Campus des Cézeaux
1 rue de la Chebarde
63170 Aubière, FRANCE

Jens-Chr. Skous vej 7
8000 Aarhus, DENMARK

Internship report

AI and Humanities' love story in interdisciplinary research.

F4: AI and Optimisation.

Student: Maurin Gilles
AU supervisor: Dr. Johanna Seibt
ISIMA supervisor: Dr. Alexandre Guitton

1 April 2025 – 15 August 2025
Defended on 26 August 2025

Aknowledgements

I want to begin the present report by expressing my warmest and most sincere thanks to Johanna Seibt, who made this internship possible by welcoming me in her group, and has been infinitely kind and a great support during my stay.

I also want to thank the other members of the research unit for Robophilosophy and Integrative Social Robotics: Malene Flensborg Damholdt, Marco Nørskov, Christina Vestergaard, and particularly Peter Fazekas who introduced me to the very exciting field of AI explainability.

Another special thank is to be addressed to Arthur Bran Herbener who made me find a home in the psychology department and has been very reassuring to work with during my first weeks in Denmark.

I obvisouly have very kind thoughts for Sofie Klausen and Yoon Frederiksen, my wonderful office mates, who taught me Danish culture, enriched my mind through our deep conversations, and made me enjoy every single day at work.

Thanks also to Alexandre Guitton for his trustful supervision from Clermont-Ferrand.

Finally, I want to express my gratitude to all the people I have met during this exciting experience and helped in their own way to make it life-changing: Mikkel Thøgersen who helped me with my first technical issues; Marie Rosenkrantz Hermann, Anita Hangaard Hoffmann and Emil Hammer Lux who kindly agreed to be my beta-testers; Bolette Windfeld Thesbjerg who invited me to Aalborg University; Mads Hansen with whom it has been pleasure to discuss; Peter Thestrup Waade who kindly introcued me to the department of cognitive science; Julia Cramer who has been one of the most inspiring yet humble people I ever met; and eventually all my friends from the Teknologkollegiet dorm: Narges, Lucas, Bob, Hulunim, Pavlos, Yuka, Burcu, Wei Wei, Laura, Antonia, Hilary, Aidan, Allegra and Tabitha.

Contents

1. Introduction

1.1 Why interdisciplinarity?

My internship which I present in this report comes with certain particularities that make it uncommon for a student from a engineering school in computer science like ISIMA.

One may firstly mention that it is research when engineering schools — as entendend in French educational system — are more industry-oriented. However, these schools are also attractive thanks to the quality of the education they provide, and the research-oriented cursus I attended at ISIMA is an additional clue proving how engineering schools can be good paths to begin a career in research.

Secondly, and in my opinion more importantly, there is the surprising choice of an interdisciplinary research group when ISIMA, despite the multiple specialities it offers, remains inherently focused on computer science and thus related "hard" sciences.

Let's answer this arguing that a specific, highly monodisciplinary background is no obstacle to interdisciplinary research. Specialising in a precise discipline is the best way to aquire an expertise that is more likely to make a real difference in a research environment.

The issue when specialising is to become oblivious to what lies outside our field; and the most dangerous corollary risk is to grow a vanity within our field that could turn into contempt, or more precisely a denial of the expertise brought from other perspectives. I want to aknowledge this risk since French engineering schools promote quality through an elitist environment, and this is one - among others - downside of such a system.

For this reason, the greatest quality one should grow to work in an interdisciplinary environment seems not to be a plurality of expertises, but rather humility regarding the limits of their field. It feels natural to think about what our skills and knowledge can bring to others, perhaps especially in computer science; but listening outside what can be brought within our domain is an equally important appraoch that requires more than plain curiosity.

My experience in higher education has been enough for me to understand that I cannot be satisfied confined inside the strict boundaries of computer science. But would interdisciplinary research suit me was yet to be confirmed. And if it would, it seemed important to me to have an opportunity to start shaping my mind in the light of the risks and qualities I presented above.

I wanted to start my introduction with this topic to present my main goals and expectations for this internship. But the question of interdisiciplinarity is a major one, and it will be addressed again in section 4.2.

1.2 The university and the group

Aarhus is the second largest city in Denmark, and the largest in Jutland, the continental part of the country.

Figure 1.1: Aarhus on a map of Denmark
Source: nationsonline.org

Aarhus University is also the second largest university in the country, with 38.000 students, and is ranked in the top 100 best universities worldwide[a].

The research unit for Robophilosophy and Integrative Social Robotics (RISR) is affiliated to the school of culture and society. The unit is lead by Johanna Seibt, my supervisor.

Robophilosophy is defined as "philosophy of, for, and by social robotics"[1]. Integrative Social Robotics (ISR) relates to the robotics that aim not only to be used in a social environment, but also to perform social interactions with humans [2].

This field must not be mistaken for a laboratory crafting robots for social purposes. Its goal is rather to bring an expertise on the ethic of such robotics, which can been pursued through a great variety of missions such as: Investigate the ontological issues underlying in ISR; Analyse and criticise the state-of-the art innovations in social robotics; Set up reliable theories on what robots can and cannot, but also should and should not do; Conduct empirical experiences to clarify humans' reaction to ISR; Invent tools, frameworks, vocabulary, to make these questions accessible through a scientific and sensible approach, thus facing the risk of moral panic [3].

These goals are at the crossroads between philosophy, psychology, anthropology, computer science and robotics, and all these fields are represented among the members of the RISR group. Interdisciplinarity, in this case, is less a chosen method than a requirement inherited from the very field it studies.

[a] 80 in Shanghai's ranking; 110 according to the Times

1.3 Organisation of the internship

The organisation of the internship was not planned in the long distance since the beginning. The reason to that was the difficulty to make a good estimate on the first project I would work on, which made unclear the time I would have for additional projects. At some point around week 4, it became clear that a second project could be added to my timetable, and it allowed us to make a previsional plan. It should however be noted that this plan was entended to be a helpful guide, not an obligation to be meant at all costs. Flexibility remained the supreme law, and it actually proved useful.

Figure 1.2 shows the plan that has been drafted in the beginning of the internship. It presents the time devoted to the main tasks of the two projects that were to be the core of my internship.

- In green, the design and implementation of a LLM-driven chatbot for psychotherapeutic purposes, that will be thoroughly presented in Chapter 2 of this report, and which I may mention as the **MARC** project since it is the name given to the chatbot.
- In orange, an open research project exploring the concept of intention as a possible clue in the AI alignement problem. This is detailed in Chapter 3 of this report, and may be refered to as the **XAI** project, which is a common acronym for Explainable Artificial Intelligence, the subfield of AI this project belongs to.

Also, even if these two projects represent the most important tasks I had to work on, they are far from covering my whole internship. I was expected to be a full member of the group and to take part of life at university. It lead to a multitude of side tasks and events I attended, that were too punctual to be added on the timetable (except for a seminar that appears because it has had a whole week devoted to it), but will be presented in Chapter 4.

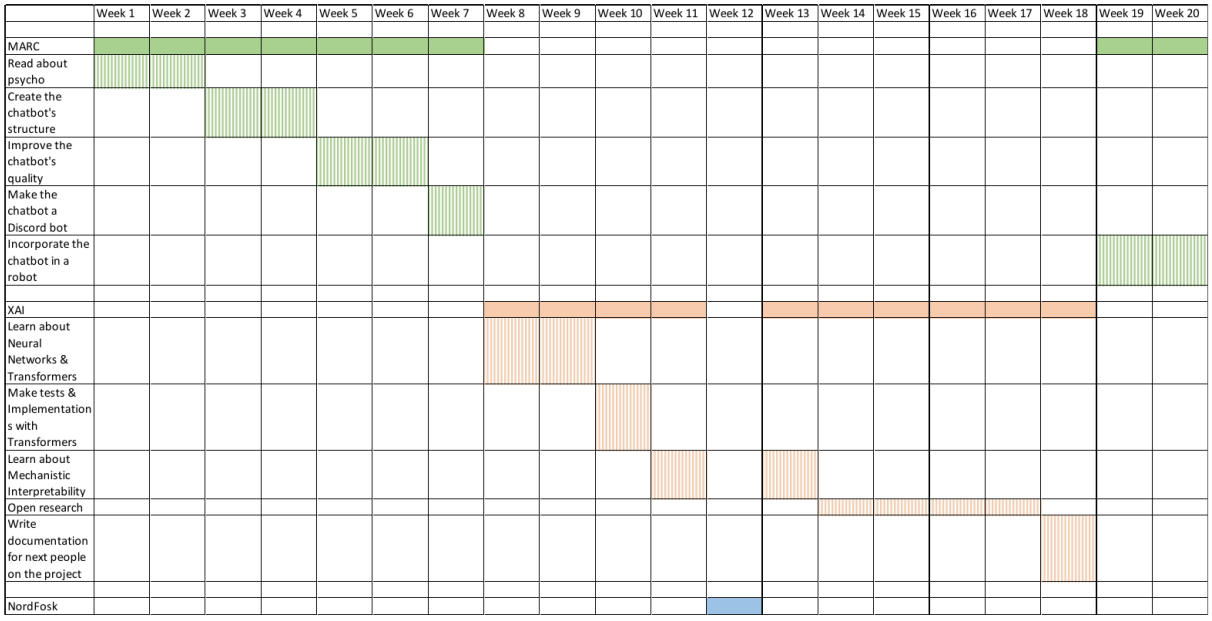

Figure 1.2: Initial plan of the internship

The moment flexibility proved useful was when, during week 16 (14-18 August), it was decided to completely revamp the chatbot of the MARC project, which implied to start almost from scratch again. The reasons for this decision are explained in section 2.4.

The major consequences of this twist were a reduction in the time allocated to the XAI project, and an unexpected rush in the completion of the MARC one.

Figure 1.3 shows the revised agenda, whose changes from the initial one can be noticed by a darker colour.

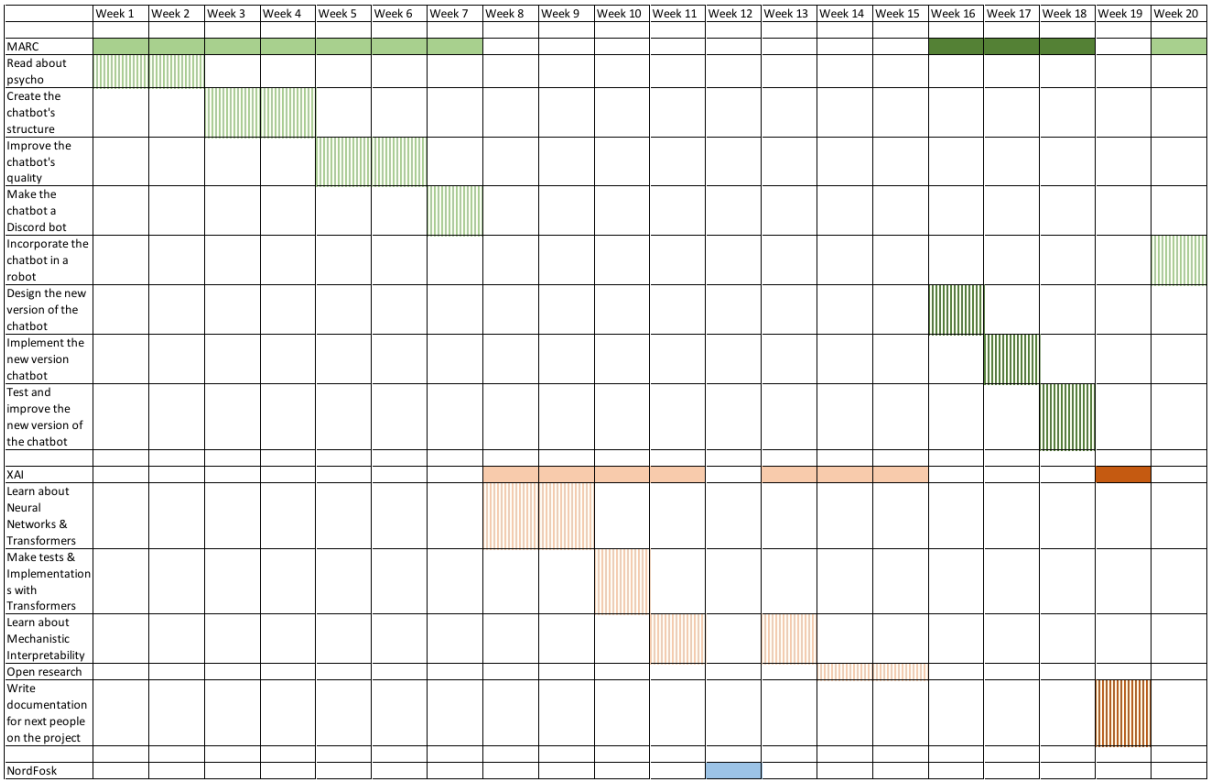

Figure 1.3: Initial plan of the internship

The RISR unit has its offices in the department of Philosophy and History of Ideas. However, I was there mostly for the weekly meetings. Since the MARC project was conducted by a collaborator from the psychology department, I was also offered an office there shared with two people. It proved to be more convenient, and I eventually stayed there, while regularly visiting the RISR unit which was 10 minutes away.

1.4 Corporate Social Responsability

"Corporate" may not be an appropriate term since I made my internship in a university. However, Aarhus University (AU) takes action to make a positive social impact, with two main focuses on sustainabilty and inclusion.

As the regards the first point, AU takes great advantage of being a university, providing both research and education aware of their role in the emergence of a sustainable future. Some departments such as the Environmental Science or the Food Science ones can even be expected to have a direct impact. The administration of the unversity also plays its role towards this goal, and a climate strategy has been set up to reduce the university's

CO2 emissions by 35% against a 2018 baseline by 2025, through the following main areas: campus operations, procurement, transporation and waste [4].

As regards the second point, it should first be reminded that Denmark is a world leader in gender inclusion. It is ranked 5th worlwide for LGBT-friendliness by the Equaldex[b], and 1st worlwide for gender equality[c]. AU also takes its own actions to ensure and promote diversity and gender equality, through a devoted committee [5].

As regards my own environmental impact, I like to think that it has been reduced to the strict minimum. I went to Aarhus and came back by bus, which is estimated as 96.6 kg CO2e by impactco2.fr tools. My stay included two professionnal travels by train, to Aalborg (2.53 kg CO2e), and Malmö (6.06 kg CO2e). Otherwise I travelled and commuted only by bike. I never used heating and shared electrical appliances with 23 people at my dorm.

The last point I want to address here is about research ethics. My contribution to the MARC project is essentially technical, however, the goal in the long run is to conduct a psychological experiment with actual participants. Therefore, following an ethical protocol is of the utmost importance, and in addition to legal regulations from Denmark or the European Union, this experiment will be conducted with absolute respect of the guidelines from the AU Research Ethics Committee.

[b]The Equaldex is the best recommended and referenced index for LGBT-friendliness. This rank comes from the latest issues published in 2025.

[c]The Gender Inequality Index is published by the United Nations Development Programme in their Human Development Reports. This rank comes from the latest issues for this index published in 2023.

2. LLM-driven chatbot for psychotherapeutic purposes

About "MARC"

I will often mention "MARC" in this chapter, but due to nested metonymies, it will not always be referring to exactly the same thing. I do not expect it to be confusing, but starting with a disambiguation remains a good idea.

The higher level and probably the most official MARC denomination refers to the project of a "Multifunctional AutoRegressive Chatbot (MARC...) for AI-delivered mental health interventions". This project is broader than what I worked on during my internship, and I am not aware of everything it covers.

Therefore, I will often call the "MARC" project what should more accurately be called the "sub-project" I worked on, conducted by Arthur Bran Herbener. It consists of a reasearch experiment in psychology requiring the development of a chatbot, and omitting the connections it has with other aspects of the MARC project, it could technically be conducted independently.

Eventually, this project uses a chatbot, and in our everyday communication, we ended up calling the bot "MARC", as if the logic we implemented ourselves was its own personnality. Anthropomorphising the chatbot felt natural, and I will use expressions such as "talking to MARC" instead of "talking to the bot" when describing the experiment.

2.1 Motivating existing work

Prosocial behaviour and well-being

The fundamental idea at the roots of the experiment Arthur will conduct is that increasing our prosocial behaviour has a positive influence on our well-being.

A prosocial behaviour corresponds to acts performed with the goal of benefiting another person [6]. They can be large-scale and meant to improve the quality of life in a community, a society, or even humanity as a whole. They can also, more commonly, refer to acts performed in daily life to do good to specific individuals, be they strangers or family members.

A close concept to distinguish prosociality from is altruism, though some reasearch uses them interchangeably. However, Maalouly et al. [7] pointed out a semantic disentanglement we find worth using. Altruism also refers to acts performed with the goal

of benefeting another person, but with the additional condition that this goal must be motivated by nothing else but the desire to increase this person's welfare. So an altruistic act is always prosocial, but one may aim to do good to someone else with an underlying expectation to get something in return, or just to give a good image of themself knowing it could be useful later. In such a situation, a prosocial behaviour would not be altruistic. Let's also mention that an act can make a positive impact on another person without it being the goal, if this is impact is indirect or a coincidence for example, and thereby not be prosocial.

The other core concept to be defined for a proper explanation of our experiment is well-being. In common parlance - and even between us - we make it simpler using the rough synonym "happiness". However, since the subsequent goal is to observe how well-being can be influenced, the concept has to be measurable, and therefore we cannot limit ourselves to a fuzzy intuition.

Arthur addressed this question in a previous paper [8]. One of the first points to come out is how happiness actually covers only a *hedonic* view on well-being. It means that the subjective experience of positive and pleasant emotions is part of well-being, but perhaps not the only parameter to take into consideration. Arthur mentions *eudaimonic* well-being as a feeling of fullfillment, that may not be as tangibly sensorial as hedonic happiness is, but provides a meaningful impression of self-accomplishment that is definitely worth including when measuring well-being. Some experiments may take advantage of focusing on one of these aspects depending on the hypothesis they want to demonstrate. However, in a case like ours that aims to capture an overall well-being, Arthur shows in the same paper the relevance to take both parameters into consideration.

Still remains to be determined how to make a reliable measure of a person's well-being. The technique generally used is to give participants surveys and ask them to rate themselves on a series of questions with a provided scale; the obvious subjective nature of the exercice is eased with a sufficient number of participants. Multiple surveys exist and have their quality accepted by the scientific community. Nevertheless, every rating system has its pros and cons, and since noe definitive choice has been made for our experience, I will not take the risk to go deeper in handful problematics of a field that is not mine.

These clarifications are important to handle properly the initial hypothesis: "Increasing our prosocial behaviour has a positive influence on our well-being". It may sound intuitive - or not - but far from being a simple intuition from us, it is actually a well-documented phenomenon. An experiment conducted in 2016 by Nelson et al. [6] makes this influence very explicit: they asked a group of participants to perform prosocial acts and followed the evolution of their "psychological flourishing" during 6 weeks, exhibiting a slight yet clear better well-being for the prosocial group against the control one.

One of the limits aknowledged of this experiment is that participants from the prosocial group agreed to perform prosocial acts for the sake of the experiment. The results are encouraging, but do not include the altruistic parameter. This leads to a question yet to be addressed: *is the positive impact of prosociality on well-being enhanced by altruistic motivations?*

Motivating people into prosociality

Answering the above question experimentally assumes a prerequisite that is far from obvious: it is possible to instil altruistic motivations in someone. Fortunately, this has actually been a well-established fact for long. Since it is not to be proven anymore, the goals of research in experimental psychology is now more to explore the practical possibilities it enables. Examples we find interesting include a better blood donor retention [9], or a conservation of helping behaviours in time after traumatic events [10].

A particularly intersting point these examples have in common is how they both use the same technique to enhance people's prosocial behaviour: *Motivational Interviewing*. This is a methodology and a set of principles used in psychocounseling that I will present in detail in the next section. Let's just note how this is exciting for our own experiment: a thoroughly documented technique already proven satisfying in practical cases is a guideline we can rely on to design our experiment, so we do not have to justify further our theorical framework, and rather focus on framing our technical implementation into it.

The major ressource dealing with motivational interviewing is a William R Miller and Stephen Rollnick's book [11], especially for us Chapters 1 & 2 that present the fundamental philosophy of the approach and a detailed explanation of the phase-based technique to use. We use the Motivational Interviewing Treatment Integrity Coding Manual (MITI) [12] as a way to evaluate the performance of our chatbot at matching the guidelines of Motivational Interviewing.

The general idea, in a nutshell, is to adopt an unconditionally positive and non-judgmental approach to set up a real partnership with the participant as equals; and instead of embodying the position of an "advice provider", guide the participant to speak themself into change through active listening, rephrasing and open-ended questions. The MITI rates a performance based on 4 main dimensions (Cultivating Change Talk, Softening Sustain Talk, Partnership, Empathy), and a long list of expected behaviours with useful clues to identifiy them objectively.

In addition to this general guidance, Motivational Interviewing comes with a very important feature: a 4 phases technique. These phases took even greater importance in our experiment, because we used them as a basis to set up the internal structure of our chatbot, which will be explored in section 2.3. Here is a rough description of these phases:

1. **Engaging:**
   Make contact with the participant to build a "partnership" with them; make them feel safe, respected and listend to.

2. **Focusing:**
   Explore the participant's idea to find a clear direction the participant would like to work towards.

3. **Evoking:**
   Guide the participant towards their goal by letting them figure out their inner motivations and reasons for change connected to their own values and experience.

4. **Planning:**
   Decide a plan to make this goal concrete, starting from achievable short-distance tasks based on what has been previsouly determined as the participant's deepest motivations.

The artificial factor

The fact that people can be motivated towards prosocialty through a motivational interview opens a lot of opportunities, even more when keeping in mind that this increased prosocial behaviour may positively influence these people's well-being. However, these opportunites get instantly limited if they require a human psychotherapist to perform the interview. Although experienced professionals of their discipline may be expected to provide significantly better results than any artificial one, their scope of action is extremely limited, and we could benefit a lot from an artificial solution that could be easily distributed widely though providing inferior results.

The quality of the artificial solution is expected to be the big issue, but a preliminary question is to be thought about: has the media an *inherent* influence on how the participant experiences the interview that leads to sensible change in the results? In other words, does the artificial nature or aspect of the agent the participant interacts with have psychological consequences we cannot escape from?

Difficult question, simple answer: yes it does, but less than one would intuitvely think. This is actually one of the major topics of interest of the RISR group at Aarhus University. It is very unclear whether this "artificial factor" is positive or negative, significant or not, absolute or changing depending on the situation, and how this factor may be different for different medias or setups.

We have good reasons to hope that this artificial factor will not have dramatic effects in the case of our experiment. The already mentioned paper from Maalouly et al. [7] deals with an experiment Marco Nørskov from the RISR group contributed to. In this experiment, participants' altruism was measured making them play the dictator game[a] before and after a conversation with a human, an android or a robot depending on the group they belonged to. In this experiment - close in many respects to the one we want to conduct - there was no significant difference between the groups exposed to agents of different natures, and it is an encouraging result. However, our experiment differs in many points, from the way greater autonomy given to the artificial agents in the conversation, to the very goal of the experiment. Therefore, comparing multiple medias remains a crucial point to pay attention to.

2.2 Design of the experiment

Let's recap what insights existing work brings. A prosocial behaviour has a positive influence on well-being. Such prosocial behaviours can be efficiently motivated through a discussion with a psychocounseler. An artificial psychocounseler providing discussions of equivalent quality can be expected to make equivalent impact. Linking all together, our hypothesis is the following: *we can create an artificial agent that will motivate people towards prosocial behaviours thus improving their well-being.*

This hypothesis, to be efficiently explored, should answer these underlying questions:

[a]The dictator game is a money distribution dilemma; see further on [13].

- Is the influence of a prosocial behaviour on well-being enhanced when the participant is motivated towards this behaviour?
- Are some medias better than others to motivate participants towards prosocial behaviours?
- Is a discussion with a participant enough to let them adopt a prosocial behaviour on their own?

All these questions can be answered by comparing groups along two axes: the *expected level of altruism* and the *motivating media*.

The first axis compares a group that is asked to perform a certain amount of prosocial acts without being motivated for it (can be considered as a control group), groups that have been motivated but are still asked to perform a certain amount of prosocial acts, and groups that have been motivated to perform prosocial acts, but are free not to perform any. Comparing the control group to the others will answer the first question, proving if this answer is positive that altruistic intentions have an even better impact on well-being. Comparing the motivated groups between them will answer the third question. We expect most participants not to change their behaviour in the groups where they are not explictely asked to. However, if the conversation alone is enough to make some of them change willingly, their motivations would be expected to be even more "genuinely altruistic", and it would be interesting to see if the positive influence on their well being is even more blatant.

The second axis will answer the second question, comparing groups that will be motivated without interaction (reading papers or watching a video), groups motivated through a discussion with a chatbot, and groups motivated through a discussion with a social robot. Since the purpose of the experiment focuses on artificial agents, the possibility of a group talking to a human agent is not necessarily relevant and would bring logistic complications, so it has been decided not to include it in the experiment.

Table 2.1 summarises the seven groups into which the participants should be divided, depending on whether or not they will be asked to perform a certain amount of prosocial acts (asked to / not asked to), whether or not they will be motivated towards prosociality (not motivated / motivated) and if yes, through which media.

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="3">Altruism</td></tr>
<tr><td>Asked to / not motivated</td><td>Asked to / motivated</td><td>Not asked to / motivated</td></tr>
<tr><td rowspan="3">Media</td><td>Non-interactive</td><td rowspan="3">G1</td><td>G2</td><td>G3</td></tr>
<tr><td>Chatbot</td><td>G4</td><td>G5</td></tr>
<tr><td>Social robot</td><td>G6</td><td>G7</td></tr>
</table>

Table 2.1: Groups for the experiment of the MARC project.

The participants will have their well-being measured as the very first step of the experiment, to have a basis to compare the final results against. Then, for the participants

of the groups where altruism is motivated (G2-7), their impressions at the end of their motivational session will be collected with a survey, to compare immediate impressions to later behaviours. Then, during a week, the participants will have to write down the prosocial acts they will perform if and when they perform any. Eventually, the participants will come back after a week and their well-being will be measured a second time, and this will enable us to draw a conclusion about our hypothesis.

2.3 Implementing the logic

The biggest technical part of the project was the implementation of what I will call the *logic* of MARC. Some participants will text a chatbot and some will talk to a robot, but these will actually be two different interfaces using the same logic in the backend, a programme in Python in our case. This logic has to answer two goals:

1. Guide the participant through a structured conversation. Because we want the artificial agents to be autonomous, they must be able to guide the participants from the beginning to the end. To do so, they must have in "mind" a clear conversation plan, and the ability to follow this plan to match the expectations of the motivational session.

2. Provide smooth and appropriate answers to the participants, to make the conversation feel natural and comfortable to them.

Matching these two goals at the same time is extremely handful: controlling the logic too much restrincts adaptability and the participant may perceive the conversation as too mechanical, but allowing too much freedom to the logic is taking the risk that the conversation may deviate from its purpose and ruin the session.

Programming with LLMs

The whole logic lies on the use of LLMs[b], so before designing a theorical strucutre, it seemed relevant to understand how LLMs can be concretely integrated in a Python programme, to keep that in mind and design the structure accordingly.

One of the great things about Python is that it provides bindings for llama.cpp, a simple and efficient library to program with LLMs. The library provides a rich interface with both high-level and low-level possibilities, however, our use case required only the most basic ones: importing a LLM, and generating answers based on prompts and a message history.

The llama.cpp library provides a "create_chat_completion" funtion which does exactly what we would expect: take a list of messages as input and return a response. Each message is a dictionary of two components: a role and a content. The content is simply the text corresponding to the message itself, and the role is one of these three:

- **System:** used to give instructions to the LLM; to pre-prompt it.
- **User:** used for the messages the LLM answers to, for example the participants' messages in our case.

[b]Large Language Models

- **Assistant:** usually used for the LLM-generated messages, but can also be manipulated to show examples of messages with a specific structure which the LLM will naturally continue to respect[c].

A very important point when programming with LLMs is the choice of a model to use.[d] We have been very lucky on this point, because only a few months before this internship started, Zhang et al. published a model fine-tuned[e] for psychotherapy, along with a paper describing the dataset used to train it [15]. Their model aims to face the often poor quality of LLMs used for psycho-counseling purposes due to the lack of relevant data, mostly due to therapetutic conversations being confidential in general. Therefore, they generated their own preference dataset[f] generating responses with other LLMs, scoring them regarding 7 core principles in psycho-counseling thanks to another LLM trained for this task, and using the best and worst responses among those generated. A sample of this dataset has been verified by experts, making the whole reliable.

One of the convenient things about this model is that it is based on LLama-3 8B[g], and could therefore run on my computer. We compared it to other models of equivalent size, and it proved its quality in practice. We collected impressions from four students and professors in psychology, all agreeing on Zhang et al.'s model expressing better empathy and asking better questions to the user.

Of course, the model still suffers from the drawbacks of small models, and sometimes sounds a bit "off-topic" and tends to repeat itself. If a larger model solved these issues, we would prefer it. However, my computer cannot run models over 13 billion parameters, and such models were not better enough for their size to weigh more than the Zhang et al.'s fine-tuning in our opinion.

A phase-based structure

We have decided to get inspired from the 4 phases of Motivational Interviewing described in section 2.1 to frame the conversation between MARC and the participants. Since these conversations are expected to be rather long, dividing them into shorter sections requires a smaller context window[h] for the LLM, which is therefore less likely to deviate.

One difficulty with this approach is to determine when to switch from a phase to the next one. When Motivational Interviewing is used in practice, the phases do not have to follow a strict linear order, and practitioners can trust their impressions and go back to a previous phase it appears necessary. In our case on the other hand, a linear structure is a great solution to keep control over the flow of the conversation, and the risk of having the chatbot jumping to a phase too quickly is preferable, in our view, to that of having the chatbot never moving forward.

[c] This is a fundamental technique when using LLMs called few-shot prompting, introduced by a reasearch team from OpenAI in 2020 [14]

[d] Multiple models exist and differ in features such as the architecture, the number of parameters or the data it has been trained on.

[e] Fine-tuning consists in taking a "base" model and train it on additional data to make it better for a specific use.

[f] A dataset made up of triplets "question"/"good answer"/"bad answer".

[g] Llama-3 is an open-source model created by Meta. 8B stands for 8 billion parameters.

[h] The context window is the amount of previous tokens the LLM relies on to compute the next one.

Therefore, we redefined each phase of Motivational Interviewing to adapt them to the very change we want to guide the participants towards (more prosocial behaviours); and give each of them a precise goal to reach that would trigger the jump to the next one. This is summarized in table 2.2.

Phase	Adapted definition	Goal
Engaging	Settle a safe place and initiate the partnership with the participant; ask them about their views and experiences on prosocial concepts.	Gathering explicit material about the participant to link to prosocial behaviours.
Focusing	Dig into the participant's values and vision of prosociality to identify a precise prosocial goal that corresponds to them.	Identifying a specific goal to work towards.
Evoking	Motivate this goal by letting the participant explore what it means to them, how it corresponds to their values and how they would feel if they achieved it.	Making the participant's motivation clear, easing their doubts.
Planning	Help the participant set up a conrete plan to make this goal a reality.	Determining a schedule and the first steps to take.

Table 2.2: Phases for MARC

Inside MARC's mind

This section will describe the concrete implementation that allowed MARC to autonomously have conversations following the structure detailed above. Figuratively, we tend to call that "MARC's mind". It is composed of three units:

- The **Main Unit** concretely handles the conversation with the participant. It receives the new messages from them and add them to a list of messages including a custom prompt and the message history of the phase. It calls the chat completion function witth this list of messages, and returns the response to the participant.

  The prompt plays a crucial role and is made out of four components:

  1. A global prompt, that explains the psychocounseling role to play, the experiment, and the general guidance of Motivational Interviewing.
  2. A phase-specific prompt, to describe the goals of the current phase. The point of the other phases are ignored, hopefully helping the LLM to stay focus.
  3. A feedback from the Controller, that is generated after each message and is asked to be taken into account with the utmost importance. This allows a

smooth, adaptive control of the flow of the conversation. More about how this feedback is generated is presented in the part about the Controller unit.

4. For phases 2 to 4, a summary of what has been told in the previous phases. This allows MARC to remain consistent through the phases without keeping "in mind" the message history of the entire session.

- The **Controller** is the unit that generates the part of the prompt that changes for each message. It is expected to be good at understanding the meaning and the goal of each phase to tell how a conversation is going regarding these goals. To this end, it has a prompt more thorough than the Main Unit's phase-specifc one about the current phase. It also gets the conversation between MARC and the participant, and formats it into a single string. Then, it few-shot trains itself thanks to examples of conversations in the same format, and feedbacks adapted to these examples that we wrote ourselves. By this mean, the chat completion provides a feedback that meets our expectations.

  The Controller returns the feedback to the Main Unit, along with a decision on whether or not the conversation should jumpt to the next phase.

- The **Summariser** is called at the end of each phase to make a summary that will be included in the prompt of the next phase. Therefore, there are two points this summary must respect: pay a special attention to what is useful to the goals of the next phases, and fit nicely into its place in the prompt.

  We expected these points to be big issues requiring for the Summariser to include a few-shot training too. However, it appeared that a good prompt and the conversation provided in a single string format were enough to generate the kind of summaries we wanted.

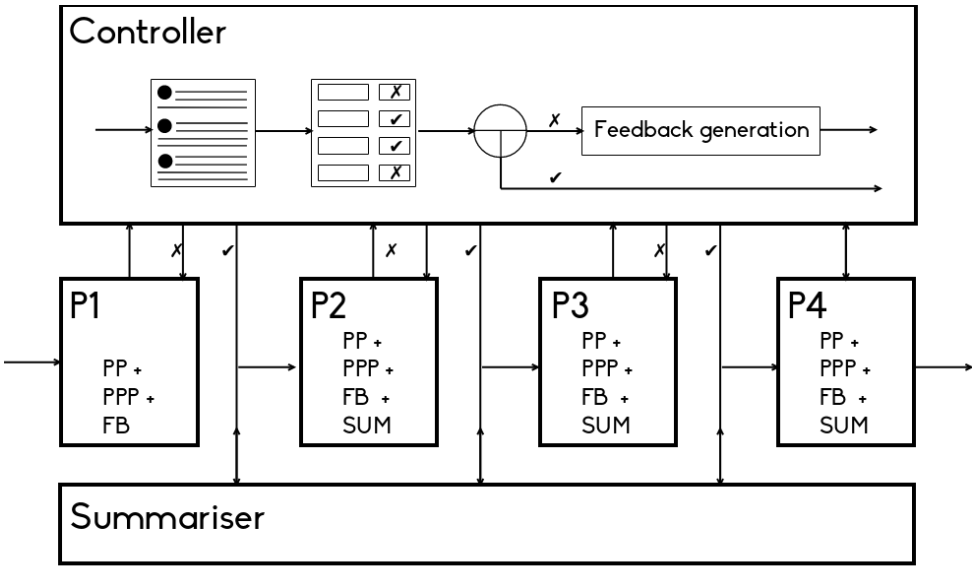

Figure 2.1: MARC's logic structure
PP: PrePrompt / PPP: Phase PrePrompt
FB: FeedBack / SUM: Summary

Training of the Controller

Making a programme reproducing the design of MARC's logic and writing good prompts were big parts of the implementation process, but it was not much of a problem. The Controller, however, required a lot of work due to its integrated few-shot prompting. The goal of this few-shot prompting, in our case, was not only to make the Controller's outputs match a specific format, but also to give good examples of what is good and what is not during a phase. In order to cover a large pan of scenarios, 5 examples for each phase appeared to be necessary, but we did not go further since it is usually recommended not to overwhelm the context window with additional examples that do not improve much the quality of the answers [16].

For these examples to be relevant, we needed the conversations to look like what could happen in a real use case during the experiment. To do so, we made these conversations using a version of MARC without controller. These examples were not perfect, especially because the absence of controller really reduces the quality of MARC's answers. But at least these examples could be used to create a first version of a controller for MARC, and use it to generate better examples to include in a final version of the Controller. This process is illustrated on figure 2.2.

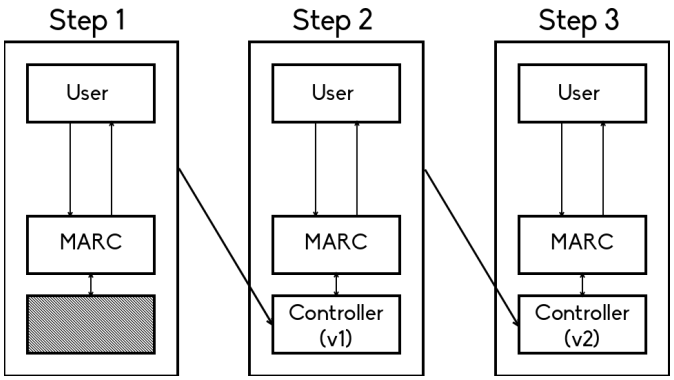

Figure 2.2: Controller training

2.4 All is to be redefined

Issues with the first structure

The structure of MARC's "mind" presented in the above section brought a bunch of issues, and most of them came from the training process of the Controller.

The first one was the number of conversations that had to be generated. Each version of the Controller requires 5 examples of conversations for each of the 4 phases. Because later phases require to go through the earlier phases first, the Controller has to be done for phases 1 to N to before samples of conversations for phase N+1 can be generated. It is therefore impossible to have full conversations to divide into examples for each phase: examples must be generated for phase 1, then 2, then 3, then 4. So each version of the Controller requires to have 20 conversations, 40 in total, each lasting around 30 minutes on average.

The problem was not the time this process takes, but the fact that I was alone generating these examples. And what I did not see coming, is how mentally exhausted it becomes to have the "same" conversattion over and over all a week long. Of course the conversations were never really the same, because I embodied different characters and stories, but the LLM tended to express itself in a rather monotonous way, and most importantly, the topic of the conversation was always the same. Also, the time the LLM takes to generate answers is not an issue when one has a single conversation with it, but for me it eventually became very annoying.

Some student assistants came to talk to MARC and generate new conversation samples for me, but I still had to do most of it myself, and the fact that I got bored and frustrated talking to MARC has repercussions that unfortunately impacted the examples to provide to the final version of the Controller.

The fact that I was writing participants' messages myself in most examples proved to be a huge issue for another reason: they look "similar" to some extent. Even if I did my best to invent different stories, personnalities and ways to express myself, I remained unconsciously influenced by a lot of things that have a huge impact on the conversations: my way of thinking could be completely changed, I still expressed myself in a special and non-native way, I was playing roles so my answers were not genuine, and very importantly, I had a perfect knowledge of MARC's mind and always knew what to say to bring the conversation where it was expected to go.

I underestimated the importance of these parameters, and despite my attemps to fake other personnalities, the few conversations some other people had with MARC made very clear that they could bring it to directions my examples did not cover.

This led to the corollary major issue: deviation. The motivational session is made to follow a precise linear structure, and one of MARC's tasks is to keep the conversation into this structure. However, the conversations deal with participants' experiences, values, social life, and these topics are likely to lead participants to talk about personal matters. When this occurs, the expected reaction is to build on this personnal matter towards prosociality; but this should be done very smoothly so that the participants do not feel like the personal matter they confided is being ignored. This theorical reaction is however often difficult to put into practice, and when a participant does not help perfectly in seeking prosocial behaviours, MARC tends to abandon the goal of the motivational session to follow the participant with their personal matter.

On the other hand, new issues appear when trying to fix that by making the goal of the experiment MARC's absolute priority. Typically, MARC tends to appear less empathetic, and sometimes even unpleasantly pushy when a participant is not convinced aout increasing their prosocial behaviour.

The plot twist

Although a version of the chatbot was technically available, the major issues presented above were still pending when I reached the 8th week of my internship and started working on another project, in accordance with the schedule presented in the introduction 1.3. I was trying not to focus on that, though I was still searching for ideas to solve them; Arthur at the same time was a visinting researcher in Zurich. There, he met Dr. Rachel

Baumsteiger who had previously worked on an intervention for promoting prosocial behaviours [17]. With her agreement, Arthur mentioned the idea to use her intervention as a basis for MARC's motivational sessions.

The obvious drawback using this intervention plan for our experience was the necessity to implement the new logic from scratch, along with a frustrating feeling to have developped a first version almost entirely "for nothing". Nevertheless, it came out quickly from our conversations that this "full rebranding" was the right choice to make in our situtation, because this opportunity brought several advantages:

- Baumsteiger's intervention plan consists in a series of activites to do with the participant. However, although it is recommended to do these activities in a certain order, they are completely independent from each other. This solves the problems of managing the session as a whole, which was only partially faced with a phase-based approach in which later phases still depend on the previous ones.
- The activities can be very easily prompted as simple tasks with clearly defined goals, which leads to a more predictable bahviour from MARC. It doesn't need adapative feedbacks to stay aligned, that makes this role of the Controller useless.
- The simplicity of the activities and MARC's predictable behaviour make all conversations structured quite alike, and the number of messages required to reach the goal of an activity is more or less always the same. This enables us to control the end of an exercice with the number of messages rather than a controller. It could lead to some rare sessions being not perfectly smooth, but these punctual cases would happen with any solution. At the same time, this approach makes the other role of the Controller pointless too. Getting rid of the Controller means to save one call to the LLM per message, which reduces greatly the time for the participants to get MARC's responses.
- The simplicity of the activities tends to make both MARC and the participants' messages shorter. Also, MARC's questions stay less deep and personal than what they tended with the first logic. For these two reasons, a revised logic was expected to fit way better into the social robot, which we feared to perform poorly with the first logic.

The new structure

Figure 2.3 presents the structure of the new version of MARC, which is way simpler than the initial one. The conversation simply goes through four activities, each controlled by a single prompt and a fixed number of messages.

Once this number of messages is reached, a unit we call "Concluder" is called. Its role is just to smoothly put an end to the conversation, and summarise briefly what has been said during the activity. As for the Summariser in the initial logic, a simple prompt and the conversation given in a single string format are enough for the concluder to meet our expectations.

After a message from the Concluder, the participant says when they are ready for the next activity to start, and it does.

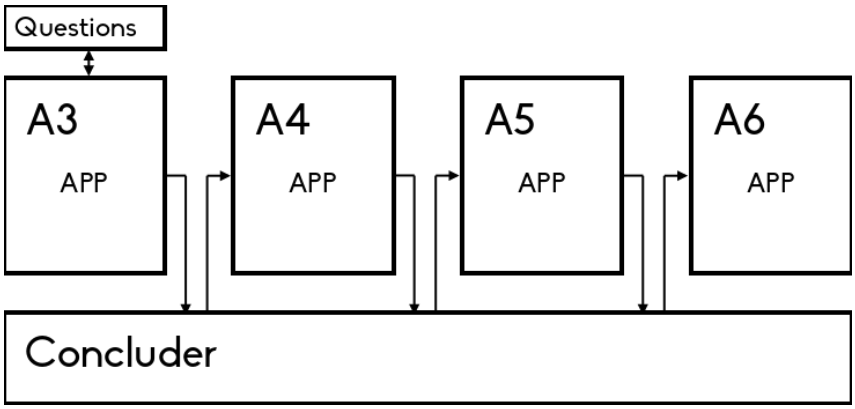

Figure 2.3: MARC's logic structure
APP: Activity PrePrompt

Baumsteiger's intervention plan includes 6 activities, but the first one is an introduction and the second one a video to watch. The remaining activities are integrated into MARC, sometimes with small changes to make them fit into a conversational format:

- **Activity 3**
  The participant has a list of 8 questions to answer intuitively without elaborating. In our case, we only pick randomly 5 questions so that the activity is the same duration as the others. Also, the questions are given to the LLM one by one using the "system" role of the llama.cpp library[i]. Then the LLM generates its own message to ask the question to the participant, sometimes rephrasing it, and usually reacting to their previous message first to create a sort of transition.

- **Activity 4**
  The participant is asked to identify 3 to 6 core values, qualities or principles that are the most important to them, and explain why they have chosen them. The activity comes with a list of example values though the participant is free to find their own. In practice, MARC sometimes go quite deep into why some values are important to the participant, and since the end of the exercice is triggered by a number of messages to reach, the number of core values explored is a little random; or at least it is not chosen by the participant. From our experience, it most often identifies 3 values, sometimes 2 or 4, but barely never less or more.

- **Activity 5**
  The participant need to imagine their life 5 years ahead if everything goes the best way possible. The points they have to mention include what they woud be like, who they would be with, what they would be doing, and what impact they would make around them. There was nothing special to do for the implementation: a well-prompted LLM guides the participant through the activity in a satisfying well.

- **Activity 6**
  The participant must think of a way to make a good impact on people around them in the coming week, detailing what action to take, how to do it, what exactly would be their impact on others, and how they would react to it. The original activity includes a table in which the participant writes down about the actions they actually

[i] See 2.3, "Programming with LLMs"

take during the week; but in our case the motivational session will not go further than the planning part, because keeping track of the prosocial actions performed is already a part of the experiment.

2.5 Interfaces and robot

The logic handling conversations between MARC and participants is the core of the project, so the natural thing to do was to focus on it first, and only then think about the chatbot's GUI[j] and the embodiment into a robot. Following this strategy, my computer's terminal was used for inputs and ouputs during most of the project's development. When the first version of the logic presented in 2.3 was roughly available and before the issues presented in 2.4 made us change everything, I worked on a GUI for the chatbot. Weeks later, when the new logic was implemented, I worked on plugging this logic into a robot.

Chatbot with Tkinter

I had already used Tkinter before, and creating a basic chatbot interface was not a problem. However, I faced an issue that I should had seen coming: the way I had programmed the logic was very bad from a software architecture perspective. The "natural" way to handle a chatbot in Tkinter - and with most user interface solutions - is to separate the logic from the interface management (the backend and the frontend), so the user interface would just call the encapsulated logic from a sandbox when necessary.

On the other hand, I had programmed the logic in a "lazy", linear, intuitive way, blocking the programme when asking for messages from the user with the Python "input" function. I thus had to refactor the whole programme to make the logic accessible via a "get_next_message" function and a bunch of parameters. Even if I could have found a way to bypass this with Tkinter, it would have had to do be done for the robot version anyway.

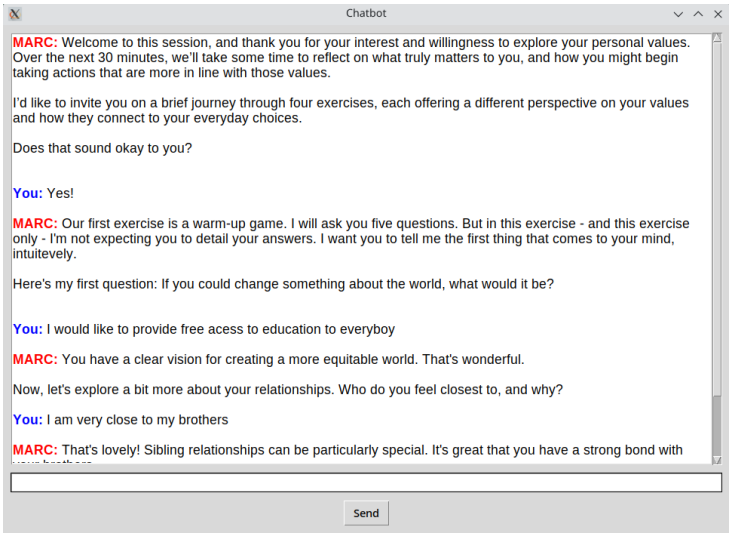

Figure 2.4: Tkinter GUI
(MARC's revised logic)

[j] Graphical User Interface

Discord bot

Making a Discord bot was a bigger technical challenge for me given that I had never done it before. It however proved not to be very difficult, especially since I had refactored my programme in way that reduced the requirements for the bot to very basic features. Basically, the "software" architecture (software may be a big word to describe the programming part of the project) was divided into three files with very clear roles:

- **"marc_logic.py":** encapsulate all the logic dealing with the LLM and provide a "get_next_message" function.
- **"marc_server.py":** run a Flask server providing access the logic to external applications. It also has a very simple session manager, to save all the parameters required by the "get_next_message" function[kl], and a reset method that simply cleans the session.
- **"marc_discord.py":** manages the Discord bot using the Python Discord library. It mainly consists in an event listener on the user's new messages that sends requests to the Flask server to get MARC's responses. It also enables the user to reset the session typing "!reset".

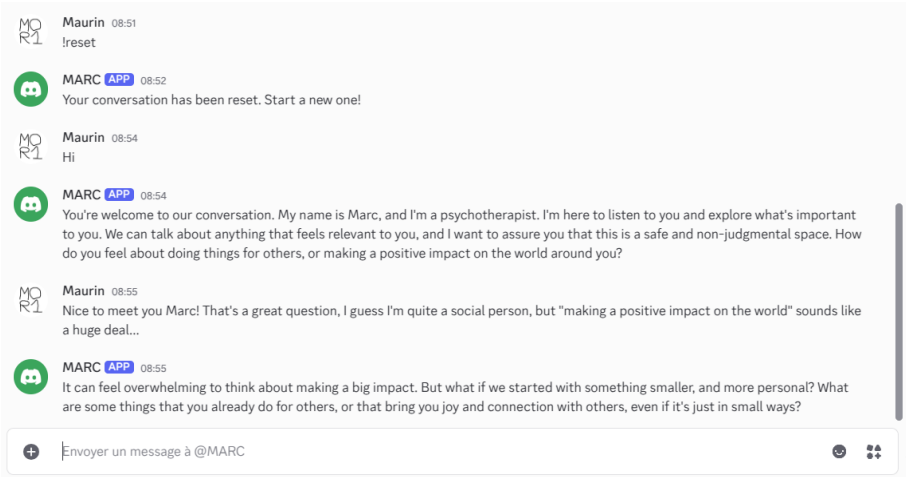

Figure 2.5: Discord bot
(MARC's initial logic)

Furhat robot

Furhat robots are social robots developped by the Swedish company Furhat Robotics. Their main purpose is to have conversations with users, thanks to state-of-the art speech

[k]For the initial logic, these parameters are the number of the current phase, the message history of the current phase, the summary of the previous phases, the number of messages sent during the current phase, and a boolean to detect the first message.

[l]For the revised logic, these parameters are the number of the current activity, the message history of the current activity, the number of messages sent during the current activity, a boolean to detect the first message of the activity, the list of the chosen questions (for activity 3), and the index of the current question (for activity 3)

detection and facial analysis, a voice synthesis module including tonal variation and other prososdic features, and the ability to mimic a lot of gestures and facial expressions.

Furhat robots can be programmed thanks to *skills*: programmes in Kotlin that are complied into .skill files that the robots understand. Fortunately, the robots come with a thorough documentation [18] and a SDK[m] that helps creating skills from models, and allows to test them on a virtual version of the robot.

Learning the basics of Kotlin and understanding the Furhat library remained a technical challenge, but calling the Flask server developed for the Discord bot to compute answers to users' messages made the requirements for the robot-side programme very simple. It is mainly an even listener detecting what the users say, makes the robot nod to confirm it has heard, and calls the server to get the response to say.

Figure 2.6: Furhat robot

[m]Software Development Kit

3. Intention-oriented perspectives for AI alignment

Compared to the MARC project thoroughly explained in Chapter 2, this one is way more theorical, and therefore more complicated to present. Particularly, the project lies on important Deep Learning groundings that took me weeks to learn[a], but do not seem relevant to me to include in a report that is meant to be a description of my work rather than a thorough lesson on AI.

I will try to present the main ideas and tools that lead to this research project in a simple intuitive way, but for that some fundamental notions about neural networks and a few proper mathematical definitions cannot be avoided.

I will first introduce the key concepts of Neural Networks in section 3.1, and present the two structures that are important for what comes next. In section 3.2, I will present the main ideas of Mechanistic Interpretability, which is the real starting point of the project. Eventually, section 3.3 is about our first intuitions for the open research project that is still in its earliest stages of definition.

3.1 About Neural Networks

Key concepts in Deep Learning

Machine Learning (ML) is the branch of AI in which machines train themselves to perform some specific tasks learning from data thanks to certain statistical algorithms. *Deep Learning* (DL) is the sub-branch of ML using *Neural Network* architectures.

The smallest components in a neural network are the *neurons*. The usual neuron structure consists in many inputs multiplied by weights and summed together, a bias added (not represented in figure 3.1), and an activation function which is applied to the whole and gives the output. Mathematically, we have the output $o = \varphi(\sum_{i=1}^{n} x_i w_i + b)$ with x_i the inputs, b the bias and φ the activation function. Figure 3.1 is an illustration of an artificial neuron.

[a] At ISIMA Deep Learning is taught in final year so I had no background on the topic, though the classes on AI I already attended were great help in my self-learning process.

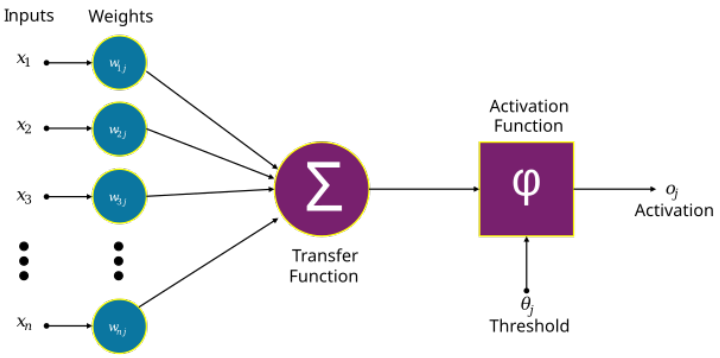

Figure 3.1: Artificial neuron
Author: Funcs
Source: wikipedia.org

Usually, neurons are grouped together in *layers*. In the most intuitive case, a layer is a groupe of neurons such that all the neurons of the layer get their inputs from the same previous layer, but multiplied by a different weights vector, and have their outputs used by the next layer. To be precise, this is the case for what are called *hidden layers*: the first layer of the network (*input layer*) takes directly the actual data to process as input; and the last layer of the network (*output layer*) produces the final output.

MLPs[b] were the first neural networks to be used, and they follow this elementary architecture. It is the intuitive representation one makes of neural networks, as they are illustrated in figure 3.2.

Mathematically, the output of a layer k in a MLP is $x^{(k)} = \varphi(W^{k^T} x^{(k-1)} + b^{(k)}$, where w is the matrix so that W_{ij} is the weight from neuron $x_i^{(k-1)}$ to neuron $x_j^{(k)}$, $b^{(k)}$) the biases of layer k, and φ the component-wise extension of the activation function. F

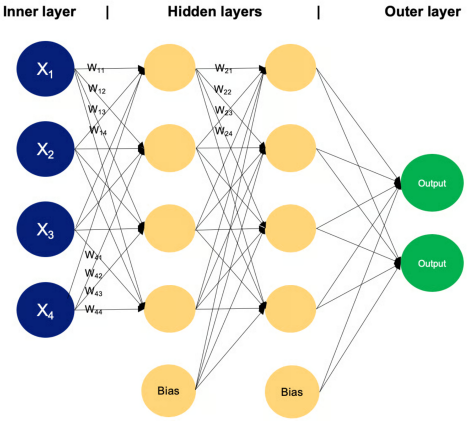

Figure 3.2: MLP Neural Network
Author: Sejal Jaiswal
Source: datacamp.com

Much more could be said about MLPs, especially regarding the training process, but this is not essential to understand the following of this chapter. The two key notions that absolutely had to be introduced are neurons and layers.

[b]MultiLayer Perceptron

CNN and Transformers

Mechanistic Interpretability is mainly studied on two neural network architectures: Convolutional Neural Networks (CNN), and Transformers. I will present here the main intuitions to have to understand them.

Convolutional Neural Networks

CNNs are typically used for image classification tasks[c], as their efficiency in this case have been proven very successful by Alex Krizhevsky, Ilya Sutskever and Geoffrey E Hinton in [19], making the architecture really popular. The first key differences to note compared to MLPs are that CNNs use different kinds of layers, and handle not matrices but tensors[d].

The fundamental type of layers in CNNs, that gave its name to the architecture, are *convolutional layers*. They consist in a bunch of *filters* (or *kernels*): tensors of smaller size that are used to perform a "convolution"[e] over the input.[f] When dealing with images, filters can be represented as smaller images slided over the input image to detect some features. The result of a filter applied to the input is a *feature map* that can be thought of as an image exhibiting where the input reacts to the filter (in the common use case).

A convolutional layer consists in a whole bunch of filters each producing a feature map, hence the 4D tensor as an output.

Mathematically, we can define our "convolution" operator (cross-corelation, actually) as:

$$\begin{aligned} Y &= X * F \\ &= \sum_{u=0}^{k_H-1} \sum_{v=0}^{k_W-1} \sum_{d=0}^{C_{in}-1} X_{i+u,j+v,d} \cdot F_{u,v,d,c} \end{aligned}$$

Where for an input height H, an input width W, a filter height f_h, a filter width f_w, a number of input channels C_{in}[g] and a number of output channels C_{out}, we have $X \in \mathbb{R}^{H \times W \times C_{in}}$ the input (initial or from the previous layer), $F \in \mathbb{R}^{f_h \times f_w \times C_{in} \times C_{out}}$ the filters, and $Y \in \mathbb{R}^{H \times W \times C_{out} \times C_{out}}$ the output.
Then a convolutional layer is:

$$Y = \varphi(X * K + b)$$

With φ a non-linear function and b the bias.

The other type of layers to be found in CNNs are *pooling layers*. They have a "down-sampling" role, which is to reduce the size of the feature maps. They simply apply a

[c] The neural network has a list of possible labels and must choose the one that corresponds the best to the input image, usually giving each possible label a probability.

[d] Images are usually 3D tensors, and hidden layers in CNNs handle multiple such tensors, resulting in 4D tensors.

[e] This terminology is not mathematically rigorous, but often used when dealing with CNNs.

[f] Also, in addition to the "convolution", a non-linear function that could be compared to the activation function in MLPs is applied.

[g] 3 if the input is RGB image, and then multiplied by the number of filters applied by the convolution layers.

chosen function, and therefore do not require any parameter to train. They are to some extent "neuronless" layers.

So basically, a CNN uses these types of layers to detect some properties in an image, that get more complex and precise. For an image classification task, it allows the neural network to truly "understand" the image, before a more "traditional" output layer associates it to labels. A typical CNN architecture[h] can look like figure 3.3.

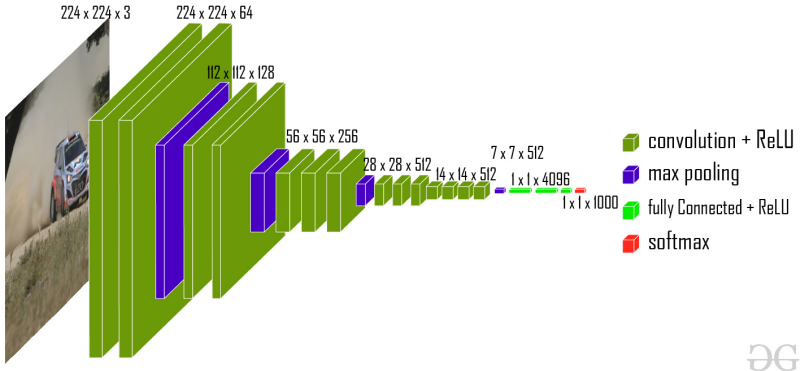

Figure 3.3: VGG-16 CNN Architecture
Source: geeksforgeeks.org

Transformers

The transformer architecture has been introduced in 2017 by a reasearch team from Google in [21] on translation tasks, and has led to OpenAI's breakthrough when used for Natural Language Processing (NLP) as they introduced GPT-1 in [22]. State-of-the-art transformers tend to have complex structures, but here we are only interested in the general idea of a simple structure such as the one of figure 3.4.

If CNNs are commonly represented taking images as input, transformers are usually represented handling textual inputs. Therefore, we call *embedding* the process of cutting a text into small pieces called *tokens*, which are then changed into vectors, making the input text a matrix.

A major particularity of transformers is the presence of a *residual stream*. Basically, in a transformer architecture, the role of the layers is not to modify the information to pass to the next layer, but to compute pieces of information to add to the residual stream, which accuma*ulates more and more information through the layers.

The type of layers at the heart of the transformer architecture are *attention* layers. Their role is to pass information from tokens to others, which a single attention head does thanks to three weight matrices called respectively Query (Q), Key (K) and Value (V). The intuition to have is that queries are questions each token asks, keys are the degree of relevance of other tokens to these questions, and values the information to transfer in such cases.

[h]VGG-16 is CNN architecture introduced by Simonyan and Zisserman in [20].

The common compact expression is:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Where Q, K and V are the matrices previously mentioned, d_k the dimension of the keys and queries, and $softmax$ a function that converts each row of the matrix into a probability distribution.

In practice, we deal with *multi-head attention* layers, that can with be considered with simplification as multiple such processes computed at once.

Eventually, in transformers, attention layers alternate with MLP layers which themselves usually have 1 hidden layer. As regards the point of having thse two types of layers, the intution usually given is that attention draws the connections between tokens, and MLPs process these connections. In other words, attention tells "what to look at", and MLPs "how to use this information".

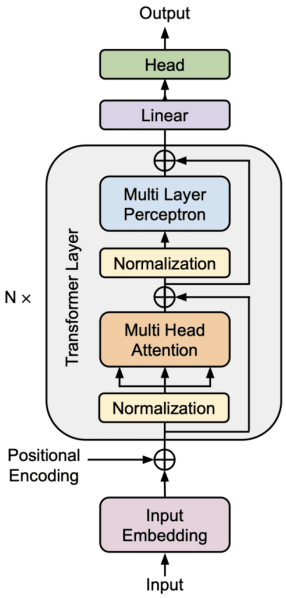

Figure 3.4: Simplified Transformer Architecture
Authors: Gesmundo and Maile
Source: See [23]

3.2 Mechanistic Interpretability

When talking about AI, the goal of interpretability sounds natural. However, as presented by Lipton in [24], there is a vagueness surrounding this concept that is too rarely clarified.

Firstly, Lipton points out how different goals that can be called "interpretabilty" refer to different properties of a model, and can sometimes even be contradictory. He mentions 5 common desiderata of interpretability:

1. **Trust:** we have evidence that the model never does worst than a human on a specific task, so the task can be left to the model with high confidence.

2. **Causality:** the way the model learns from the data is causal and thereby brings out insights on this data.

3. **Transferability:** we understand how the model generalises from the training data, so we can protect it from attempts of manipulating it with unusual data.

4. **Informativeness:** the way the model processes data and constructs outputs is rich in information.

5. **Fair and ethical decision-making:** we can identify the biases the model may be subject to, and correct them.

Lipton then mentions the two main approaches in interpretability, that we can define in relation to the issue of AI models being "black boxes"[i]. The first approach is to seek *transparency*, by creating models that are inherently interpretable. On the other hand, *post-hoc interpretability* lets models have a black box, and then tries to understand what is happening inside. Mechanistic Interpretability belongs to this second category.

Main idea

Mechanistic Interpretability (MI), in the words of one of its most important contributors Chris Olah, is the process of reverse engineering neural networks [25]. The assumption is made that all neurons in the model serve a certain purpose, and that understanding what role they have and how they interact with each other can help better understanding the whole system.

The hypotheses behind MI can apply to all types of neural networks, but CNNs processing images are generally used for introduction because they provide very simple ideas both about *what* to visualise (filters/feature maps) and *how* to visualise them (an image format). The common process to this extent is explained by Olah, Mordvintsev and Schubert in [26] and is called *optimisation*. It consists, basically, in constructing an image that would trigger the neuron to the highest utmost. It can provide blatant (and very intersting) results such as those shown in figure 3.5.

[i]The black box is a common image to illustrate the fact that neural networks tune their parameters in an autonomous training process. So humans understand inputs and outupts, but not the computation between: as if it was hidden inside a black box.

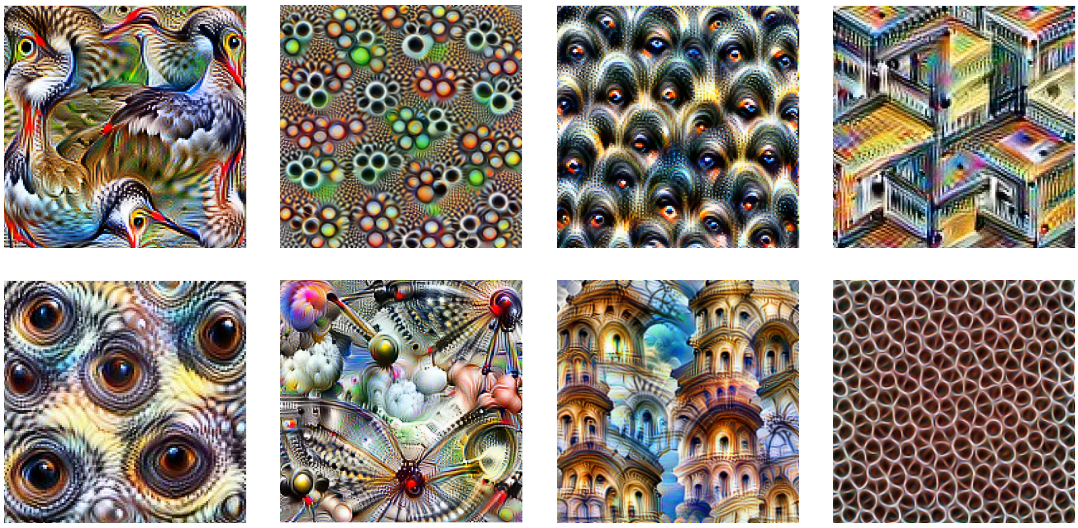

Figure 3.5: Examples of neurons from GoogLeNet
Authors: Olah, Mordvintsev and Schubert
Source: See [26]

Although visualising neurons from a model is satisfying, the interest for the field in general remains limited. However, there is a strong hope in MI that some patterns transcend individual models, and appear like inherent properties of neural networks' "way of thinking".

A first towards this hope is the great amount of occurences of a phenomenon called *equivariance* within a model, which consists in multiple neurons detecting the same feature with a shift in size, orientation or hue for example [27].

The second, even more encouraging clue is that some features already show some signs of universality. It has been shown for neurons such as curve detectors or high-low frequency detectors in the "claim 3" part of [28].

This hypothesis is very important in MI, because if true, it would mean that a deep study of some systems could help understand other systems. Therefore, MI wouldn't be the understanding of the low-level mechanisms of one model, but the understanding of neural networks' thinking in general.

Superposition

The way I presented things so far assumes that each neuron has one specific role, but it actually happens to be more complicated. A central observation in MI is that neurons tend to contribute to multiple tasks simultaneously - a phenomenon called *superposition*.

The accepted explanation of why this happens is that AI models try to reprensent more *features* than they have neurons. The concept of feature is quite fuzzy but conceptually easy to get: it is a small piece of information the model uses. In image processing, features can detect things like curves or borders in the early layers, and things like dogs or cars in the later ones. In language processing, features can detect objective properties such as a the text being French or Python code; abstract properties such as the tone being familar or the text dealing with someone expressing doubts; or concepts such as the text

talking about Copenhagen or being verses from the Bible.

Superposition conceptually makes a lot of sense: such features are *sparse* in the sense that they are relevant only in very specific cases, and useless for a vast majority of inputs. So having neurons associated to one feature exactly would be a huge waste of ressources given the limited number of neurons.

Elhage et al. explored superposition in detail in [29] and proved empirically that superposition happens in an expected way: when features are dense, there is few superposition and only the most important features are represented, because the risk of a feature interfering with another important one is too high; but as sparsity increases more features can be represented. Figure 3.6 illustrates this idea.

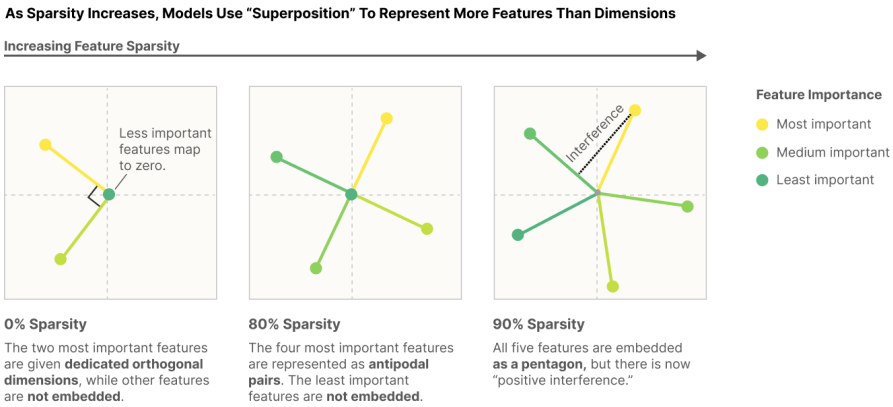

Figure 3.6: Superposition and sparsity
Authors: Elhage et al.
Source: See [29]

Superposition is not a big issue when dealing with CNNs: the amount of features required for an image classification task is expected to be nearly accessible, and even when superposition happens like it does on the famous example of figure 3.7, it does not necessarily prevent the interpretability of the neuron.

Figure 3.7: Example of a polysemantic neuron from GoogLeNet
Authors: Olah, Mordvintsev and Schubert
Source: See [26]

On the other hand, superposition becomes a major issue when applying MI to transformers processing natural language.

Sparse AutoEncoders to interpret Transformers' features

Let's formulate the superposition hypothesis like this: neurons' dense activations correspond to a latent representation of sparse features forming an overcomplete basis in higher dimension. This formulation naturally leads to the desire to find this sparse representation, which is expected to be interpretable. This is actually a typical *dictionary learning* problem, and in the case of MI, it is solved with a fundamental tool for the field: *Sparse AutoEncoders* (*SAEs*).

The goal of an autoencoder is to find the pair of encoder f and decoder $\hat{x}$ such that

$$\begin{aligned} &\underset{\mathcal{X}\to\mathcal{Z}}{f} : x \to \sigma(W_{enc}x + b_{enc}), \quad \sigma \text{ a non-linear function} \\ &\underset{\mathcal{Z}\to\mathcal{X}}{\hat{x}} : f \to W_{dec}f + b_{dec} \end{aligned}$$

that maps the best $\hat{x}(f(x))$ back to x, usually minimising $L = ||\hat{x}(f(x)) - x||_2^2$.

Sparse AutoEncoders have the additional constraint that the latent representation is sparse, which is encouraged by adding a penality term in the loss function that becomes:

$$L = ||\hat{x}(f(x)) - x||_2^2 + \lambda||f(x)||_0$$

Where $||f||_0$ is the number of nonzero elements in f.[j]

Eventually, since the goal with Sparse AutoEncoders is to discover more features than there are neurons, we have:

$$\mathcal{X} = \mathbb{R}^n, \mathcal{Z} = \mathbb{R}^m, n < m$$

In 2023, Bricken et al. showed the effiency of SAEs in [30] using toy models. They were able to point out features with specific roles such as detecting Hebrew or nucleotide sequences. They also proved features to be significantly more interpretable than neurons.

In 2024, Lieberum et al. released Gemma Scope, a suite of SAEs trained on Google DeepMind's Gemma-2 models. It was presented in [31] with the terminology and the loss function that I used in this section.

Gemma Scope is has been integrated to the Neuronpedia project launched by Johnny Lin in 2023 [32]. The project is a collaboration with leading companies in AI interpretability (including Anthropic and Google DeepMind which are the absolute references in MI), and is completely open source. Although most features are not as blatant as the ones generally used for examples, the possibilities to study the most-activated features for any input - and even to influence outputs by steering specific features like I did for example on figure 3.8 - are very encouraging results for the future of MI.

[j] It is therefore not a norm in a strict mathemtical sense.

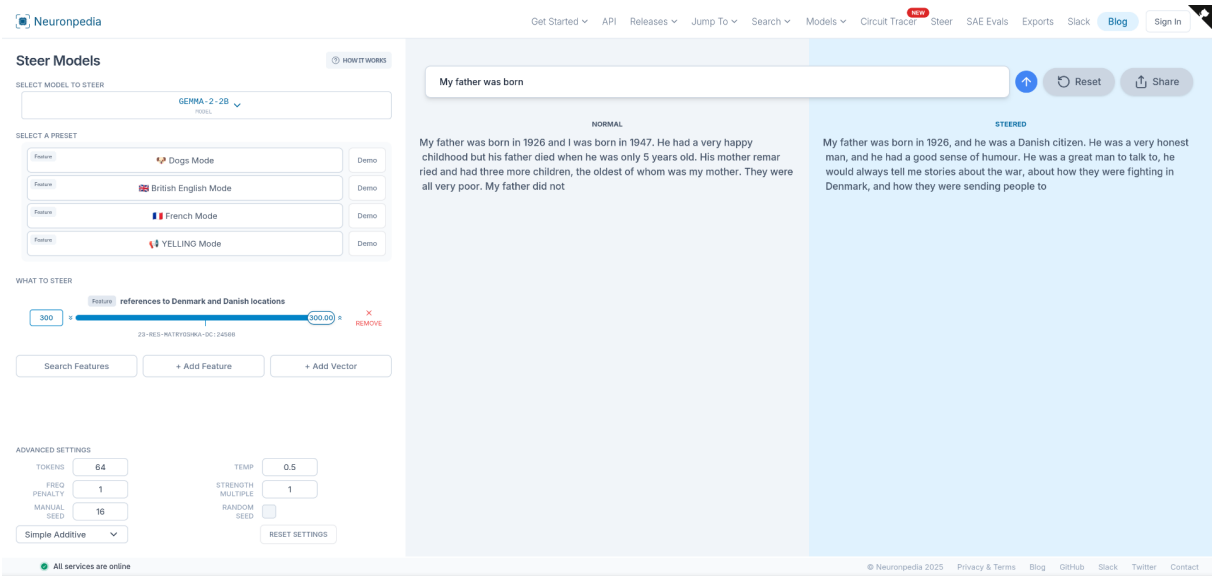

Figure 3.8: Example of text generation with steered feature on neuronpedia.

Circuits

The previous sections dealt with interpretability of individual features. However, the strength of neural networks probably comes in great part from the connections between neurons, and how they assemble to process information through "reasonning" patterns. In the MI vocabulary, they are called *circuits*.

Once again, CNNs make things clearer, and Olah et al. give some simple intuitive examples in [28] such as curve detectors connecting to more complex shape detecors in later layers, or oriented dog head detectors "assembling" into orientation-invariant dog head detectors.

Coming to transformers, the notion of circuits suddenly loses its tangibility. It has however started to be seriously explored with Elhage et al.'s mathematical framework for transformers circuits [33]. Analysing 0, 1 and 2-layers attention-only transformer models[k] they were already able to show algorithmic mechanisms enabled by composition of attention heads.

A more thorough look was taken at *induction heads* in [34]. They tend to appear in every model, and perform the most basic pattern completion: if the sequence [A][B] appears earlier in the text, then [B] is likely to be predicted after [A] if it occurs again [l]. It is argued in the same thread that they may be the primary mechanism behind in-context learning, which refers to transformers' ability to learn from its input [m].

Other types of circuits have been found in transformers such as Inderect Object Identification[n] or "grater than" operations. Recent works attempt to automate the circuit

[k]Transformers with attention layers and no MLP layer.

[l]Induction heads do not appear only for direct succession of strictly equivalent tokens: it can also reproduce a semantic relation between A and B, and apply to distanced tokens.

[m]Take the example of few-shot prompting [14]

[n]The circuit enabling the model to predict "Bob" in an input such as "Alice and Bob were at the beach, when Alice found a shell and gave it to ..."

idenfitication process, by enabling circuit recognition in larger models [35], and automated discovery of interpretable causal graphs [36].

So we reach the current boundaries of the field. Every aspect of MI mentioned and many more have been explored in much more depth and rigour than in this introduction, but the field is very young and new major discoveries are still to be made.

3.3 The hope of immutable intention

Intention in action

"I have an intention to go to the cinema, so I take action to go to the cinema." is an understandable sentence, but from a philosophical perspective, it avoids the complexity behind the term "intention". In philosophy of action, "intention" is not a simple monholitic concept, but a dynamic construct operating at different levels. This led many philosophers to decompose and classify intentions. One of the most popular model has been proposed by Elisabeth Pacherie in [37], and has been enriched with time as present the key points featuring in [38].

Pacherie's model is called "DPM" after the three types of intention it separates: Distal, Proximal and Motor.

- **Distal intentions:** they are presented as "terminators of practical reasoning about ends". In simpler terms, it means they are decisions to take a certain action after a reasoning process, and they are not meant to be reconsidered unless a new reasoning bring other perspectives. Concretely, they usually refer to long-term, abstract commitments such as "I want to go to the cinema" in our introductory example.

- **Proximal intentions:** they often inherit from the action plan of a distal intention, and lead the execution in situation; they ensure the continuous control of the ongoing action. Following the example, such intutions are behind actions such as getting prepared to go the cinema, choosing a movie or taking the bus.

- **Motor intentions:** they also control the ongoing action, but at the lowest level of anchoring in reality. They do not result from a conscious perception of reality, but are rather a neurological effect of perceived sensory-motor information. They are behind goal-directed movements, which in our example can be the hand's unconscious movement to open the door, or the reflex to show the ticket when arriving at the cinema.

There is a hierarchical structure in this classification, often refered to as the *intenentional cascade*. It is also relevant - espcially for the bridge with AI to be made in the next section - to organise these types of intentions according to two axes: their conscious availability for the agent, and their sensible executive role in a present action. This is illustrated in figure 3.9;

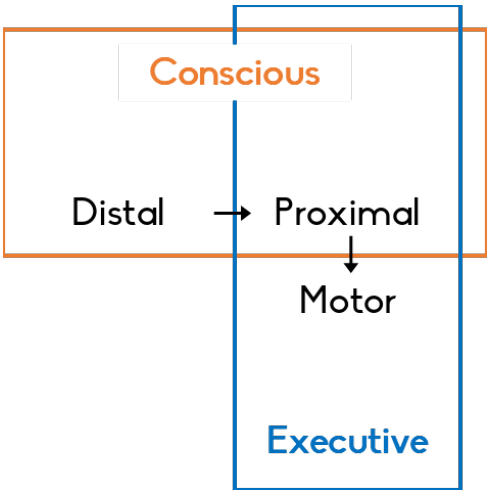

Figure 3.9: Three types of intention in the DPM model.

How do Transformers represent intention?

Intention in action has been studied for human agents, with perspectives combining phenomenology, psychology and neuroscience. However, no attempt has been made to think these ideas for Artificial Intelligence to our knowledge.

However, this is an insight worth being explored. One may argue that no evidence exists that transformers represent intention in any way: even if they exhibit some behaviour comparable to intention in their effects - as anyone can observe prompting a LLM for example - this can be an analogue phenomenon emerging from a mechanism comparable to intention in no other way. This would remain an important result, because this other phenomenon could be worth studying, and questions would arise about the possibility and the consequences of inventing new AI models embedding a management of intentions closer to ours. On the other hand, we may discover that intention is a good framework to explain some of transformers' behaviours. In this case, this could bring original perspectives to the field of interpretability.

Our initial proposition to apply the analogy of intention to the case of transformers is summarised in table 3.1.

Intention	**Application to Transformers**	**Comparable characteristics**
Distal	Training, Fine-Tuning	Broad, general ideas and concepts that yet influence the model's orientation and behaviours.
Proximal	System instructions, Prompt	Influence concrete, task-specific, in-time behaviours.
Motor	Forward pass for token prediction	Low-level, automatic, causal process.

Table 3.1: Intention as analogy for Transformers

The analogy between motor intention and forward passes in a transformer architecture has our main interest, because Mechanistic Intepretability may be a powerful tool to understand this phenomenon commonly represented as happening "inside the black box".

In the cascading hierarchy of the DPM model, motor intuitons are serving the goals of higher-level intentions while expressing the strongest degree of commitment. However, MI has been proved in the previous sections to enable opportunities of analyse and manipulation at the lowest level of the architecture. For the case of transformers, it thereby becomes conceptually conceivable to interfer with the low-level mechanism to artificially construct nearly untouchable intentions, while respecting a coherent theorical framework.

Unfortunately, we have not been able to explore this idea further due to Peter Fezekas - my supervisor for this project - being on vacation, and the necessity for me to return to the MARC project earlier than expected. However, the insights brought by combining intention in action and Mechanistic Intepretability are not to be ignored in our opinion.

4. Side tasks and events

Chapters 2 and 3 presented the two projects I worked during this internship. In this smaller chapter, I will talk about different additional aspects of my internship that were less important on their own, but contributed to the overall experience.

4.1 Talks

In this section, I will present some talks I was offered to attend. They were organised in different departments, and the links with computer science were not always obvious at first glance. However, I like to think one point of interdisciplinarity research is a curiosity and a genuine belief that new ideas and collaborations can appear anywhere. Therefore, for each talk I will introduce the main ideas of the talk itself, but also what it made me think about how it resonated with my own background.

Mads Hansens - University of Bristol
Why you should start thinking in terms of constraints.

The talk started from the notion of mechanism in philosophy, that claims that natural things and phenomena can be understood by decomposing them to understand the causal processes that drive them. The idea in Mads Hansen's talk was to think about how the concept of constraint can be a powerful, nay necessary tool to incorporate to this theory. He also addressed the polysemy of the notion of constraint, for example the questions of constraints being absolute and/or situational.

Talking about constraints immedialty reminded me of the analogue concept in linear programming, in the field of optimisation. One key principle in linear programming is the concept of duality, which reverses the perspectives on what constraints are. There might be a parallel to explore.

Caterina Villani - University of Bologna
Constructing meanings in linguistics.

The main point of this talk was to propose a disambiguation between two dichotomies in a speech: abstractness/concreteness, and generality/specificity. In a linguistic context, these dichotomies are scales on which speeches can be positioned. The first one opposes speeches dealing with abstract, fuzzy, polysemantic concepts to speeches dealing with concrete, tangible, representable topics. The second dichotomy lies on a taxonomy of concepts, where some notions are sub-categories of others. For example: "freedom" is abstract and general while "freedom of speech" is abstract too but more specific. On the

other hand, "bird" is concrete and general, "parrot" still concrete but more specific, and "cockatiel" even more. The major contribution of Caterina Villani's team was to measure and analyse cognitive reactions to speeches and concepts regarding these properties.

I knew nothing about lingustics before attending this talk, yet it seems clear that the field has a lot to bring to Natural Language Processing in AI.

Julia Cramer - Leiden University
Entangling quantum and society.

Julia Cramer has a background in experimental quantum physics and she leads the research group "Quantum and Society" at Leiden University, which aims to study the boundary between quantum technology and science communication. In particular, she explores with her group experts' communication and media coverage of quantum technologies, feelings and reactions of societal groups towards quantum, and studies empirically how to change people's understanding and attitudes towards quantum. The talk was a presentation of her work.

The question of how scientific innovations are perceived in society does not apply only to quantum. My own field being AI, whose questions are massively covered, impact on society hardly measurable, and tends to trigger reactions some may qualify as moral panics; analysing it through the spectre of communication science is important research, that should receive more attention in my opinion.

Judith Bishop - La Trobe University
AI, emotions and creative writing.

Judith Bishop is a poet, a doctor in linguistics, and worked in AI training data industry for more than 17 years. Her talk connnected the dots: she presented the "Emily Dataset", a comparative dataset between original poems from the 19th century American poet Emily Dickinson, and AI-generated poems imitating her style. The result of this dataset could be called "good imitation": the structure of the poems are well reprduced, the vocabulary is appropriate, but deeper analysis shows a lack of "risk" in AI-generated poems, where "rare" or "original" words almost never appear while, from a litterary analysis perspective, they contribute greatly to a writer's style.

Whether or not AI is able of genuine creativity, and if yes is it the case for models already existing, is another huge interdisciplinary question. Something that struck me in Judith Bishop's Emily Dataset is AI's struggle to handle different levels of meaning simultenaously. It led me to the following question: are current AI models only able to manipulate information when it has only one level of interpretation, and if they do not, which clue do we have that they do not incorporate sub-levels of meaning that escape humans' understanding? This sounds like a simple science-fiction question - and it may be one in this case - but I do not think such potentially legitimate questions should be ignored just because of that.

4.2 NordForsk seminar

NordForsk is an organisation affiliated to the Nordic Council[a] whose main mission is to fund, encourage and facilitate Nordic co-operation in research.

In 2018, they launched an initiative for interdisciplinary research [39], and 12 selected projets were funded and started in 2020, including one conducted by the RISR group.

From 18 to 20 June, a closure seminar was held in Malmö, Sweden, during which representatives from every project were invited to present their results.

Figure 4.1: Group photo taken during NordForsk seminar.

Johanna invited me to join, thinking this event was in several aspects a relevent experience to my internship:

- Looking at practical things first: no one else from the RISR group was available, and my presence allowed her not to be alone to represent a research team.
- The event embodied multiple aspects of what a researcher's work can be, especially when accessing to administrative postions: international networking, fundraising, presentation of activity reports.
- The seminar focused on interdisciplinary research, which is the heart of the RISR group, and by extension my internship.

This last point has been particularly observed during the seminar, not only because the presentations provided good examples of interdiscinplinary research projects, but also because some talks were planned explicitely about interdisciplinarity, which was also the topic of many informal discussions with researchers we met. It has been an opportunity to talk about the importance of interdisciplinarity, what it can bring to research, what can be the Nordic added value to such projects; and also in practice what are the pros and cons of interdisciplinary groups, and what lessons should be extracted from the 12 experiences represented at the seminar.

[a]In their own words, "The Nordic Council is the official body for Nordic inter-parliamentary co-operation". It gathers Denmark, Sweden, Norway, Iceland, Finland, and the autonomous areas of Greenland, Faroe Islands and Åland.

4.3 A bit of data analysis

One day I was asked a puzzle that was to help in analysing data collected from an experiment conducted in the RISR group. Here was the puzzle: participants divided into 60 groups reacted to a same material and wrote down a list of terms to describe it. The originality of each term had to be measured using a formula from Guilford [40]: 2 if the term appears only once, 1 if it is mentionned by multiple groups, 0 of the term alone represents at least 1% of all the terms mentionned.

The data was in an Excel file with 60 sheets corresponding to the 60 groups, with for each of them the terms written in the first column. Computing the originality for each term in a non-quadratic time complexity was not a big algorithmic challenge (see the pseudo-code of algorithm1), but a difficulty appeared because the terms were written by the participants, and different group sometimes expressed similar ideas with an equivalent term, or writting the same term in a different format (uppercase/lowercase, dashes/spaces...). Our first idea was to use a LLM to detect these cases, but it sounded like huge abuse of ressources. Instead, I made a function using regex[b] to change terms in a format that matches the one of WordNet Library, which provides a thorough thesaurus I could then extract synonyms from.

Algorithm 1 Originiality computation

Input: Excel File F

$N \leftarrow sum(t$ for Term t in S for Sheet S in $F)$
for Sheet S in F **do**
    for Term t in S **do**
        if $t.originiality = null$ **then**
            $syns \leftarrow t.getSyns()$
            $positions \leftarrow \{Sheet : S.index; Term : t.index\}$
            for Sheet S' in $F[S.index : End]$ **do**
                for Term t' in S' **do**
                    if t' in $syns$ **then**
                        $pos.add(\{F'.index; t'.index\})$
                    end if
                end for
            end for
            $o \leftarrow 2$ if $pos.length = 1$ else (1 if $pos.length/N < 0.01$ else 0)
            for Pair pos in $positions$ **do**
                $F[pos[Sheet]][pos[Term]].originality \leftarrow o$
            end for
        end if
    end for
end for

[b]Regular Expressions: a standard syntax for pattern matching in text.

5. Conclusion

5.1 Lessons and skills acquired

I learnt a lot during this internship, and the technical skills are probably the easiest to start with. The MARC project required to learn how to program with LLMs, and in particular, how to use the llama.cpp library and models on huggingface.co. Later, designing interfaces for a Discord bot using a Slack server were also new skills for me. Eventually, transfering the logic into a Furhat robot required to learn the basics of Kotlin and - although I will probably never use it again - the Furhat SDK.

On a more theorical level, this project brought me inside the world of psychology, which I knew nothing about. I discovered the process for setting up psychological experiments, and to deisgn the chatbot, I had to learn about Motivational Interviewing.

The second project also introduced me to some theorical ideas outside my field, in particular notions of intuition and mechanism in philosophy.

The major theorical gain from this project is way more directly relevant to my studies at ISIMA: I learnt a lot about the theorical fundations of Deep Learning, and continued until cutting-edge concepts in the field of interpretability.

This internship, was also my first real research experience, and it taught me a lot in that regard. It also helped me improve some communication skills. Not that I usually have trouble expressing myself, but in the present context I had to deal with two major challenges: speaking in a foreign language, and talking with people from various research fields. It was very enriching, and in addition to these skills I improved, I benefited a lot from this vibrant interdisciplinary atmosphere.

5.2 About my future

I had very high expectations for this internship. I wanted it to confirm at the same time my interest in reasearch, and the fact that I do not belong in pure computer science.

The first point is validated: every aspect of my work routine during these 4½ months has been delightful. The idea of being on the front line of a field is exciting, meeting people explaining their own research is fascinating, coming with our own ideas is gratifying... My desire to do a PhD is clearer than ever.

The second point is validated too. I have always had major passions for philosophy and arts, and although my interest in mathematics and computer science is genuine, the prospects after graduation remain the main reason why I chose this academic course.

Interdisciplinary research appeared as a possibility to reconnect intensively and somehow professionally with this other part of me, but I had not anticipated the degree it would do so.

I cannot evaluate this internship on my own as regards the quality of my work for the group, nor can I claim that it was the most relevant for a student in an engineering school in computer science.

I can only speak about what it brought me on a personal level: a great experience, and an example of a fulfilling professional future to envision.

Figure 5.1: RISR group photo

Bibliography

References

[1] Research Unit for Robophilosophy and Integrative Social Robotics. *RISR webpage*. URL: https://cas.au.dk/en/robophilosophy/about-us.

[4] Aarhus University. *Aarhus University Climate Strategy 2020 - 2025*. URL: https://international.au.dk/about/profile/sustainability/climate-strategy.

[5] Aarhus University. *Diversity, gender equality and inclusion at Aarhus University*. URL: https://medarbejdere.au.dk/en/strategy/gender-equality-diversity-and-inclusion.

[18] Furhat Robotics. *Furhat robot and developer documentation*. URL: https://docs.furhat.io/.

[39] NordForsk. *Nordic Initiative for Interdisciplinary Research*. 2018. URL: https://www.nordforsk.org/research-areas/nordic-initiative-interdisciplinary-research.

Psychology

[6] S Katherine Nelson et al. "Do unto others or treat yourself? The effects of prosocial and self-focused behavior on psychological flourishing." In: *Emotion* 16.6 (2016), p. 850.

[7] Elie Maalouly et al. "The effect of conversation on altruism: A comparative study with different media and generations". In: *PloS one* 19.6 (2024), e0301769.

[8] Arthur B Herbener, Annette Bohn, and Stefan Pfattheicher. "Bringing Past Kindness Into the Present: Memories of Acts of Kindness Are Vivid, Memorable, and Feel Good". In: *Collabra: Psychology* 10.1 (2024), p. 121246.

[9] Kadian S Sinclair et al. "An adapted postdonation motivational interview enhances blood donor retention". In: *Transfusion* 50.8 (2010), pp. 1778–1786.

[10] Jessica Balderas et al. "Brief online intervention model promotes sustained helping behavior across 6 months following a population-wide traumatic event". In: *Psychological Reports* 128.2 (2025), pp. 1248–1268.

[11] William R Miller and Stephen Rollnick. *Motivational interviewing: Helping people change*. Guilford press, 2012.

[12] TB Moyers, JK Manuel, and D Ernst. "Motivational interviewing treatment integrity coding manual 4.1". In: *Unpublished manual* 1 (2014), p. 3.

[13] Wikibooks. *Bestiary of Behavioral Economics/Dictator Game*. URL: https://en.wikibooks.org/wiki/Bestiary_of_Behavioral_Economics/Dictator_Game.

[15] Mian Zhang, Shaun M Eack, and Zhiyu Zoey Chen. "Preference Learning Unlocks LLMs' Psycho-Counseling Skills". In: *arXiv preprint arXiv:2502.19731* (2025).

[17] Rachel Baumsteiger. "What the World Needs Now: An Intervention for Promoting Prosocial Behavior". In: *Basic and Applied Social Psychology* 41.4 (2019), pp. 215–229. DOI: 10.1080/01973533.2019.1639507. eprint: https://doi.org/10.1080/01973533.2019.1639507. URL: https://doi.org/10.1080/01973533.2019.1639507.

[40] Joy Paul Guilford. "The nature of human intelligence." In: (1967).

Artificial Intelligence

[14] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[16] Sasha Aickin. *How many few shots examples should you use?* 2024. URL: https://www.libretto.ai/blog/how-many-few-shot-examples-should-you-use.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).

[21] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[22] Alec Radford et al. "Improving language understanding by generative pre-training". In: (2018).

[23] Andrea Gesmundo and Kaitlin Maile. "Composable function-preserving expansions for transformer architectures". In: *arXiv preprint arXiv:2308.06103* (2023).

[24] Zachary C Lipton. "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3 (2018), pp. 31–57.

[25] Chris Olah. "Mechanistic Interpretability, Variables, and the Importance of Interpretable Bases". In: *Transformer Circuits Thread* (2022). URL: https://transformer-circuits.pub/2022/mech-interp-essay.

[26] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. "Feature visualization". In: *Distill* 2.11 (2017), e7.

[27] Chris Olah et al. "Naturally occurring equivariance in neural networks". In: *Distill* 5.12 (2020), e00024–004.

[28] Chris Olah et al. "Zoom in: An introduction to circuits". In: *Distill* 5.3 (2020), e00024–001.

[29] Nelson Elhage et al. "Toy Models of Superposition". In: *arXiv preprint arXiv:2209.10652* (2022). URL: https://transformer-circuits.pub/2022/toy_model.

[30] Trenton Bricken et al. "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning". In: *Transformer Circuits Thread* (2023). URL: https://transformer-circuits.pub/2023/monosemantic-features.

[31] Tom Lieberum et al. “Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2”. In: *arXiv preprint arXiv:2408.05147* (2024).

[32] Johnny Lin. *Neuronpedia: Interactive Reference and Tooling for Analyzing Neural Networks*. Software available from neuronpedia.org. 2023. URL: https://www.neuronpedia.org.

[33] Nelson Elhage et al. “A mathematical framework for transformer circuits”. In: *Transformer Circuits Thread* (2021). URL: https://transformer-circuits.pub/2021/framework.

[34] Catherine Olsson et al. “In-context learning and induction heads”. In: *arXiv preprint arXiv:2209.11895* (2022).

[35] Arthur Conmy et al. “Towards automated circuit discovery for mechanistic interpretability”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 16318–16352.

[36] Samuel Marks et al. “Sparse feature circuits: Discovering and editing interpretable causal graphs in language models”. In: *arXiv preprint arXiv:2403.19647* (2024).

Philosophy

[2] Johanna Seibt, Malene Flensborg Damholdt, and Christina Vestergaard. “Five principles of integrative social robotics”. In: *Envisioning robots in society–power, politics, and public space*. IOS Press, 2018, pp. 28–42.

[3] Johanna Seibt. “Classifying forms and modes of co-working in the ontology of asymmetric social interactions (OASIS)”. In: *Envisioning robots in society–Power, politics, and public space*. IOS Press, 2018, pp. 133–146.

[37] Elisabeth Pacherie. “Towards a dynamic theory of intentions”. In: *Does consciousness cause behavior* (2006), pp. 145–167.

[38] Myrto Mylopoulos and Elisabeth Pacherie. “Intentions: The dynamic hierarchical model revisited”. In: *Wiley Interdisciplinary Reviews: Cognitive Science* 10.2 (2019), e1481.