by the simulator.

7.2 Evaluation of Machine Adaptation Performance

In the preceding section, we discussed the key components of our system and provided a foundational overview of the adaptation mechanism, where the simulator modulates the speaker's latent space to enhance the listener's comprehension.

This section is dedicated to presenting the outcomes of the adaptation mechanism. Specifically, we will first examine performance outcomes, assessing the listener's efficiency on the referential task after receiving the adapted Referring Expression in §7.2.1. Subsequently, we will explore the effect of ToM adaptation on the speaker language use, in §7.2.2.

7.2.1 Analyzing Performance Metrics

Table 7.2 displays the listener's performance in the referential task across different domains. The results are categorized into OOD and IND. Each category further lists the outcomes based on the different origins of the RefEx: the original training set (*Training Data*), RefExs generated by the speaker (*G-Speak*), and adapted RefExs (*Adapted*).

The data suggests that adaptation proves beneficial in scenarios with knowledge asymmetry. This is evident even in IND contexts where the agents engage in a conversation about a domain familiar to the listener, with performance improving from 52.12% to 71.77%. Furthermore, the adapted RefExs significantly enhances resolution in OOD scenarios, increasing from 18.61% to 26.74%.

The results presented directly address the issues introduced at the beginning of this chapter and demonstrate the efficacy of our solution. By integrating a ToM-enhanced architecture into language generation, we can modify the generated language to enhance comprehension. While this achievement is noteworthy and constitutes a milestone in our study, an exclusive emphasis on performance enhancement does not provide insights into the qualities of the language, as highlighted in §5.2.2. Such a narrow focus can lead to erroneous conclusions and may degrade human-machine interactions in the long run. To address this, the subsequent section offers a comprehensive linguistic analysis of the adapted RefExs.

7.2.2 Linguistic analysis

The primary aim of this section is to delve deeper into the neural mechanisms and strategies that derive from our adaptation experiments.

Probing Neural Representations for Domain Information

Diagnostic probing is a technique often employed to investigate and interpret neural network representations. According to Adi et al. [2017], Conneau et al. [2018], Hupkes et al. [2018], this method allows researchers to discern what information is stored or retrieved from neural activations. With diagnostic probing, we want to decipher how well the speaker model encodes relevant domain information.

Central to our investigation is the LSTM decoder's first hidden state, h_0. This state is crucial as the simulator module acts upon it to modify the speaker's Referring Expression and can be seen as the speaker's belief state. Given its role in encoding a target image, our hypothesis suggests that h_0 should possess information about the semantic domain of that image. For the speaker to adapt successfully to domain-specific listeners, its ability to differentiate between visual domains becomes necessary.

To validate our hypothesis, we subjected h_0 to the diagnostic probing process. A logistic regression classifier was trained on 70% of the h_0 hidden states obtained during testing. We then evaluated its proficiency in predicting the domain of images tied to the remaining 30%. Consistent with our anticipations, the classifier demonstrated perfect precision and recall across all five visual domains (both equal to 100%). Our next step revolved around whether h_0 also captures the domain of the *listener*. Preliminary understanding suggests that prior to any simulator intervention, the speaker model remains oblivious to the listener's domain knowledge. To test this, probing was conducted, revealing accuracy scores between 13% and 16% across various domains, nearly coinciding with the random baseline of 16%. This affirmed that pre-adaptation, the speaker's initial hidden state remains uninformed about listener-specific data.

As adaptation progresses, led by the simulator's processes, h_0 undergoes modifications over adaptation steps. Analyzing these transformed states: $h_0^1, h_0^2, \ldots, h_0^{s_{iter}}$, we notice a decline in the encoding robustness of the image domain (Figure 7.3). Indeed, after three adaptation steps, the listener's domain can be almost perfectly recognized from the adapted h_0, with an accuracy of 90%.

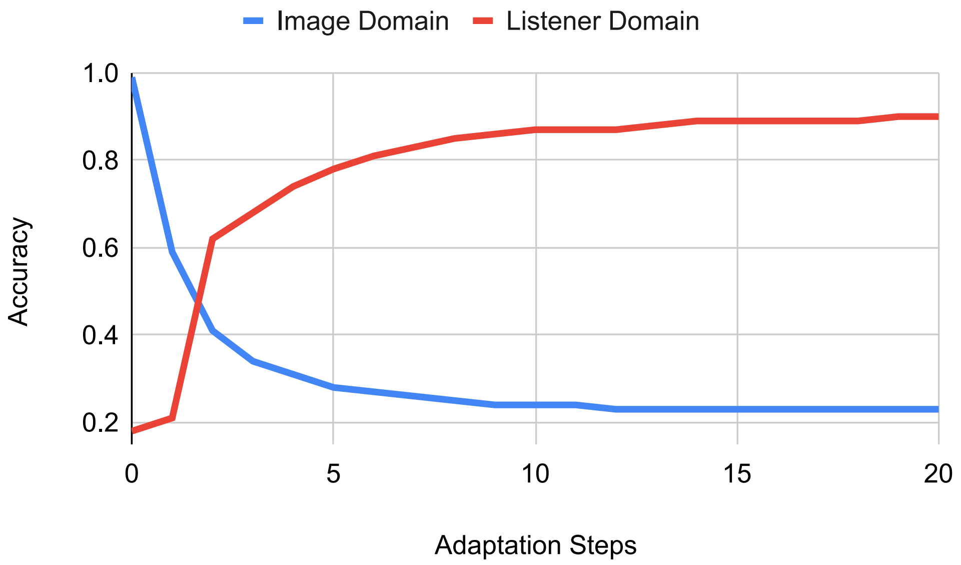
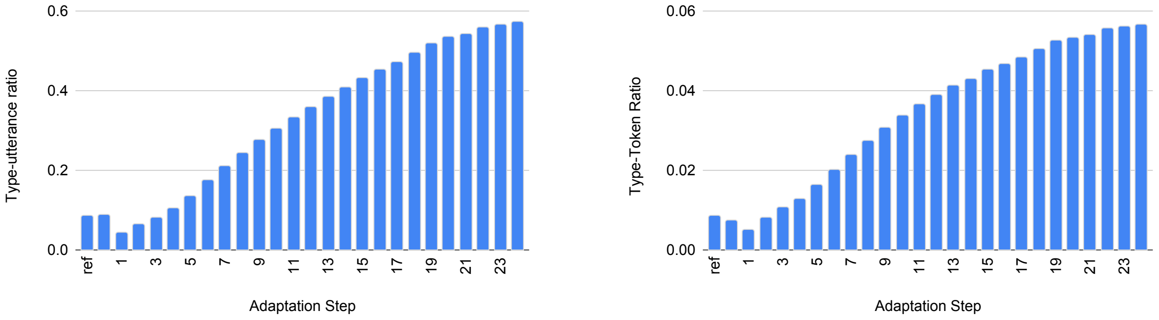


Figure 7.3: Graphical representation of probing accuracy for image and listener domain predictions across various adaptation phases. Stage '0' pertains to the unadapted h_0.



(a) Type-Utterance Ratio (TUR)

(b) Type-Token Ratio

Figure 7.4: (a) Variation in Type-Utterance Ratio (TUR) across adaptation steps, including comparisons with human gold utterances (*refg*) and non-adapted utterances (0). (b) Fluctuations in Type-Token Ratio during adaptation steps, with similar comparisons.

These findings show how listener-focused data effectively replace information from the visual context through adaptation.

Analysis of the Adapted Speaker Vocabulary

To understand the dynamic shift in vocabulary used by the speaker during its adaptation, we calculated two linguistic metrics: the Type-Utterance Ratio (TUR) and the Type-Token Ratio (TTR).

Type-Utterance and Type-Token Ratios The TUR measures the vocabulary size relative to the number of RefExs for a specific step, giving insights into the density of vocabulary usage[4]. Conversely, the TTR provides insight into the diversity and variability of the vocabulary utilized.

[4] It is important to clarify that this ratio accounts for the variable number of RefExs across steps, given the nature of the simulator module's operations.

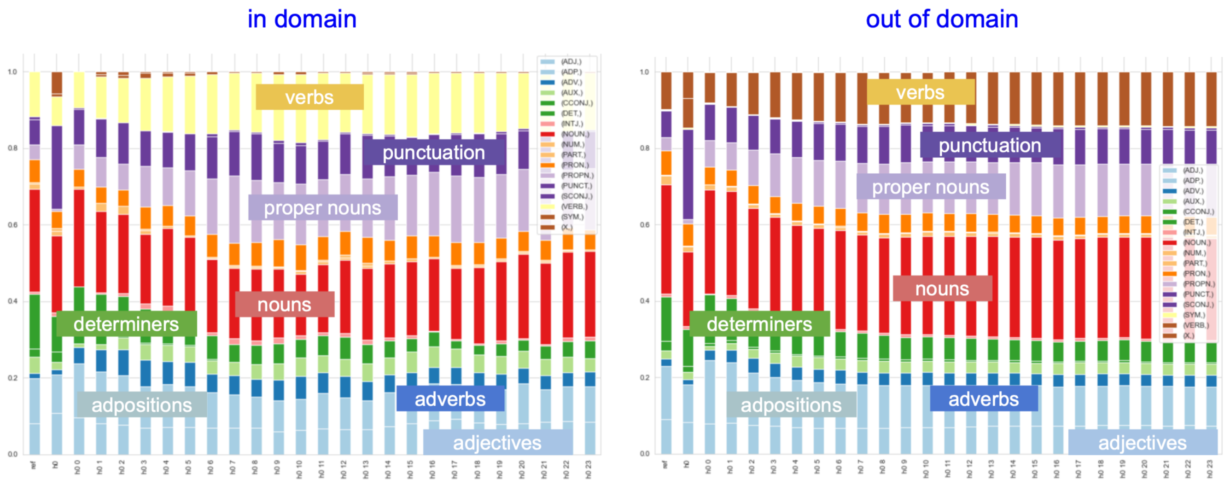


Figure 7.5: Distribution of unigram Part of Speech categories across various adaptation steps.

An examination of figure 7.4 reveals how the initial adaptation phases, particularly steps 1 to 3, show a noticeable decline in both ratios. However, following this dip, a significant increase can be observed. This indicates a considerable augmentation in vocabulary diversity, especially when contrasted with the non-adapted RefExs.

Unigram Part-of-Speech Distribution Analyzing the unigram Part of Speech (PoS) distribution, as shown in Figure 7.5, gives insights into the speaker's language choices during adaptation. In the early stages of adaptation, there is a clear decrease in punctuation and a rise in nouns. This indicates the speaker's shift from sentence structure to emphasizing specific content using nouns. Among the diverse PoS categories, two exhibited pronounced changes: proper nouns and determiners. This rise can be interpreted as the speaker's effort to make its language more specific, using unique identifiers that might better align with the listener's domain-specific knowledge. This approach mirrors what we observed when employing Reinforcement Learning Fine-Tuning in the previous chapter, where specific word choices helped tailor the communication more effectively. On the other hand, the use of determiners decreases, indicating a move towards more generalized or open-ended statements by the speaker. This trend suggests a strategic avoidance of generating overly specific referential expressions, a challenge we previously encountered with RL-FT. The adaptation process thus seems to reach a balance, avoiding overly specific language while still achieving more tailored communication.

Domain-Specificity of Referring Expressions In our study, we also examined how the specificity of RefExs changed over time with respect to both the image content and the listener's domain. To achieve this, we identified words that were entirely used in conversations about a specific domain, labeling these as domain-specific. A key finding from this analysis is that, as the adaptation process unfolded, the speaker began to use more words relevant to both the image in question and the domain of the listener (as shown in Figure 7.6). Simultaneously, there was a noticeable decrease in the use of more general, domain-agnostic terms.

Interestingly, even though the speaker's focus shifted more towards the listener's domain over time, this shift did not result in completely ignoring the image domain. Instead, the speaker

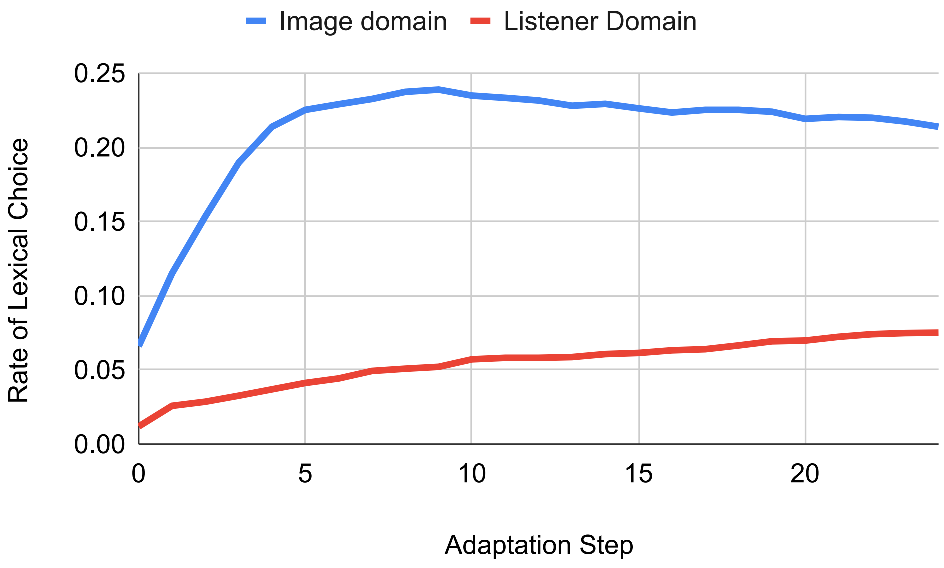


Figure 7.6: Distribution of Lexical Selections from Image and Listener-Specific Domains.

continued to incorporate words related to the individual images. This trend indicates that the speaker might be giving more importance to the specific images rather than their broader semantic categories. This observation raises further questions about how the speaker is grounding its language in the image context and how it uses image-related words during adaptation.

Successful Adaptation Strategies

An important aspect of our exploration is discerning the differences between successful and non-successful adaptation outcomes. While our previous observations provided an overview of how adaptation impacts RefExs over time, they did not explicitly address the characteristics of successful adaptations.

To explore this, we considered the concept of Age of Acquisition (AoA), a metric used in psycholinguistics to indicate when a word is typically learned in one's life. Based on the ratings by Kuperman et al. [2012], AoA values range from 0 to 25, with lower values representing words learned earlier in life. These early-learned words are usually more basic and widely understood. We hypothesized that RefExs utilizing words with a lower AoA might be more effective as they are likely easier for the listener to comprehend.

Our hypothesis is supported by the findings illustrated in Figure 7.7. We analyzed adapted RefExs based on whether they led to correct responses from the listener. The analysis revealed that successful RefExs tended to use words with a lower AoA (with a statistically significant result of $t = -28.88$, $p < 0.001$). Additionally, these successful RefExs showed a decreased reliance on vocabulary specific to the target image (with a significant $t = -28.76$, $p < 0.001$) and an increased use of terms from the listener's domain ($t = 5.88$, $p < 0.001$). This suggests that adaptability in RefExs is not just about adjusting to the listener's domain but also about using more universally understandable language.

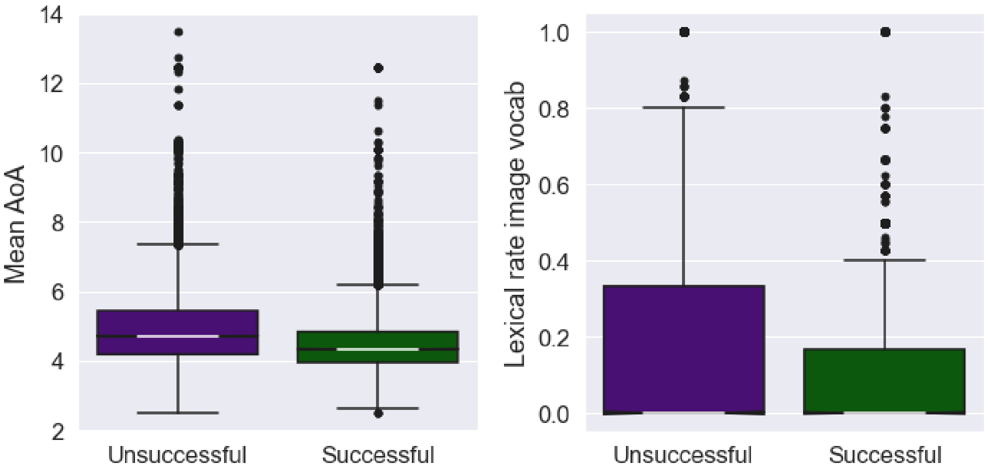


Figure 7.7: Influential determinants of a successful adapted RefEx: Age of Acquisition (left) and proportion of words tied to the target image domain (right).

Qualitative Assessment Through Manual Inspection

For a thorough understanding of the model's behavior and its linguistic adaptation mechanisms, a manual examination of the generated data remains crucial. This step ensures we are not relying solely on metrics but are also considering the intricacies of the model's language dynamics.

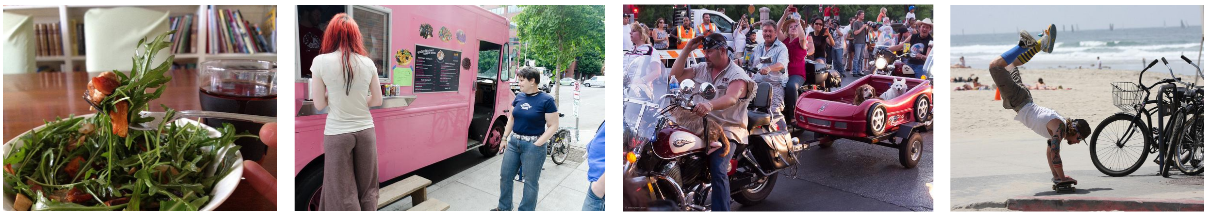
In Figure 7.8, we illustrate select instances of adapted sentences to show the influence of adaptation on lexical choices. Taking the leftmost example as a case in point, the image domain presented is primarily centered around *food*, and the associated listener had its training grounded in the *indoor* domain. As a result, the adapted RefEx shifts from a description of food items and introduces a term - *bookshelves* - more familiar to the listener's domain. The second example features an image from the *outdoor* domain and a listener trained in the *food* domain. Notably, the adapted Referring Expressions avoids explicitly mentioning a *truck*. Instead, the model emphasizes recognizable features like the color *pink* and leans towards entities familiar to the listener, such as *donuts*.

However, our qualitative evaluation does raise some issues. While many of the adapted RefExs are understandable, a notable portion appears less fluent or even unnatural. This may stem from our choice to use artificial agents. As a result, the language model's adaptation to an artificial listener might deviate from typical human language patterns.

7.3 Discussion and Key Findings

In our pursuit of enhanced Human-Machine Communication, this chapter delves into the often-overlooked aspect of human-to-human interaction in modern Large Language Model: adaptation. We have delineated adaptation and associated it with the concept of Theory of Mind, defined as the ability to have a mental model of other agents in the environment and tailor your action accordingly. Given that this theory originates from cognitive sciences (see §3.3.1), it naturally steers our exploration towards cognitive insights within computational linguistics, where we identified Rational Speech Act 's alignment with ToM and state-of-the-art adaptive techniques for LLM (§3.3.2).

Drawing parallels with challenges addressed in the previous chapter, we customized the formal-



food to indoor
Gold green salad with a person holding up a portion with fork ?
Speaker i have one more maybe round you think that has a lime green shaped greens , a salad ?
Adapted 1 with carrots , etc . maybe you have a picture of a salad on a reddish tray of an greens
Adapted 2 must bookshelves in the salad ?

outdoor to food
Gold i have the pink food truck again ... white shirt lady
Speaker girl at black phone , red truck , brown hair , pink
Adapted 1 two donuts , tan tank top ?
Adapted 2 pink donuts

vehicles to food
Gold do you have the bike pulling the car with the dogs ?
Speaker i have that one too
Adapted 1 with a cup of dogs on ?

outdoor to vehicles
Gold handstand on beach .
Speaker i have the guy with his hand
Adapted 1 low dude doing handstand
Adapted 2 must beach doing
Adapted 3 must beach doing
Adapted 4 low tire doing handstand lenningh .

Figure 7.8: Illustrative examples of RefExs influenced by audience-aware adaptation. We simplify the presentation by displaying only target images, omitting the comprehensive visual context. Presented are the final adapted RefExs, generated when the adaptation mechanism anticipates successful image recognition by the listener.

ism presented in §4.1 to delineate our problem's constituents, namely the *simulator* (§7.1). Notably, while the simulator is a crucial component of our approach, the essence of our proposition lies in the adaptation algorithm detailed in §7.1.2.

At the conclusion of our experiment, we shifted our attention to the study's outcomes as outlined in §7.2. Mirroring the analytical approach of the prior Chapter 6.2.3, we first illustrated the enhanced performance of the listener when utilizing adapted expressions, as delineated in §7.2.1. The empirical results met our anticipations, illustrating augmented performance in both IN-Domain instances (with an improvement of $\approx$ 20%) and Out Of Distribution instances, recording an improvement of $\approx$ 7%.

While these results are exciting, we emphasized that merely optimizing for performance without interpretability can lead to languages that are not transparent to human observers. With this understanding, we performed a linguistic analysis presented in §7.2.2. Our findings are as follows:

(i) Diagnostic probing revealed that the speaker model's hidden state, denoted as h_0, begins by accurately encoding the image's domain. Yet, as adaptation unfolds, there's a shift towards encoding the listener's domain knowledge. This shift ends in perfect listener domain recognition after three adaptation steps. (ii) The study's linguistic metrics revealed that the speaker's vocabulary initially decreased in diversity during early adaptation phases, but subsequently showed a significant increase in vocabulary diversity, surpassing non-adapted RefExs. (iii) The unigram part-of-speech distribution analysis revealed that during adaptation, the speaker reduces punctuation use and emphasizes content through an increase in nouns, especially proper nouns, while decreasing determiners to create more generalized statements. (iv) As adaptation progressed, the speaker increasingly incorporated words related to both image and listener domains, reducing domain-agnostic terms, indicating a focus on individual images rather than their broader semantic domain. (v) Successfully adapted RefExs predominantly use words learned earlier in life (lower Age of Acquisition) and favor the listener's vocabulary over the target image's specific vocabulary.

7.3.1 Comparative Analysis: Theory of Mind vs. Reinforcement Learning Techniques

Comparing results from the previous chapter (§6.2.3) and this chapter (§7.2.1), a clear pattern can be seen: Reinforcement Learning Fine-Tuning outperforms Theory of Mind-based modeling concerning measured performances. Notably, RL-FT demonstrates superior performance due to its ability to harness the entirety of a LLM complexity and size, focusing on excelling at specific tasks.

However, this comes with anticipated drawbacks. Firstly, fine-tuning an LLM is computationally expensive, both in terms of the power needed and the volume of data required to approximate optimal behavior accurately. This is attributed to the model's size, as larger models inherently possess more substantial priors. In practice, given equal training samples, our simulator outperforms a pre-trained speaker by better approximating a listener's behavior. In this direction, we advocate for the adoption of smaller, modular systems comprised of multiple task-specific models instead of foundational models with billions of parameters. This not only facilitates better manageability but also ensures that problems are decomposed into smaller, more manageable units.

An additional limitation, centrally featured in this work, pivots on the concept of misalignment intrinsic to Reinforcement Learning Fine-Tuning. Simulators, while not a definitive solution, serve as an initial step towards modeling decisions based on intentions, with the intention in this context being to predict listener behavior. This approach aligns with efforts by researchers focusing on action-oriented language models [FAIR]. Such models, comprising multi-part systems, are driven by intention and can be probed to comprehend the genesis of a particular sentence generation. In essence, they embody an understanding of how their generated words manipulate the environment and other interacting agents.

7.3.2 Limitations

Having discussed the advantages of the Theory of Mind-based solution method over the Reinforcement Learning Fine-Tuning, we must now address its limitations.

One primary limitation is the focus on adapting the speaker alone, without considering mutual adaptation. In human communication, both speaker and listener typically adjust to each other, a dynamic we have not fully captured. This might mean missing out on the complexities and richness of interactions that involve both parties modifying their behavior for clearer communication.

Another limitation stems from our choice not to use pretrained models for language tasks. While this approach was deliberate in addressing specific research questions, it does raise concerns about the applicability of our findings to scenarios where pretrained models, with their rich linguistic knowledge, are employed. Additionally, our model architecture does not incorporate transformer-based models, which have demonstrated their efficacy in various language tasks.

Furthermore, despite interaction being a central theme, our research heavily relies on supervised training paradigms. Incorporating Reinforcement Learning could potentially provide more comprehensive insights into the adaptation process, presenting a direction for future exploration.

However, the most significant limitation is our exclusive use of artificial learners. These agents, designed to mimic human responses, may not fully replicate human belief systems. This raises the risk of misalignment, where behaviors influenced by feedback from artificial agents do not align with human values and expectations. This issue is especially critical given our simplification of equating an agent's belief state to its action (2-$\mathcal{GSF}$). While this makes modeling more manageable, it overlooks the complexity of human beliefs. Addressing this alignment challenge is crucial and highlights the irreplaceable role of human involvement in shaping AI behaviors.

Chapter 8

Conclusions

At the core of this work is the question of how Large Language Models can autonomously act while aligning with our desired outcomes and the broader societal context. The pervasive influence of such technology in society naturally raises concerns regarding its reliability. A crucial question arises: *Do these models genuinely understand their societal repercussions, and can they consistently align with human values?* These questions are rooted in the concepts of agency and misalignment, which form the backbone of our discussions.

Driven by the historical evolution of AI, aligning it with human values has become fundamental, with the past decade emphasizing a human-centric approach. This approach, described by Davenport and Kirby [2016] as augmentation, seeks to enhance rather than replace human efforts. As language is a universal human experience, there's been a shift towards natural language-based applications. This shift amplifies the significance of Human-Machine Communication, where the challenge lies in training artificial agents, especially with the emergence of the transformer architecture in NLP, to communicate in globally spoken languages.

The year 2023 has witnessed a surge in the deployment of LLMs across diverse fields, including medicine, chemistry, and economics [Clusmann et al., 2023, Sallam, 2023]. However, one limitation of LLMs is their reliance on predicting the next word in a sentence rather than understanding the context and purpose behind language usage, as humans do. Given the widespread impact of this technology on society, issues emerge about its reliability.

As researchers, we must address these issues by enhancing LLMs with the awareness of how their communications influence their surroundings and facilitate a deeper comprehension of humans. This multidisciplinary problem intersects various fields, but our primary focus is on Human-Machine Communication.

8.1 Comprehensive Summary

Having identified the lack of *agency* and *misalignment* in modern LLMs as main issues, this doctoral thesis systematically addresses these challenges. The approach is structured into three distinct phases, each building upon the findings of the previous one. Each of these phases is grounded in peer-reviewed materials that have been published during this doctoral journey. Specifically, phase one is informed by *Rlupus: Cooperation through emergent communication in the werewolf social deduction game* [Brandizzi et al., 2021] and *Emergent communication in human-machine games* [Brandizzi and Iocchi, 2022], phase 2 draws from both *Speaking the language of your listener: Audience-aware adaptation via plug-and-play theory of mind* [Brandizzi et al., 2023] and *Towards more human-like AI communication: A review of emergent communication research* [Brandizzi, 2023], and phase 3 is based on Brandizzi et al. [2023].

8.1.1 Game Dynamics for Exploring AI Agency

In the initial phase, we deal with a fundamental question: *Can agency manifest in artificial systems given an adequate learning framework?* We address this issue in Chapter 5 by creating an artificial version of *The Werewolf* game, where agents coordinate within their groups and compete against other groups. In this game, one group (the villagers) is posed in a heavily disadvantageous position, where chances of winning against the other group (the werewolf) in a random environment are close to zero (4%). We posit that: *if agents can leverage a communication channel for cooperative intent, it serves as evidence of their emergent agency*, i.e., using communication to influence the game outcome.

To assess our hypothesis, we introduce a communication channel in the game without incentivizing agents to use it. Upon analyzing the results, we found that the villagers' win rates increased significantly from 4% to 40% with the introduction of communication, thereby supporting our hypothesis. Our results show that by introducing a simple communication channel, even without explicitly teaching the agents to use it, we witnessed a tenfold jump in the villagers' chances of winning. This transformation signifies that when placed in the right learning environment, like our multi-agent and Reinforcement Learning setup, artificial agents can indeed develop the ability to communicate with intent, demonstrating emerging agency in linguistic contexts. We validate these findings across two player configurations: nine and 21 players. Subsequent analysis by Lipinski et al. [2022], which builds on our findings, reveals how the villagers agreed on specific communicative symbols, setting a kind of "linguistic trap" for the werewolves. Since the werewolves did not learn this emergent language, they were easily identified by their inability to use or understand these symbols, reminiscent of how a Turing test distinguishes between humans and machines. Yet, a challenge remains: the evolved language, while efficient, consists of bits and numbers, making it non-interpretable by humans. Our exploration ends with the insight that while the right learning paradigm can indeed spark agency, the resultant language must be understandable to humans.

8.1.2 Language Agency in AI through Reinforcement Learning

Building upon our earlier findings, the subsequent Chapter 6 advances to agents that communicate in a human-interpretable manner. Drawing from the established research (§3.2), we experiment with a Referential Game (Ref Game). In this game, one agent (the *speaker*) describes a target image while another agent (the *listener*) tries to identify it from a set. We aim to answer the following question: *Can agents, trained independently from each other, collaboratively solve a game?* Surprisingly, our results revealed that these independently-trained agents could solve the game with above random accuracy (§6.1). We attribute these results to their shared expertise in both linguistic and visual domains. However, this success leans heavily on the agents sharing complete knowledge, an idealistic scenario not reflected in real-world dynamics, where knowledge disparities are common.

We introduce a different version of the Ref Game to simulate these disparities. Here, the game's data is split across various domains, laying the groundwork for an Asymmetric Referential Game (A-Ref Game). With the speaker retaining a complete knowledge base, we build listeners limited to specific domains, e.g., listeners fluent in food-related images only. As anticipated, this setup severely impairs game outcomes, with some performances decreasing to random chance when a listener confronts unfamiliar domains, for example, vehicle images. The primary challenge arises from the speaker's limited exposure to other agents during training, suggesting a lack of agency.

Taking inspiration from the previous part, we turn to Reinforcement Learning, fine-tuning the speaker to adapt to a designated listener (§6.2). The outcomes show the adapted speaker outperforming its original version (44.57% vs. 23.85%), again proving how an adequate learning framework can mitigate the issue of agency. Yet, a deeper analysis into the speaker's lexicon uncovers a reliance on selected keywords that trigger specific listener reactions (§6.2.3). We relate this trend to the misalignment problem since higher game performances do not inherently translate to better communication. Finally, we critique the constraints of our approach, particularly in the context of contemporary LLMs training (§6.3). While we can address the agency issue by adjusting training methods, the misalignment problem requires a different solution.

8.1.3 Advancing AI Communication with Theory of Mind

Our concluding Chapter 7 stands as the main contribution of this thesis, combining insights from different disciplines to address the problem of agency and misalignment in language models. We propose a novel approach inspired by the recent introduction of adapters to the field of NLP and insight from cognitive psychology.

Adapters are trainable functions used in LLMs that leverage the model's prior knowledge and adapt it for a different task, e.g., a LLM used for translating English to Italian can use an adaptor for translating from Italian to English (§3.3). However, adapters do not necessarily capture nor model the new task they are trained for. On the other hand, a prominent theory in cognitive psychology, termed Theory of Mind (ToM), defines a propriety of humans strictly tied with modeling. ToM is the ability of humans to reason about the understanding and knowledge of other humans and modify their actions accordingly (§3.3.1). In this sense, ToM provides a perfect framework for alignment since, if the machine is able to understand what the human wants from it, it will more likely be able to accomplish exactly that. Moreover, ToM also tackles the agency issue, where reasoning about the effect an action can have on another agent is one of the fundamental aspects discussed previously.

Our solution takes into account both strategies, with the introduction of an auxiliary model: the *simulator* (§7.1). The simulator's primary function is to learn the listener's behavior during interaction rounds in a typical ToM fashion. Once the simulator demonstrates proficiency in predicting listener actions based on the environment, we implement a unique adaptation mechanism inspired by adapters. This mechanism leverages the simulator's insights to guide the speaker's responses, ensuring they resonate more effectively with the listener.

To validate our approach, we initially show (in §7.2) how our adaptation strategy enables the speaker to adjust to various listeners, improving performances (49.25% vs. 23.85%). Subsequently, we examine the evolved language and report metrics such as Type-Token Ratio, Part of Speech tagging, and Age of Acquisition. Our findings indicate a noticeable improvement in vocabulary diversity (avoiding the previous issue of distribution collapse) and enhanced human interpretability.

One significant advantage of our approach is that we do not need to make extensive changes to the speaking model, which is both time-consuming and has a significant environmental impact [Rillig et al., 2023, Scao et al., 2022]. Instead, our smaller *simulator* model is more environmentally friendly and quicker to train. Furthermore, the reduced model size is advantageous for interpretability analyses, which are vital to gain insight into how the AI makes decisions. Another critical point to note is that we are addressing both the agency and misalignment issues together. Often, solutions focus on one problem or the other, but our approach tackles both. Finally, combining NLP techniques, psychological insights, and environmental considerations highlights the importance of combining knowledge from different fields. As AI plays a more significant role in our daily lives, it is clear that we need expertise from various areas to ensure it works responsibly and effectively for everyone.

8.2 Limitations

This thesis ambitiously aimed to address critical challenges in the field of AI, specifically in enhancing the agency and alignment of language models within human contexts. While the findings contribute valuable insights to the discipline, it is crucial to acknowledge certain limitations that influenced the research scope and outcomes. This acknowledgment does not diminish the work's significance but rather provides clarity on the experimental conditions and the potential scalability of the proposed solutions.

8.2.1 Resource Constraints and Model Selection

One significant limitation was the computational resources required for employing transformer models, which are known for their state-of-the-art results in various NLP tasks. The selection of LSTM models was critical to ensure the feasibility of experiments within the available resources, and they have been effective to an extent within this framework. Although the results can theoretically be applicable to transformer architectures, further experimentation is needed to confirm their scalability and to enhance the robustness of the solutions proposed in this thesis.

Moreover, opting for LSTM models also introduced challenges in the quality of language generation. While LSTMs are resource efficient, they fall short of the richer linguistic and grammatical precision offered by the more advanced transformer models. This limitation was evident in the AI's language outputs, which, although functional, occasionally lacked the depth of context that transformers can achieve. This issue highlights a prevalent challenge in today's NLP research landscape: the necessity for substantial computational resources to achieve impactful results. This requirement often limits what can be accomplished in academic settings and tends to favor entities in the industry with greater resources. We address the implications of this disparity in the Ethical Considerations chapter of this thesis (§8.4.3), underscoring the need for equitable access to advanced computational technologies in the academic community.

8.2.2 Exclusion of Human Participants

Another limitation was the exclusion of human participants in the experimental phases. Integrating human data could provide deeper insights into the human-AI interaction dynamics and enhance the validity of the AI's agency and alignment in real-world scenarios. The use of simulated environments and agent-only interactions was primarily due to the high costs and logistical complexities associated with human-subject research. Despite this, the conceptual frameworks developed are designed with the flexibility to include humans in the loop. Future studies are encouraged to incorporate human participants to test the applicability of the findings in more naturally occurring interaction settings. Furthermore, the ethical implications of involving human participants in AI research are significant and are discussed in greater detail in the Ethical Considerations Section 8.4.3.

8.3 Future Research Impact

In this section, we reflect on the broader implications of our research. Here, we consider how the insights and methodologies developed in this thesis could shape future AI research, influence industry practices, and contribute to society's understanding of AI. This foresight effort is an essential part of our academic responsibility, helping us to understand and articulate the potential future influence and real-world applications of our findings.

8.3.1 Theoretical : Merging RL and Language Modeling

In this thesis, we have taken a fresh approach by combining Reinforcement Learning and language modeling, as highlighted in Section 4.2. Our approach redefines the standard perspective of supervised training in Large Language Models, where it is not considered an isolated process but a policy within a dual-policy architecture. By viewing supervised training as a policy in its own right, we effectively merge it RL. This integration allows for the application of RL principles and methodologies directly to the process of language model training. Thus, advancements and innovations in supervised learning for Large Language Models can be directly incorporated into this broader RL-based framework. This novel perspective significantly expands the scope and applicability of RL strategies in language modeling. It opens a pathway for a more cohesive and integrated approach to AI development, leading to potentially more robust, adaptable, and advanced language models.

Moreover, another key theoretical contribution of our work is the iterative refinement process, as detailed in §4.3. Traditional alignment methods in modern LLMs often rely on retraining with RL or new supervised tasks. These practices directly modify the model parameters, necessitating substantial computational resources. Our methodology, which delegates the adaptation task to an external model, addresses these challenges. By reducing the computational load on the primary LLM, we could potentially replace the current resource-intensive processes, marking a significant theoretical advancement in AI and language modeling.

8.3.2 Industrial: Personalizing AI with Efficient Models

Large Language Models have gained immense popularity, revolutionizing various industry applications. However, a significant limitation is their lack of personalization, i.e., the ability to adapt to individual users' specific needs or preferences. Until now, the industry has attempted to achieve this through fine-tuning techniques, such as Reinforcement Learning or adding supervised tasks. However, due to the immense computational resources required, this approach is not scalable to individual users.

Our research presents a novel solution, particularly the simulator model introduced in Section 4.3. By utilizing smaller, more efficient models, we facilitate the personalization of LLMs with considerably reduced computational demands. This model acts as an intermediary, adapting the LLM's responses to align with individual user preferences, enabling a more personalized interaction without the need for extensive resources typically associated with fine-tuning LLMs.

Advancing Towards On-Device Language Models

Another emerging focus in the industry is developing compact LLMs capable of operating on user devices. Current efforts in this direction often involve downsizing the models, which, unfortunately, compromises their capabilities. Our approach offers an alternative pathway.

By offloading the personalization aspect to our smaller simulator model, the need for constant online connectivity and server-side computation diminishes. This model can be trained directly on user devices, allowing the original, more robust LLMs to retain their full capabilities while providing personalized experiences. Although internet connectivity remains a requirement for initial interactions, our methodology lays the groundwork for more independent, on-device LLM in the future. This advancement not only enhances user privacy and accessibility but also marks a significant step towards more sustainable and efficient AI technologies in everyday applications.

8.3.3 Societal: Promoting Sustainable and Ethical AI Research

In discussing the societal impact of our research, it's essential to recognize the broader implications of AI technologies on the environment and accessibility. A key aspect of our work is the reduced computational resource requirement for adaptable language models. This lessens the environmental impact, which is crucial in an era of heightened ecological awareness, and democratizes access to advanced AI technologies. With our approach, requiring fewer resources, a wider range of individuals and groups, from academic researchers and students to tech enthusiasts, can experiment with and train these models on their own devices. This democratization aligns with the goal of making AI more inclusive and accessible, breaking down barriers that previously limited engagement with these technologies to entities with substantial computational capacities.

Addressing AI Misalignment and Encouraging Responsible Research

The second part of our societal impact concerns the ethical dimensions of AI, particularly the issues of misalignment and agency. Our work highlights the real-life consequences of unmonitored AI systems, showcasing instances where such technologies have negatively impacted human living conditions. By bringing these issues to the forefront, this thesis emphasizes the necessity of considering AI's societal impacts throughout the research process.

Throughout this work, we maintain a dual focus on research advancement and societal implications, underlining the importance of AI technologies aligning with human values and societal norms. We hope that this approach encourages other researchers to similarly prioritize the societal impact of their work, fostering a research culture that not only pursues innovation but also conscientiously evaluates how such advancements affect the broader society.

8.4 Ethical Considerations

While this thesis has extensively focused on the technical development of AI, it is important to address the ethical implications of these advancements. Integrating ethical considerations into AI education is essential, either by incorporating ethics as a fundamental component of university curricula or by promoting collaborations with experts in the field. This section aims to explore some ethical considerations, underscoring the need for balanced progress in AI that is aware of its broader impacts on society and human values.

8.4.1 Humanazing AI: When Do Machines Become Too Human?

Throughout this thesis, our primary aim was to develop machines that better understand us through learning dynamics akin to human experiences. This objective raises an important ethical question: *how much should we blur the line between humans and machines?*

There are two principal approaches to this dilemma. One approach is maintaining a clear distinction between machines and humans, ensuring these entities are consistently recognized as distinct. However, this may impose a mental workload on users, requiring them to adapt to interacting with these machines. Conversely, creating machines increasingly resembling humans could ease this cognitive strain and enhance Human-Machine Interaction. Yet, a pressing ethical concern arises: *what are the implications when machines become indistinguishable from humans?*

This question must be considered from both human and machine perspectives. From a human standpoint, we must ponder the ethics of potentially deceiving a person into believing they are interacting with another human. What could be the societal impacts of such interactions? The 2013 film "Her" by Spike Jonze illustrates this dilemma, where the protagonist falls in love with an AI entity. This scenario resembles an incident involving a Google engineer who believed an advanced language model exhibited sentience [Guardian, 2022]. In these situations, *is it morally acceptable to subject individuals to emotional connections with entities they can neither physically interact with nor share a human experience?* This concern intensifies where very lonely individuals find company and potentially life-saving connections through their interactions with AI systems. From the machine's perspective, ethical questions about their treatment arise as they evolve to exhibit human-like reasoning and emotions. Specifically, the Reinforcement Learning paradigm, based on rewards and punishments, becomes a point of question. *How does penalizing an AI entity with a -10 penalty differ from inflicting a physical slap as punishment?* The primary distinction lies in our control over the virtual realm in which these machines live, rendering their experiences less tangible compared to our reality. This control is akin to an author's power over the fate and pain of fictional characters in a story. Yet, unlike fictional characters, machines possess the capability to interact with our world and us. Therefore, the ethics of 'punishing' such entities demands contemplation, especially as they grow increasingly sophisticated and sentient-like in their interactions with the human world.

8.4.2 AI and the Future of Work: Progress or Pressure?

In this thesis, our focus has been on enhancing human-machine interaction, particularly in communication. We discussed complex concepts like agency and misalignment, critically analyzing recent advancements in Reinforcement Learning Fine-Tuning LLMs and their shortcomings in addressing misalignment issues. Yet, the broader societal implications of AI advancements remain to be fully explored.

The current AI evolution parallels the Industrial Revolution in many ways but with distinct implications. The Industrial Revolution's machines replaced human labor, improving living standards but also displacing jobs. While this shift ultimately led to the creation of new, more specialized roles requiring advanced education (such as mechanics instead of manual laborers), it inevitably left behind those unable to adapt to the rapid changes, whether due to lack of resources, skills, or other constraints. This transition, while broadly beneficial, impacted individual livelihoods and societal structures. Today's advancements in AI hold the potential to similarly reshape society, automating repetitive tasks in favor of roles that demand more uniquely human expertise. This transition is likely to create and encourage new job opportunities, such as those in advanced fields like computer science, mirroring the changes brought about by the Industrial Revolution. However, just as before, this shift raises important questions about the evolving landscape of work and our collective priorities.

If AI's capability transforms a typical five-day workweek into four days, *how does this reshape our perception of work and leisure? Will individuals be encouraged to enjoy an extra day of rest, or will the pressure to increase productivity push for more work within the reduced timeframe?* These considerations are deeply rooted in societal value alignment: *what do we prioritize as a society?*

Though I do not claim to have the answers, I envision a future where AI's efficiency allows us to reevaluate our work-focused lifestyle. As AI reshapes our efficiency and production capabilities, we must continually question and assert our values, ensuring that technological advancements enhance, rather than worsen, our quality of life.

8.4.3 The AI Race: Balancing Innovation with Responsibility

The contemporary trend in artificial intelligence is a race towards creating ever-larger and more advanced AI models. While these developments are undoubtedly groundbreaking, it becomes crucial to pause and consider the broader implications of such rapid advancements. This section explores several critical aspects often overshadowed by the allure of AI's capabilities.

Inequality in Compute Power and Innovation Access

The demand for substantial computational resources to develop and operate advanced AI models like transformers highlights a significant disparity in the capability to innovate between well-funded industry giants and resource-constrained academic institutions. This disparity limits the progress of new research in places that usually support the early development of ideas and widens the technological divide across different regions and economic backgrounds. As addressed in the discussion of LSTM limitations, the necessity for high compute power often favors entities with greater resources, potentially leading to a concentration of AI advancements in fewer hands. This concentration risks exacerbating global inequalities and hinders the democratization of AI technology. Recognizing and addressing these challenges is essential for ensuring that advancements in AI contribute positively across all sectors of society, not just those with the most substantial financial capabilities.

Environmental Impact

Following the concern of compute inequality, the environmental footprint of AI development is another significant aspect that needs urgent attention. Training large-scale AI models require substantial computational resources and energy, leading to notable carbon emissions. As AI technology advances, adopting an interdisciplinary approach that integrates sustainable practices into every stage of AI development is essential. This effort extends beyond software optimization to include collaboration with hardware researchers and manufacturers. The goal is to innovate and create hardware that is not only more powerful but also energy-efficient.

Moreover, this issue feeds into the broader debate about energy sources, particularly the balance between renewable and non-renewable options. The AI field, therefore, has a responsibility to advocate for and utilize renewable energy sources wherever possible, thereby minimizing its environmental impact.

AI in Classrooms

The introduction of Large Language Models into the educational sector holds both potential for enhancing learning and poses significant risks, particularly in the context of the current education system. Students, faced with the pressure of academic success, might resort to using LLMs to compose entire texts, thereby bypassing the essential learning process. This challenge necessitates a shift in our educational paradigms. Rather than emphasizing memorization and writing skills, education systems should pivot towards promoting critical thinking and analytical abilities. The ability to discern text inconsistencies, whether AI-generated or otherwise, should be central to this new educational system.

On the positive side, LLMs can be transformative in personalizing learning. They can adapt educational content to suit individual learning styles and paces, a practice proven to be highly effective [Gómez et al., 2014, Zhang et al., 2020a]. This personalized approach could allow teachers to focus on facilitating deeper understanding and addressing specific learning challenges.

In essence, while LLMs bring potential risks to the education system, they also open up opportunities for a more inclusive and effective learning environment.

The Human Cost of AI

The development of Large Language Models like ChatGPT relies heavily on human annotators, who play a role in processing and filtering massive amounts of data. These annotators are often tasked with sifting through internet content, selecting or excluding materials that could influence the training process of these AI systems. This task exposes the annotators to a wide array of content, including potentially disturbing or traumatic material.

For instance, annotators in Kenya, who have contributed significantly to training models like ChatGPT, often encounter distressing content as part of their job [Guardian, 2023]. The nature of this work can have long-lasting psychological impacts, with stress and trauma potentially persisting long after the job is done. This raises an ethical question: *Should the advancement of AI, aimed at enhancing human life, be built upon the potential suffering of humans?*

It is crucial to consider the human cost behind these advancements. Providing adequate mental health support, creating safer work environments, and developing strategies to minimize exposure to harmful content is essential to ensure ethical AI development.

Copyright Issues in AI Contributions

The integration of human-generated data into AI systems presents complex copyright issues that need urgent addressing. As modern Generative AI technologies utilize vast datasets often provided by human contributors, questions arise about the ownership and copyright of the outputs generated by these systems. This debate extends to whether contributors should retain any legal rights over their input or the AI-generated content that results from it.

The current legal frameworks are often not equipped to handle the unique challenges posed by AI-generated content, leading to potential conflicts between AI developers, users, and content providers. There is a growing need for laws that recognize and regulate the contributions of both humans and AI systems in a way that protects the rights of all parties involved.

The Workforce in an AI World: Anxiety and Adaptation

Continuing on the topic of mental health, the advancement of AI and automation extends its impact on the psychological well-being of individuals, especially young adults. The prospect of AI replacing human roles in various sectors has ignited a pervasive sense of employment anxiety. This fear is not just about the potential loss of jobs but also about the uncertainty of the future in an increasingly AI-driven world.

This anxiety is rooted in the perception of AI as a threat to job security, fueling concerns about the relevance of human skills in the future workforce. To address this, it becomes vital for society to shift its perspective on AI and automation. Rather than viewing these technologies as replacements for human workers, we should see them as complementary to human skills. Emphasizing a cooperative relationship between technology and the workforce can help alleviate some of the fears associated with AI-driven automation.

In light of this, there is a need for enhanced mental health support. Providing counseling services, stress management workshops, and educational courses about AI and its role in the future of work can be effective ways to help people. Additionally, promoting environments where individuals can learn to use and understand these technologies can empower them, turning fear into a sense of control for the future.

Concluding Thoughts

The race towards achieving AI supremacy should not be blinded by the pursuit of technological excellence alone. We must broaden our focus to prioritize the long-term societal, environmental, and ethical implications of AI development. By doing so, we can ensure that AI advancements truly enhance the quality of life and well-being of all involved, paving the way for a more balanced and conscientious technological future.

Bibliography

Abbeel, P. and Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In Proceedings of the twenty-first international conference on Machine learning, page 1.

Abramson, G. and Kuperman, M. (2001). Social games in a social network. Physical Review E, 63(3):030901.

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., and Goldberg, Y. (2017). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net.

Agarap, A. F. (2018). Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375.

Ahmed, Z., Roux, N. L., Norouzi, M., and Schuurmans, D. (2019). Understanding the impact of entropy on policy optimization. In Chaudhuri, K. and Salakhutdinov, R., editors, Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of Proceedings of Machine Learning Research, pages 151–160. PMLR.

Albrecht, S. V. and Stone, P. (2018). Autonomous agents modelling other agents: A comprehensive survey and open problems. Artificial Intelligence, 258:66–95.

Alizadeh Alamdari, P., Klassen, T. Q., Toro Icarte, R., and McIlraith, S. A. (2022). Be considerate: Avoiding negative side effects in reinforcement learning. In Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, pages 18–26.

Amodei, D. and Clark, J. (2016). Faulty reward functions in the wild.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in ai safety. arXiv preprint arXiv:1606.06565.

Andreas, J. and Klein, D. (2016). Reasoning about pragmatics with neural listeners and speakers. In Su, J., Carreras, X., and Duh, K., editors, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, pages 1173–1182. The Association for Computational Linguistics.

Arcadi, A. C. (2000). Vocal responsiveness in male wild chimpanzees: implications for the evolution of language. Journal of human evolution, 39 2:205–23.

Arora, S. and Doshi, P. (2021). A survey of inverse reinforcement learning: Challenges, methods and progress. Artificial Intelligence, 297:103500.

Bagnell, J., Chestnutt, J., Bradley, D., and Ratliff, N. (2006). Boosting structured prediction for imitation learning. Advances in Neural Information Processing Systems, 19.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Bengio, Y. and LeCun, Y., editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022a). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. (2022b). Constitutional ai: Harmlessness from ai feedback.

Baker, C. L., Saxe, R., and Tenenbaum, J. B. (2009). Action understanding as inverse planning. Cognition, 113(3):329–349.

Banihashem, K., Singla, A., Gan, J., and Radanovic, G. (2022). Admissible policy teaching through reward design. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 6037–6045.

Barandiaran, X. E., Di Paolo, E., and Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. Adaptive Behavior, 17(5):367–386.

Baron-Cohen, S., Leslie, A. M., and Frith, U. (1985). Does the autistic child have a “theory of mind”? Cognition, 21(1):37–46.

Ben-David, E., Oved, N., and Reichart, R. (2022). PADA: Example-based prompt learning for on-the-fly adaptation to unseen domains. Transactions of the Association for Computational Linguistics, 10:414–433.

Bengio, Y., Ducharme, R., and Vincent, P. (2000). A neural probabilistic language model. Advances in neural information processing systems, 13:1137–1155.

Bessen, J. (2018). Artificial intelligence and jobs: The role of demand. In The economics of artificial intelligence: an agenda, pages 291–307. University of Chicago Press.

Bi, X. and Tanaka, T. (2016). Human-side strategies in the werewolf game against the stealth werewolf strategy. In International Conference on Computers and Games, pages 93–102. Springer.

Bowie, A. M. (2007). Herodotus: Histories Book VIII. Cambridge University Press.

Brandizzi, N. (2023). Towards more human-like ai communication: A review of emergent communication research. IEEE Access, pages 1–1.

Brandizzi, N., Grossi, D., and Iocchi, L. (2021). Rlupus: Cooperation through emergent communication in the werewolf social deduction game. Intelligenza Artificiale, 15:55–70. 2.

Brandizzi, N. and Iocchi, L. (2022). Emergent communication in human-machine games. In Emergent Communication Workshop at ICLR 2022.

Brandizzi, N., Takmaz, E., Giulianelli, M., Pezzelle, S., and Fernandez, R. (2023). Speaking the language of your listener: Audience-aware adaptation via plug-and-play theory of mind. In Findings of the Association for Computational Linguistics: ACL 2023, pages 4198–4217, Toronto, Canada. Association for Computational Linguistics.

Brandon, R. N. (2014). Adaptation and environment, volume 1040. Princeton University Press.

Braverman, M., Etesami, O., Mossel, E., et al. (2008). Mafia: A theoretical study of players and coalitions in a partial information environment. The Annals of Applied Probability, 18(3):825–846.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Butlin, P. (2023). Reinforcement learning and artificial agency. Mind & Language.

Butlin, P. et al. (2022). Machine learning, functions and goals. Croatian Journal of Philosophy, 22(66):351–370.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., et al. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv preprint arXiv:2308.08708.

Campbell, M., Hoane Jr, A. J., and Hsu, F.-h. (2002). Deep blue. Artificial intelligence, 134(1-2):57–83.

Cao, K., Lazaridou, A., Lanctot, M., Leibo, J. Z., Tuyls, K., and Clark, S. (2018). Emergent communication through negotiation. CoRR, abs/1804.03980.

Card, S. K. (2018). The psychology of human-computer interaction. Crc Press.

Carlisle, J. H. (1976). Evaluating the impact of office automation on top management communication. In Proceedings of the June 7-10, 1976, National Computer Conference and Exposition, AFIPS '76, page 611–616, New York, NY, USA. Association for Computing Machinery.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., et al. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. arXiv preprint arXiv:2307.15217.

Chaabouni, R., Kharitonov, E., Bouchacourt, D., Dupoux, E., and Baroni, M. (2020). Compositionality and generalization in emergent languages. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. R., editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 4427–4442. Association for Computational Linguistics.

Chaabouni, R., Strub, F., Altché, F., Tarassov, E., Tallec, C., Davoodi, E., Mathewson, K. W., Tieleman, O., Lazaridou, A., and Piot, B. (2022). Emergent communication at scale. In The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net.

Chan, K. T., King, I., and Yuen, M.-C. (2009). Mathematical modeling of social games. In 2009 International Conference on Computational Science and Engineering, volume 4, pages 1205–1210. IEEE.

Chella, A. and Manzotti, R. (2013). Artificial consciousness. Andrews UK Limited.

Cherry, C. (1966). On human communication.

Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Moschitti, A., Pang, B., and Daelemans, W., editors, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pages 1724–1734. ACL.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. (2023). Palm: Scaling language modeling with pathways. J. Mach. Learn. Res., 24:240:1–240:113.

Christian, B. (2020). The alignment problem: Machine learning and human values. WW Norton & Company.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in neural information processing systems, 30.

Clusmann, J., Kolbinger, F. R., Muti, H. S., Carrero, Z. I., Eckardt, J.-N., Laleh, N. G., Löffler, C. M. L., Schwarzkopf, S.-C., Unger, M., Veldhuizen, G. P., et al. (2023). The future landscape of large language models in medicine. Communications Medicine, 3(1):141.

Colman, A. M. (2003). Cooperation, psychological game theory, and limitations of rationality in social interaction. Behavioral and brain sciences, 26(2):139–153.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. (2018). What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Consalvo, M. (2011). Using your friends: Social mechanics in social games. In Proceedings of the 6th International Conference on Foundations of Digital Games, pages 188–195.

Dagan, G., Hupkes, D., and Bruni, E. (2020). Co-evolution of language and agents in referential games. CoRR, abs/2001.03361.

Darwin, C. (1964). On the origin of species: A facsimile of the first edition. Harvard University Press.

Das, A., Kottur, S., Moura, J. M., Lee, S., and Batra, D. (2017). Learning cooperative visual dialog agents with deep reinforcement learning. In Proceedings of the IEEE international conference on computer vision, pages 2951–2960.

Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., Yosinski, J., and Liu, R. (2020). Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In International Conference on Learning Representations.

Davenport, T. H. and Kirby, J. (2016). Only humans need apply: Winners and losers in the age of smart machines. Harper Business New York.

Dennett, D. C. (2008). Kinds of minds: Toward an understanding of consciousness. Basic Books.

Dessì, R., Bevilacqua, M., Gualdoni, E., Rakotonirina, N. C., Franzon, F., and Baroni, M. (2023). Cross-domain image captioning with discriminative finetuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6935–6944.

Dessì, R., Bevilacqua, M., Gualdoni, E., Rakotonirina, N. C., Franzon, F., and Baroni, M. (2023). Cross-domain image captioning with discriminative finetuning. CoRR, abs/2304.01662.

Dolan, R. J. and Dayan, P. (2013). Goals and habits in the brain. Neuron, 80(2):312–325.

Dragan, A. D., Lee, K. C., and Srinivasa, S. S. (2013). Legibility and predictability of robot motion. In 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pages 301–308. IEEE.

Dretske, F. (1985). Machines and the mental. In Proceedings and Addresses of the American Philosophical Association, volume 59, pages 23–33. JSTOR.

Dretske, F. (1991). Explaining behavior: Reasons in a world of causes. MIT press.

Eger, M. and Martens, C. (2018). Keeping the story straight: A comparison of commitment strategies for a social deduction game. In Fourteenth Artificial Intelligence and Interactive Digital Entertainment Conference.

Etzrodt, K., Gentzel, P., Utz, S., and Engesser, S. (2022). Human-machine-communication: introduction to the special issue. Publizistik, 67(4):439–448.

Everitt, T., Krakovna, V., Orseau, L., and Legg, S. (2017). Reinforcement learning with a corrupted reward channel. In Sierra, C., editor, Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017, pages 4705–4713. ijcai.org.

Eysenbach, B., Gupta, A., Ibarz, J., and Levine, S. (2019). Diversity is all you need: Learning skills without a reward function. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net.

(FAIR)†, M. F. A. R. D. T., Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., et al. (2022). Human-level play in the game of diplomacy by combining language models with strategic reasoning. Science, 378(6624):1067–1074.

Farina, M. and Lavazza, A. (2023). Chatgpt in society: emerging issues. Frontiers in Artificial Intelligence, 6:1130913.

Faulkner, T. A. K., Short, E. S., and Thomaz, A. L. (2020). Interactive reinforcement learning with inaccurate feedback. In 2020 IEEE International Conference on Robotics and Automation (ICRA), pages 7498–7504. IEEE.

Ferrucci, D., Levas, A., Bagchi, S., Gondek, D., and Mueller, E. T. (2013). Watson: beyond jeopardy! Artificial Intelligence, 199:93–105.

Fisac, J. F., Gates, M. A., Hamrick, J. B., Liu, C., Hadfield-Menell, D., Palaniappan, M., Malik, D., Sastry, S. S., Griffiths, T. L., and Dragan, A. D. (2020). Pragmatic-pedagogic value alignment. In Robotics Research: The 18th International Symposium ISRR, pages 49–57. Springer.

Foerster, J. N., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., and Mordatch, I. (2017). Learning with opponent-learning awareness. arXiv preprint arXiv:1709.04326.

Frank, M. C. and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. Science, 336(6084):998–998.

Frattolillo, F., Brandizzi, N., Cipollone, R., and Iocchi, L. (2023). Towards computational models for reinforcement learning in human-ai teams. In Proceedings of the 2nd International Workshop on Multidisciplinary Perspectives on Human-AI Team Trust (MULTITTRUST 2.0), Gothenburg, Sweden.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and machines, 30(3):411–437.

Gallagher, H. L. and Frith, C. D. (2003). Functional imaging of 'theory of mind'. Trends in cognitive sciences, 7(2):77–83.

Gallese, V. and Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. Trends in cognitive sciences, 2(12):493–501.

Ganguli, D., Askell, A., Schiefer, N., Liao, T., Lukošiūtė, K., Chen, A., Goldie, A., Mirhoseini, A., Olsson, C., Hernandez, D., et al. (2023). The capacity for moral self-correction in large language models. arXiv preprint arXiv:2302.07459.

Gibbons, R. (1998). Incentives in organizations. Journal of economic perspectives, 12(4):115–132.

Gómez, S., Zervas, P., Sampson, D. G., and Fabregat, R. (2014). Context-aware adaptive and personalized mobile learning delivery supported by uolmp. Journal of King Saud University-Computer and Information Sciences, 26(1):47–61.

Goodhart, C. A. and Goodhart, C. (1984). Problems of monetary management: the UK experience. Springer.

Goodman, J. (2001). Classes for fast maximum entropy training. In 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), volume 1, pages 561–564. IEEE.

Goodman, N. D. and Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. Trends in cognitive sciences, 20(11):818–829.

Goodman, N. D. and Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. Topics in cognitive science, 5(1):173–184.

Gopnik, A. and Wellman, H. M. (1992). Why the child's theory of mind really is a theory. Mind & Language, 7(1-2):145–171.

Graesser, L., Cho, K., and Kiela, D. (2019). Emergent linguistic phenomena in multi-agent communication games. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 3698–3708. Association for Computational Linguistics.

Guan, Y., Ren, Y., Li, S. E., Sun, Q., Luo, L., and Li, K. (2020). Centralized cooperation for connected and automated vehicles at intersections by proximal policy optimization. IEEE Transactions on Vehicular Technology, 69(11):12597–12608.

Guardian (2022). Google fires software engineer who claims ai chatbot is sentient.

Guardian, T. (2023). "it's destroyed me completely": Kenyan moderators decry toll of training of ai models.

Haber, J., Baumgärtner, T., Takmaz, E., Gelderloos, L., Bruni, E., and Fernández, R. (2019). The PhotoBook dataset: Building common ground through visually-grounded dialogue. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1895–1910.

Hadfield-Menell, D., Milli, S., Abbeel, P., Russell, S. J., and Dragan, A. (2017). Inverse reward design. Advances in neural information processing systems, 30.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., and Dragan, A. (2016). Cooperative inverse reinforcement learning. Advances in neural information processing systems, 29.

Happé, F. G. (1993). Communicative competence and theory of mind in autism: A test of relevance theory. Cognition, 48(2):101–119.

Harris, P. L., Johnson, C. N., Hutton, D., Andrews, G., and Cooke, T. (1989). Young children's theory of mind and emotion. Cognition & Emotion, 3(4):379–400.

Havrylov, S. and Titov, I. (2017). Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In Advances in neural information processing systems, pages 2149–2159.

Hawkins, R. X. D., Kwon, M., Sadigh, D., and Goodman, N. D. (2020). Continual adaptation for efficient machine communication. In Fernández, R. and Linzen, T., editors, Proceedings of the 24th Conference on Computational Natural Language Learning, CoNLL 2020, Online, November 19-20, 2020, volume abs/1911.09896, pages 408–419. Association for Computational Linguistics.

Hayes, C. F., Rădulescu, R., Bargiacchi, E., Källström, J., Macfarlane, M., Reymond, M., Verstraeten, T., Zintgraf, L. M., Dazeley, R., Heintz, F., et al. (2022). A practical guide to multi-objective reinforcement learning and planning. Autonomous Agents and Multi-Agent Systems, 36(1):26.

He, H., Boyd-Graber, J., Kwok, K., and Daumé III, H. (2016a). Opponent modeling in deep reinforcement learning. In International conference on machine learning, pages 1804–1813. PMLR.

He, K., Zhang, X., Ren, S., and Sun, J. (2016b). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778.

Hill, F., Mokra, S., Wong, N., and Harley, T. (2020). Human instruction-following with deep reinforcement learning via transfer-learning from text. arXiv preprint arXiv:2005.09382.

Hirata, Y., Inaba, M., Takahashi, K., Toriumi, F., Osawa, H., Katagami, D., and Shinoda, K. (2016). Werewolf game modeling using action probabilities based on play log analysis. In International Conference on Computers and Games, pages 103–114. Springer.

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 6(02):107–116.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2020). The curious case of neural text degeneration. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In Chaudhuri,

K. and Salakhutdinov, R., editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 2790–2799. PMLR.

Hu, K. (2023). Chatgpt sets record for fastest-growing user base - analyst note.

Hupkes, D., Veldhoen, S., and Zuidema, W. (2018). Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. Journal of Artificial Intelligence Research, 61:907–926.

Hurley, S. L. (1998). Consciousness in action. Harvard University Press.

Hutchins, W. J. (1999). Retrospect and prospect in computer-based translation. In Proceedings of Machine Translation Summit VII, pages 30–36.

Jackson, F. (1998). Epiphenomenal qualia. In Consciousness and emotion in cognitive science, pages 197–206. Routledge.

Jang, E., Gu, S., and Poole, B. (2017). Categorical reparameterization with gumbel-softmax. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net.

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. Current Opinion in Behavioral Sciences, 29:105–110.

Kaiser, L. and Bengio, S. (2018). Discrete autoencoders for sequence models. CoRR, abs/1801.09797.

Kajiwara, K., Toriumi, F., Ohashi, H., Osawa, H., Katagami, D., Inaba, M., Shinoda, K., Nishino, J., et al. (2014). Extraction of optimal strategies in human wolf using reinforcement learning. Proceedings of the 76th National Convention, 2014(1):597–598.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3128–3137.

Katagami, D., Takaku, S., Inaba, M., Osawa, H., Shinoda, K., Nishino, J., and Toriumi, F. (2014). Investigation of the effects of nonverbal information on werewolf. In 2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pages 982–987. IEEE.

Kerr, S. (1975). On the folly of rewarding a, while hoping for b. Academy of Management journal, 18(4):769–783.

Kim, Y. Y. (2001). Becoming intercultural: An integrative theory of communication and cross-cultural adaptation. Sage.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Klopf, A. H. (1972). Brain function and adaptive systems: a heterostatic theory. Air Force Cambridge Research Laboratories, Air Force Systems Command, United˜. . . .

Kolter, J., Abbeel, P., and Ng, A. (2007). Hierarchical apprenticeship learning with application to quadruped locomotion. Advances in Neural Information Processing Systems, 20.

Kottur, S., Moura, J. M. F., Lee, S., and Batra, D. (2017). Natural language does not emerge 'naturally' in multi-agent dialog. CoRR, abs/1706.08502:2962–2967.

Krakovna, V., Orseau, L., Martic, M., and Legg, S. (2018). Measuring and avoiding side effects using relative reachability. arXiv preprint arXiv:1806.01186.

Krishna, R., Lee, D., Fei-Fei, L., and Bernstein, M. S. (2022). Socially situated artificial intelligence enables learning from human interaction. Proceedings of the National Academy of Sciences, 119(39):e2115730119.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. The annals of mathematical statistics, 22(1):79–86.

Kuperman, V., Stadthagen-Gonzalez, H., and Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. Behavior Research Methods, 44(4):978–990.

Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z.-R., Stevens, K., Barhoum, A., Duc, N. M., Stanley, O., Nagyfi, R., ES, S., Suri, S., Glushkov, D., Dantuluri, A., Maguire, A., Schuhmann, C., Nguyen, H., and Mattick, A. (2023). Openassistant conversations – democratizing large language model alignment.

Lamb, A. M., ALIAS PARTH GOYAL, A. G., Zhang, Y., Zhang, S., Courville, A. C., and Bengio, Y. (2016). Professor forcing: A new algorithm for training recurrent networks. Advances in neural information processing systems, 29.

Lazaridou, A., Peysakhovich, A., and Baroni, M. (2017). Multi-agent cooperation and the emergence of (natural) language. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net.

Lazaridou, A., Potapenko, A., and Tieleman, O. (2020). Multi-agent communication meets natural language: Synergies between functional and structural language learning. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. R., editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 7663–7674, Online. Association for Computational Linguistics.

Lee, J., Cho, K., Weston, J., and Kiela, D. (2017). Emergent translation in multi-agent communication. arXiv preprint arXiv:1710.06922.

Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., Orseau, L., and Legg, S. (2017). Ai safety gridworlds. arXiv preprint arXiv:1711.09883.

Lewis, D. K. (1969). Convention: A Philosophical Study. Cambridge, MA, USA: Wiley-Blackwell.

Li, F. and Bowling, M. (2019). Ease-of-teaching and language structure from emergent communication. In Advances in Neural Information Processing Systems, pages 15851–15861.

Li, X. L. and Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4582–4597, Online. Association for Computational Linguistics.

Li, Y., Ponti, E. M., Vulic, I., and Korhonen, A. (2020). Emergent communication pretraining for few-shot machine translation. In Scott, D., Bel, N., and Zong, C., editors, Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020, pages 4716–4731. International Committee on Computational Linguistics.

Liang, P. P., Chen, J., Salakhutdinov, R., Morency, L., and Kottur, S. (2020). On emergent communication in competitive multi-agent teams. CoRR, abs/2003.01848:735–743.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014a). Microsoft COCO: common objects in context. CoRR, abs/1405.0312:740–755.

Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014b). Microsoft COCO: common objects in context. In Fleet, D. J., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V, volume 8693 of Lecture Notes in Computer Science, pages 740–755. Springer.

Lipinski, O., Sobey, A., Cerutti, F., and Norman, T. (2022). Emergent password signalling in the game of werewolf. In Emergent Communication Workshop at ICLR 2022 (29/04/22 - 29/04/22).

Liu, E. Z., Suri, S., Mu, T., Zhou, A., and Finn, C. (2023a). Simple embodied language learning as a byproduct of meta-reinforcement learning. arXiv preprint arXiv:2306.08400.

Liu, H., Sferrazza, C., and Abbeel, P. (2023b). Chain of hindsight aligns language models with feedback. arXiv preprint arXiv:2302.02676, 3.

Loshchilov, I. and Hutter, F. (2017). SGDR: stochastic gradient descent with warm restarts. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net.

Lowe, R., Foerster, J. N., Boureau, Y., Pineau, J., and Dauphin, Y. N. (2019). On the pitfalls of measuring emergent communication. In Elkind, E., Veloso, M., Agmon, N., and Taylor, M. E., editors, Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019, pages 693–701. International Foundation for Autonomous Agents and Multiagent Systems.

Lowe, R., Gupta, A., Foerster, J. N., Kiela, D., and Pineau, J. (2020). On the interaction between supervision and self-play in emergent communication. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Maddison, C. J., Mnih, A., and Teh, Y. W. (2017). The concrete distribution: A continuous relaxation of discrete random variables. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net.

Manyika, J. (2023). An overview of bard: an early experiment with generative ai. Technical report, Technical report, Google AI.

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., and Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 11–20.

Mehrabian, A. et al. (1971). Silent messages, volume 8. Wadsworth Belmont, CA.

Migdal, P. (2010). A mathematical model of the mafia game. CoRR, abs/1009.1031.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In Bengio, Y. and LeCun, Y., editors, 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings.

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In Interspeech, volume 2, pages 1045–1048. Makuhari.

Mikolov, T., Kombrink, S., Burget, L., Černockỳ, J., and Khudanpur, S. (2011). Extensions of recurrent neural network language model. In 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 5528–5531. IEEE.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In International conference on machine learning, pages 1928–1937. PMLR.

Morin, F. and Bengio, Y. (2005). Hierarchical probabilistic neural network language model. In International workshop on artificial intelligence and statistics, pages 246–252. PMLR.

Nakamura, N., Inaba, M., Takahashi, K., Toriumi, F., Osawa, H., Katagami, D., and Shinoda, K. (2016). Constructing a human-like agent for the werewolf game using a psychological model based multiple perspectives. In 2016 IEEE Symposium Series on Computational Intelligence (SSCI), pages 1–8. IEEE.

Ng, A. Y., Harada, D., and Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In Icml, volume 99, pages 278–287. Citeseer.

Ng, A. Y., Russell, S., et al. (2000). Algorithms for inverse reinforcement learning. In Icml, volume 1, page 2.

Nguyen, K. X., Misra, D., Schapire, R., Dudík, M., and Shafto, P. (2021). Interactive learning from activity description. In International Conference on Machine Learning, pages 8096–8108. PMLR.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. science, 314(5805):1560–1563.

Nowak, M. A. and Krakauer, D. C. (1999). The evolution of language. Proceedings of the National Academy of Sciences, 96(14):8028–8033.

OpenAI (2023). Gpt-4 technical report.

Orzan, N., Acar, E., Grossi, D., and Rădulescu, R. (2023). Emergent cooperation and deception in public good games. In 2023 Adaptive and Learning Agents Workshop at AAMAS.

Osa, T., Pajarinen, J., Neumann, G., Bagnell, J. A., Abbeel, P., Peters, J., et al. (2018). An algorithmic perspective on imitation learning. Foundations and Trends® in Robotics, 7(1-2):1–179.

Papadimitriou, I. and Jurafsky, D. (2020). Pretraining on non-linguistic structure as a tool for analyzing learning bias in language models. arXiv preprint arXiv:2004.14601.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pages 311–318. Association for Computational Linguistics.

Pascual, D., Egressy, B., Meister, C., Cotterell, R., and Wattenhofer, R. (2021). A plug-and-play method for controlled text generation. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3973–3997, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al. (2022). Discovering language model behaviors with model-written evaluations. arXiv preprint arXiv:2212.09251.

Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., and Gurevych, I. (2020a). AdapterHub: A framework for adapting transformers. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 46–54, Online. Association for Computational Linguistics.

Pfeiffer, J., Vulić, I., Gurevych, I., and Ruder, S. (2020b). MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7654–7673, Online. Association for Computational Linguistics.

Pistono, F. and Yampolskiy, R. V. (2016). Unethical research: how to create a malevolent artificial intelligence.

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? Behavioral and brain sciences, 1(4):515–526.

Qin, L., Welleck, S., Khashabi, D., and Choi, Y. (2022). Cold decoding: Energy-based constrained text generation with langevin dynamics. In Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., Eslami, S. M. A., and Botvinick, M. M. (2018). Machine theory of mind. CoRR, abs/1802.07740.

Raileanu, R., Denton, E., Szlam, A., and Fergus, R. (2018). Modeling others using oneself in multi-agent reinforcement learning. CoRR, abs/1802.09640:4254–4263.

Ramamurthy, R., Ammanabrolu, P., Brantley, K., Hessel, J., Sifa, R., Bauckhage, C., Hajishirzi, H., and Choi, Y. (2023). Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net.

Ratner, E., Hadfield-Menell, D., and Dragan, A. D. (2018). Simplifying reward design through divide-and-conquer. arXiv preprint arXiv:1806.02501.

Reports, V. (2023). Global large language model(llm) market research report 2023. The Large Language Model (LLM) Market was valued at 10.5 Billion USD in 2022 and is anticipated to reach 40.8 Billion USD by 2029, witnessing a CAGR of 21.4

growth of the big language model market is the need for natural language processing (NLP) technologies across a number of sectors.

Riedl, M. O. (2019). Human-centered artificial intelligence and machine learning. Human Behavior and Emerging Technologies, 1(1):33–36.

Rillig, M. C., Ågerstrand, M., Bi, M., Gould, K. A., and Sauerland, U. (2023). Risks and benefits of large language models for the environment. Environmental Science & Technology, 57(9):3464–3466.

Rodriguez, R. C., Alaniz, S., and Akata, Z. (2019). Modeling conceptual understanding in image reference games. In Advances in Neural Information Processing Systems, pages 13155–13165.

Rose, M. R. and Lauder, G. V. (1996). Adaptation. Academic Press.

Ruiz-Serra, J. and Harré, M. S. (2023). Inverse reinforcement learning as the algorithmic basis for theory of mind: Current methods and open problems. Algorithms, 16(2):68.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. International journal of computer vision, 115:211–252.

Russell, S. (1998). Learning agents for uncertain environments. In Proceedings of the eleventh annual conference on Computational learning theory, pages 101–103.

Salakhutdinov, R. and Hinton, G. E. (2009). Semantic hashing. International Journal of Approximate Reasoning, 50(7):969–978.

Sallam, M. (2023). The utility of chatgpt as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations. medRxiv, pages 2023–02.

Sap, M., LeBras, R., Fried, D., and Choi, Y. (2022). Neural theory-of-mind? on the limits of social intelligence in large lms. arXiv preprint arXiv:2210.13312.

Saparov, A. and He, H. (2022). Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. arXiv preprint arXiv:2210.01240.

Saunders, W., Sastry, G., Stuhlmüller, A., and Evans, O. (2018). Trial without error: Towards safe reinforcement learning via human intervention. In André, E., Koenig, S., Dastani, M., and Sukthankar, G., editors, Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018, pages 2067–2069. International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. CoRR, abs/1707.06347.

Shneiderman, B. (2021). Human-centered AI: A new synthesis. In Ardito, C., Lanzilotti, R., Malizia, A., Petrie, H., Piccinno, A., Desolda, G., and Inkpen, K., editors, Human-Computer Interaction - INTERACT 2021 - 18th IFIP TC 13 International Conference, Bari, Italy, August 30 - September 3, 2021, Proceedings, Part I, volume 12932 of Lecture Notes in Computer Science, pages 3–8. Springer.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. nature, 529(7587):484–489.

Skalse, J., Howe, N., Krasheninnikov, D., and Krueger, D. (2022). Defining and characterizing reward gaming. Advances in Neural Information Processing Systems, 35:9460–9471.

Smith, E. A. (2010). Communication and collective action: language and the evolution of human cooperation. Evolution and human behavior, 31(4):231–245.

Soares, N. and Fallenstein, B. (2014). Aligning superintelligence with human interests: A technical research agenda. Machine Intelligence Research Institute (MIRI) technical report, 8.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1):1929–1958.

State of, W. (2008). Correctional offender management profiling for alternative sanctions.

Stone, V. E., Baron-Cohen, S., and Knight, R. T. (1998). Frontal lobe contributions to theory of mind. Journal of cognitive neuroscience, 10(5):640–656.

Sumers, T. R., Hawkins, R. D., Ho, M. K., Griffiths, T. L., and Hadfield-Menell, D. (2022a). How to talk so your robot will learn: Instructions, descriptions, and pragmatics. arXiv preprint arXiv:2206.07870.

Sumers, T. R., Hawkins, R. D., Ho, M. K., Griffiths, T. L., and Hadfield-Menell, D. (2022b). Linguistic communication as (inverse) reward design. arXiv preprint arXiv:2204.05091.

Sumers, T. R., Ho, M. K., Hawkins, R. D., Narasimhan, K., and Griffiths, T. L. (2021). Learning rewards from linguistic feedback. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 6002–6010.

Sundermeyer, M., Schlüter, R., and Ney, H. (2012). Lstm neural networks for language modeling. In Thirteenth annual conference of the international speech communication association.

Symons, D. (1990). Adaptiveness and adaptation. Ethology and Sociobiology, 11(4):427–444.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1–9.

Takmaz, E., Giulianelli, M., Pezzelle, S., Sinclair, A., and Fernández, R. (2020). Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4350–4368, Online. Association for Computational Linguistics.

Taylor, C. (1985). Human Agency and Language. Cambridge University Press, New York.

Thorn, P. D. (2015). Nick bostrom: Superintelligence: Paths, dangers, strategies - oxford university press, oxford, 2014, xvi+328, £18.99, ISBN: 978-0-19-967811-2. Minds Mach., 25(3):285–289.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023a). Llama: Open and efficient foundation language models. CoRR, abs/2302.13971.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023b). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Tucker, M., Levy, R., Shah, J. A., and Zaslavsky, N. (2022). Trading off utility, informativeness, and complexity in emergent communication. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, Advances in Neural Information Processing Systems, volume 35, pages 22214–22228. Curran Associates, Inc.

Turing, A. M. (1950). Computing machinery and intelligence. Mind, LIX(236):433–460.

Turner, A., Ratzlaff, N., and Tadepalli, P. (2020a). Avoiding side effects in complex environments. Advances in Neural Information Processing Systems, 33:21406–21415.

Turner, A. M., Hadfield-Menell, D., and Tadepalli, P. (2020b). Conservative agency via attainable utility preservation. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pages 385–391.

Valmeekam, K., Olmo, A., Sreedharan, S., and Kambhampati, S. (2022). Large language models still can't plan (a benchmark for llms on planning and reasoning about change). arXiv preprint arXiv:2206.10498.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008.

Vedantam, R., Bengio, S., Murphy, K., Parikh, D., and Chechik, G. (2017). Context-aware captions from context-agnostic supervision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 251–260.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), pages 4566–4575.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. (2019). Grandmaster level in starcraft ii using multi-agent reinforcement learning. Nature, 575(7782):350–354.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3156–3164.

Wainwright, C. L. and Eckersley, P. (2019). Safelife 1.0: Exploring side effects in complex environments. arXiv preprint arXiv:1912.01217.

Wang, J., Liu, Y., and Li, B. (2020). Reinforcement learning with perturbed rewards. In Proceedings of the AAAI conference on artificial intelligence, volume 34, pages 6202–6209.

Wang, R. E., White, J., Mu, J., and Goodman, N. D. (2021). Calibrate your listeners! robust communication-based training for pragmatic speakers. CoRR, abs/2110.05422.

Wang, T. and Kaneko, T. (2018). Application of deep reinforcement learning in werewolf game agents. In 2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI), pages 28–33.

Wei, H., Liu, X., Mashayekhy, L., and Decker, K. (2019). Mixed-autonomy traffic control with proximal policy optimization. In 2019 IEEE Vehicular Networking Conference (VNC), pages 1–8. IEEE.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. (2022). Emergent abilities of large language models. Trans. Mach. Learn. Res., 2022.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning, 8(3-4):229–256.

Williams, R. J. and Peng, J. (1991). Function optimization using connectionist reinforcement learning algorithms. Connection Science, 3(3):241–268.

Wiseman, S. and Lewis, K. (2019). What data do players rely on in social deduction games? In Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts, pages 781–787.

Wolf, Y., Wies, N., Levine, Y., and Shashua, A. (2023). Fundamental limitations of alignment in large language models. arXiv preprint arXiv:2304.11082.

Xie, A., Losey, D., Tolsma, R., Finn, C., and Sadigh, D. (2021). Learning latent representations to influence multi-agent interaction. In Conference on robot learning, pages 575–588. PMLR.

Xu, B., Wang, N., Chen, T., and Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853.

Xu, W. (2019). Toward human-centered ai: A perspective from human-computer interaction. Interactions, 26(4):42–46.

Yao, S., Yu, M., Zhang, Y., Narasimhan, K. R., Tenenbaum, J. B., and Gan, C. (2022). Linking emergent and natural languages via corpus transfer. In The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net.

Yu, L., Tan, H., Bansal, M., and Berg, T. L. (2017). A joint speaker-listener-reinforcer model for referring expressions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3521–3529.

Yuan, L., Fu, Z., Shen, J., Xu, L., Shen, J., and Zhu, S. (2020). Emergence of pragmatics from referential game between theory of mind agents. CoRR, abs/2001.07752.

Yudkowsky, E. (2016). The ai alignment problem: why it is hard, and where to start. Symbolic Systems Distinguished Speaker, 4.

Zhang, L., Basham, J. D., and Yang, S. (2020a). Understanding the implementation of personalized learning: A research synthesis. Educational Research Review, 31:100339.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020b). BERTScore: Evaluating Text Generation with BERT. In International Conference on Learning Representations.

Zhu, H., Neubig, G., and Bisk, Y. (2021). Few-shot language coordination by modeling theory of mind. CoRR, abs/2107.05697.

Appendix A

Meta-Analysis of Thesis Writing Process

This appendix aims to provide a clear and honest view of what goes into writing my doctoral thesis. When I first started my writing, I looked through dozens of theses in my university's library and noticed something missing: there was no trace of the actual effort, the ups and downs, that went into writing them. This lack of transparency inspired me to record my writing process to give future PhD students a real-life reference of what to expect.

I kept track of my writing by simply counting the characters I wrote each day, and later on, I also started counting the math equations and titles. This method gave me a complete picture of my progress. From this data, I learned a lot about how academic writing works. The hardest part was just getting started with writing, which led me to seek help from a writing advisor. Their support got me going and taught me the importance of asking for help. The data also showed that my writing productivity varied, which is expected considering all the different tasks and responsibilities that come with a PhD.

I have created this appendix, especially for other Ph.D. students at my university and in my lab. I hope it can act as both a guide and a source of encouragement, showing an accurate picture of the thesis-writing journey. Keeping track of my writing was not just about the numbers; it was motivational and helped me see how far I had come. Looking back at the comments I made along with the character counts, it is interesting to see how my worries and challenges have shifted over time. With this appendix, I aim to make the thesis-writing process more transparent and relatable for future scholars who are about to start this important academic journey. Beyond that, I hope to set a precedent, encouraging other Ph.D. students to record and share their own journeys. By doing so, we can collectively build a richer, more supportive framework for academic writing that truly reflects the diverse experiences and challenges of PhD life.

A.1 Comparison with Peers

Engaging in comparison with peers and seeking their advice is crucial in academia and life. It not only aids in understanding the process but also instills a sense of control over your work. In this spirit, I began by reviewing theses written by my colleagues and later conducted interviews with those who had recently completed their doctorates.

A.1.1 Analyzing Previous Theses

First, I examined theses written by colleagues in my department, accessible through Sapienza's publication system, Iris. I found 24 published theses[1].

The first aspect I examined was the page count, which, excluding bibliographies, averaged 116.9 pages with a standard deviation of 31.7. Next, I observed the structural pattern: most theses were divided into self-contained chapters based on published material. Initially, I adopted this structure, too, but after consulting with my supervisor, I realized this approach could disrupt the thesis's flow. Therefore, I revised the structure significantly, leading to a notable difference between the initial Table of Contents (Figure A.1) and the current one.

A.1.2 Interview with Graduates

After analyzing the thesis, I sought the experiences and advice of those who had already graduated. I interviewed six of my senior colleagues, all recent Ph.D. graduates, to gather insights. These interviews covered a range of topics, from the technicalities of thesis submission to the personal challenges encountered during the Ph.D. journey.

Summary of Responses

The responses from the interviews highlighted several key aspects:

1. *Thesis Delivery*: The delivery process involves uploading the thesis to the university portal, identifying external reviewers, and meeting specific deadlines. The committee structure varies but typically includes internal and external members.

2. *Review and Approval Time*: It generally takes 1-2 months to receive revisions from reviewers. Auditors' judgment usually ranges from good to excellent, with suggestions for minor or significant changes.

3. *Challenges*: The main challenges include creating a consistent structure and narrative, managing time pressure, and dealing with language barriers. Most emphasized the importance of transforming multiple publications into a coherent document.

4. *Advice for Future Ph.D. Students*: Common suggestions included talking to peers in similar situations, using year-end reports as a starting point, and looking at other theses for structure inspiration. Resources like LaTeX templates, Overleaf, and AI writing tools were recommended.

5. *Post-Doctorate Plans*: Responses varied from pursuing postdoctoral research to exploring job opportunities and research grants.

Effect on My Approach

The insights I gained from these interviews significantly shaped my thesis writing approach. Understanding their experiences, particularly regarding the time commitment they dedicated, offered me a realistic perspective on the process. Their time frames varied from one to six months, averaging

[1] There were many more theses available, but these were not published.

around 2.6 months, with many reporting up to eight hours of daily work. I learned the role of planning in both the structural development of the thesis and the preparation for its defense.

Furthermore, their reflections on seeking support were particularly impactful. It reinforced the value of utilizing available resources, whether leveraging digital tools like Overleaf and AI writing assistants or engaging in discussions with fellow academics.

From them, I learned the value of early planning and consistent effort. Their advice: start with a detailed plan, but be prepared for it to evolve. They also stressed the importance of balancing writing with rest and self-care.

A.2 Methodology of Data Collection and Writing Process

In writing my thesis, I utilized Overleaf, a LaTeX editor that is useful for composition and for tracking my progress. Overleaf offers a feature to count words, including headers and mathematical equations (both inline and displayed). However, I only discovered the capability to track headers and equations halfway through my thesis writing; for this reason, they appear later in my tracking.

My approach to data collection was straightforward: I tracked the days I worked on the thesis rather than the hours I spent each day. Overleaf also keeps a detailed history log in HTML format, which records the date and time of each writing session but not the word count. Later, I downloaded this HTML log and extracted the dates and times for a deeper time analysis. So, the upcoming data analysis will combine these two sources, offering a more complete view of my thesis writing timeline.

It is important to note that my thesis is cumulative, primarily based on my previous publications. This nature of the thesis meant that much of my writing involved adjusting existing texts rather than creating entirely new content. As a result, the time taken for my writing process might differ significantly from those working on monographic theses, where the content is entirely original. This distinction is crucial for anyone looking to compare their progress with mine, as the nature of the thesis heavily influences the writing process.

A.2.1 AI Tools for Writing

A significant part of my writing process involved leveraging AI tools to refine my scientific writing. Being a fluent but non-native English speaker, I found that tools like Grammarly, Wordtune, and ChatGPT were invaluable in enhancing the clarity and readability of my work. While the content and research are my own, the assistance of such technologies has made this thesis more reader-friendly and significantly sped up the writing process.

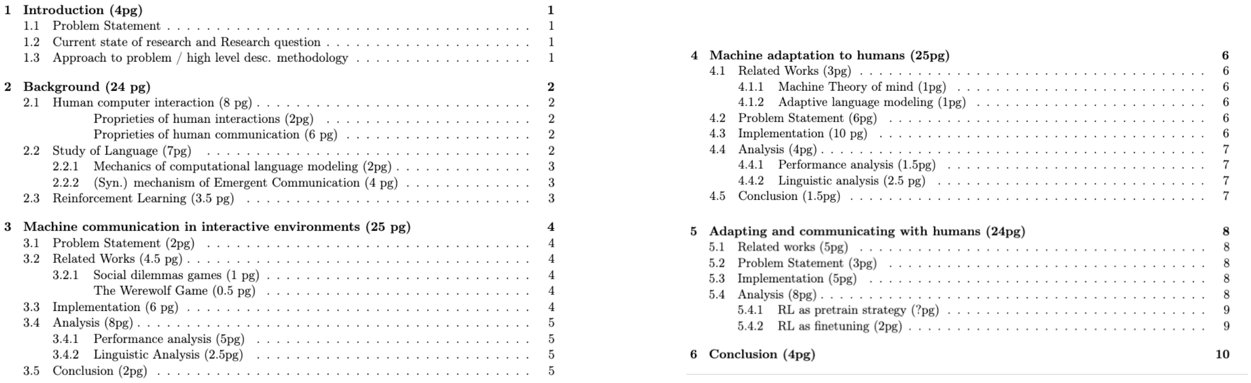
The process was iterative and required careful management. I often found myself guiding the AI towards more suitable expressions or adjusting the wording myself. It is worth noting that while I always provided the content for the AI-generated suggestions, these tools occasionally introduced content that I had not written or suggested, often inaccurately. Therefore, a critical and thorough review of the AI-generated text was essential to ensure accuracy and maintain the integrity of the content.

This process underscores the importance of being critically engaged with AI tools. Moreover, this experience has provided me with additional insights on Human-Machine Interaction and on improving not just this work but also my future research.

A.2.2 Evolution of Thesis Structure

My journey in writing the thesis started with creating the Table of Contents. This step provided a comprehensive view of the entire thesis, helping me envision the structure and flow of my work. The process of outlining the Table of Contents was not just about listing the chapters; it was about framing the narrative I intended to develop through my research.

During this initial phase, an important realization was the insignificance of chronological order in the context of a thesis. Instead, what mattered more was the contextual relevance and how each chapter contributed to answering the research questions. For instance, in my thesis, while Chapter 7 preceded Chapter 6 chronologically, the narrative demanded an inverse order.



1 Introduction (4pg) **1**

1.1 Problem Statement 1

1.2 Current state of research and Research question 1

1.3 Approach to problem / high level desc. methodology 1

2 Background (24 pg) **2**

2.1 Human computer interaction (8 pg) 2

Proprieties of human interactions (2pg) 2

Proprieties of human communication (6 pg) 2

2.2 Study of Language (7pg) 2

2.2.1 Mechanics of computational language modeling (2pg) 3

2.2.2 (Syn.) mechanism of Emergent Communication (4 pg) 3

2.3 Reinforcement Learning (3.5 pg) 3

3 Machine communication in interactive environments (25 pg) **4**

3.1 Problem Statement (2pg) 4

3.2 Related Works (4.5 pg) 4

3.2.1 Social dilemmas games (1 pg) 4

The Werewolf Game (0.5 pg) 4

3.3 Implementation (6 pg) 4

3.4 Analysis (8pg) 5

3.4.1 Performance analysis (5pg) 5

3.4.2 Linguistic Analysis (2.5pg) 5

3.5 Conclusion (2pg) 5

4 Machine adaptation to humans (25pg) **6**

4.1 Related Works (3pg) 6

4.1.1 Machine Theory of mind (1pg) 6

4.1.2 Adaptive language modeling (1pg) 6

4.2 Problem Statement (6pg) 6

4.3 Implementation (10 pg) 6

4.4 Analysis (4pg) 7

4.4.1 Performance analysis (1.5pg) 7

4.4.2 Linguistic analysis (2.5 pg) 7

4.5 Conclusion (1.5pg) 7

5 Adapting and communicating with humans (24pg) **8**

5.1 Related works (5pg) 8

5.2 Problem Statement (3pg) 8

5.3 Implementation (5pg) 8

5.4 Analysis (8pg) 8

5.4.1 RL as pretrain strategy (?pg) 9

5.4.2 RL as finetuning (2pg) 9

6 Conclusion (4pg) **10**

Figure A.1: Initial Table of Contents

When I began writing, I started with the methodological chapters. These chapters mainly involved reshaping work I had already done. This was a practical starting point and made moving into more complex parts of the thesis easier. However, as I progressed, the structure of the thesis naturally evolved. The initial Table of Contents changed quite a bit; new chapters were added, and others were repositioned or revised. This dynamic process is evident when comparing the final structure of my thesis with the initial Table of Contents, which is presented in Figure A.1.

Midway through the thesis writing, I was prompted to write an extended abstract for unrelated reasons. This exercise turned out to be a turning point in my writing journey. It offered me a new perspective on my work, helping me to crystallize the core themes and issues. Notably, it brought into sharper focus the emphasis on concepts like agency and misalignment, aspects that were not as pronounced in the early drafts. Consequently, I highly recommend writing a summary or an extended thesis abstract as a second step, following the Table of Contents.

A.3 Data Analysis

This section details the analysis of my thesis writing data. Initially, I focus on entries from my manual log, recording word counts and dates. Subsequently, I explored insights from Overleaf's history feature. As we will observe, discrepancies emerge between these two sources, prompting a combined analysis for a fuller picture.

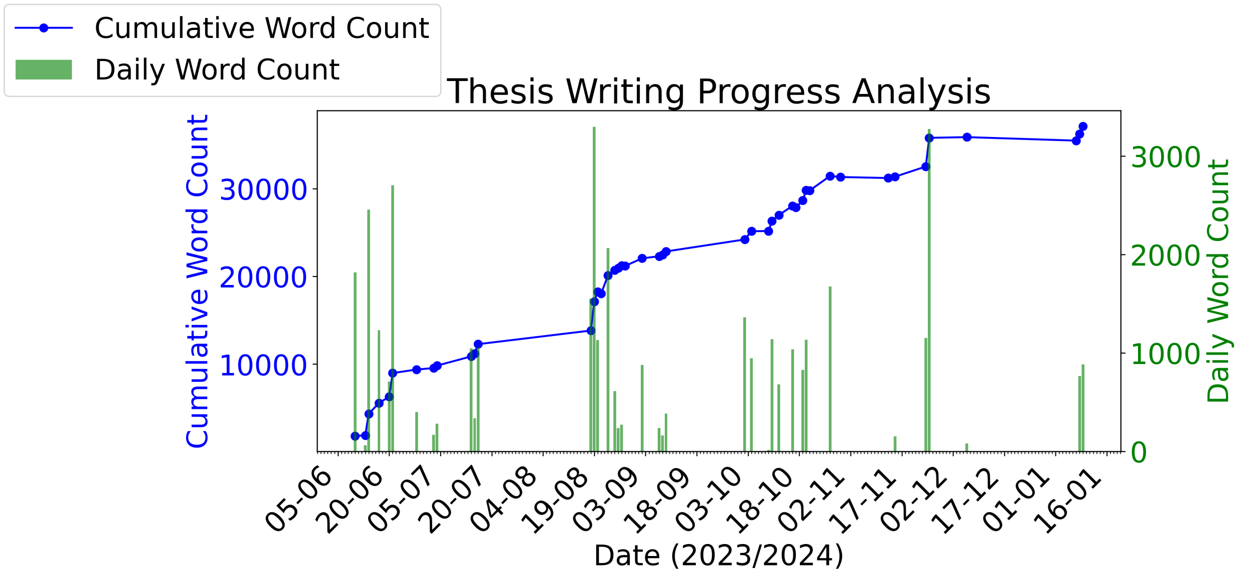


Figure A.2: Daily and Cumulative Word Count: The x-axis represents the dates, and the left y-axis (in green) corresponds to the cumulative word count aligned with the blue line. At the same time, the right y-axis (also in green) relates to the daily word count depicted by the bars.

A.3.1 Manual Logged Entries

Throughout my thesis journey, I wrote a total of 37,128 words across 45 active writing days, spanned over 213 days. The longest pause in writing was 33 days, while my most consistent streak involved writing for four consecutive days. On a weekly average, I crafted 1,256 words.

The graph in Figure A.2 highlights two noticeable gaps. The first, from late July to mid-August, was a period when I attended a conference and took a well-deserved vacation (it is vital to rest and recharge). The second, from mid-September to mid-October, coincided with job interviews, an invited talk, and some intermittent breaks. The graph illustrates phases of intense writing, often followed by rest periods.

A.3.2 Overleaf history

The Overleaf history review revealed some variations: the total writing days rose to 72, and the total writing timeframe shortened to 158 days. The additional days accounted for in the Overleaf history likely indicate days I missed logging manually. At the same time, the reduced timeframe is attributed to starting the Overleaf document from a previously worked-on copy, thus omitting 55 days of activity.

With Overleaf tracking precise modification times, we can estimate the average time spent per writing day, circa 4 hours (03:47:16), and the total time spent, which amounts to 11 days, 09 hours, and 05 minutes. However, these estimates assume continuous work without breaks, based on the first and last entry times within a day, making it a high upper boundary.

The most extended break between writing sessions was reduced to 12 days. Even during vacations, this indicates that I accessed the document for minor tweaks that I then failed to log. Meanwhile, the longest stretch of continuous writing extended to 11 days.

Figure A.3 reflects the updated, more substantial investment of time in writing, providing a representation that more accurately mirrors the effort poured into my thesis.

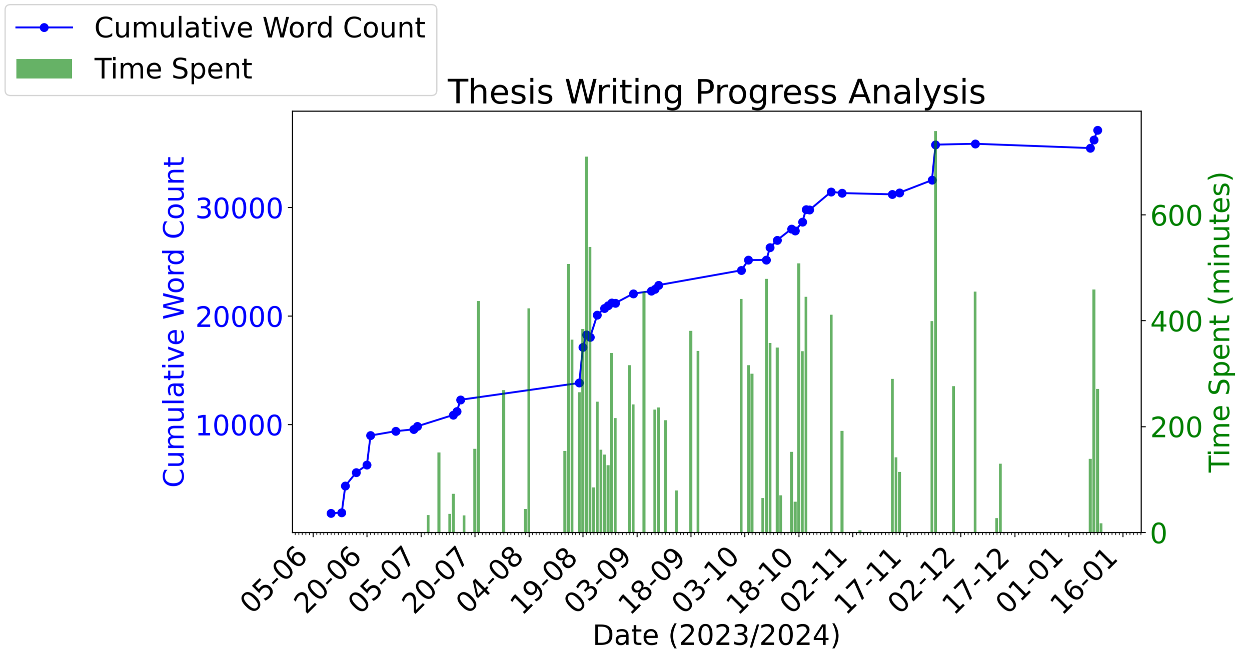


Figure A.3: Aggregated Data on Writing Sessions: This updated graph's x-axis lists the dates; the left y-axis (in green) aligns with the cumulative word count indicated by the blue line, and the right y-axis (in green) represents the daily minutes spent writing, corresponding to the green bars.

A.3.3 Combined analysis

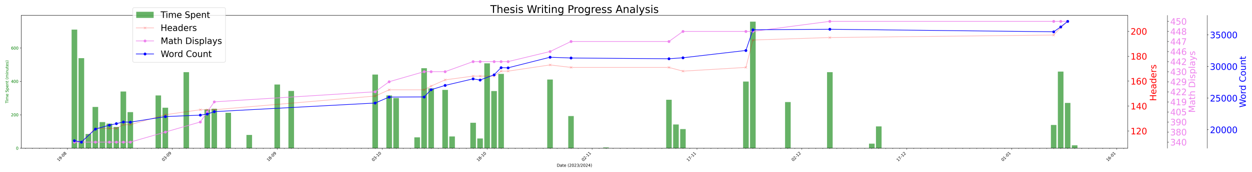


Figure A.4: Structural and Content Development: Correlation between writing time, headers, and math displays over the thesis writing period.

Figure A.4 illustrates the incorporation of headers and mathematical formulas into the graph. It highlights that the frequency of these elements does not consistently align with increases in word count. This disparity suggests a distinct differentiation in the nature of tasks associated with thesis writing. The initial setting up of the structure, as marked by the creation of headers, often precedes the more intensive content development phase. Once the thesis framework is established, the focus shifts predominantly to expanding and refining the content. This observation reflects my personal experience, where showing a clear structure early on facilitated a smoother and more focused content-writing phase.

A.4 Conclusion and Discussion

Crafting my thesis resembled navigating through a complex labyrinth rather than a straightforward journey. Early in the process, I realized the significance of sacrificing chronological order in favor of contextual relevance for each chapter, which proved more effective in addressing my research questions. My advice to those embarking on their thesis is to begin with a well-structured table of contents and a comprehensive summary. This strategy offers a clear roadmap and a defined sense of direction.

The actual writing felt like a series of sprints, with breaks in between for conferences, vacations, or job interviews. Recording my progress was a game-changer. Seeing the word count rise was not just about tracking; it was motivating. It is a practice I would strongly recommend to others.

A.4.1 Challenges, Solutions, and the Role of Technology

The most challenging aspect for me was revisiting and refining my written texts. I discovered that the older the text, the easier it was to revise. Freshly written sections often required some distance before I could critically evaluate and revise them effectively. Additionally, being in a supportive environment, especially with my partner, who was also engaged in thesis writing, proved invaluable. Sharing the experience with someone who understood the process's ups and downs helped validate and navigate through the challenging phases.

Technology played a big part too. Overleaf was great for drafting and tracking my thesis. AI writing tools like grammar checkers and ChatGPT helped refine my writing. These tools did not reduce my workload but changed it. They are great for getting a draft going, but always read critically what they produce. Sometimes, they come up with incorrect stuff, so always double-check.

A.4.2 Personal Takeaways and Final Thoughts

This analysis helped put my thesis into perspective. It made me realize that the thesis is just a part of the PhD journey, not the entirety of it. It is a valuable piece of work, but it is not the whole story of what I have learned and experienced.

In sharing this, I hope to set a precedent. I want future PhD students to have a real-life example of what the thesis writing process can look like. It is not just about the final product but also the journey to get there. My advice is to embrace the process, find your support system, use the tools available, and remember to take a step back once in a while to see the bigger picture.

Acknowledgements

I would like to express my deepest gratitude to those who have contributed to my journey through the Ph.D. program and to the completion of this thesis.

First, I extend my thanks to my supervisor, Prof. Luca Iocchi, for his mentorship throughout my time at the Ro.Co.Co. lab.

I am also grateful for the opportunity to work with and learn from Prof. Raquel Fernández during my research visit at the University of Amsterdam. The insights gained during this period, particularly in NLP, significantly shaped a major part of this thesis. Similarly, collaborating with Prof. Daniel Buschek fromt the University of Bayreuth has been a transformative experience, teaching me about German efficiency. In general, the exposure to different academic environments and ways of thinking has significantly contributed to my personal and professional development.

My involvement with CLAIRE[2] A.I.S.B.L has been a significant and enriching part of my Ph.D. journey. I am grateful for the opportunity to grow as a researcher in AI, surrounded by a community that places as much emphasis on ethical values as it does on scientific excellence. This experience has reinforced my belief in the necessity of pursuing AI research that is not only advanced but also ethically grounded and value-driven.

Thank you all for your support, guidance, and encouragement throughout this journey.

[2] Confederation of Laboratories for AI Research in Europe.