

Department of Computer Science

University College London

**Computational Persuasion using
Chatbots based on
Crowdsourced Argument Graphs &
Concerns**

Lisa Andreevna Chalaguine

Supervisor: Anthony Hunter

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science at
University College London, January 2023



I hereby declare that this thesis and the work reported herein was composed by and originated entirely from me. Information derived from the published and unpublished work of others has been acknowledged in the text and references are given in the list of sources.

Lisa Andreevna Chalaguine (2023)

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial license. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, and that they do not use it for commercial purposes. For any reuse or redistribution, researchers must make clear to others the license terms of this work.

Abstract

As computing becomes involved in every sphere of life, so too is persuasion a target for applying computer-based solutions. Conversational agents, also known as chatbots, are versatile tools that have the potential of being used as agents in dialogical argumentation systems where the chatbot acts as the persuader and the human agent as the persuadee and thereby offer a cost-effective and scalable alternative to in-person consultations

To allow the user to type his or her argument in free-text input (as opposed to selecting arguments from a menu) the chatbot needs to be able to (1) “understand” the user’s concern he or she is raising in their argument and (2) give an appropriate counterargument that addresses the user’s concern.

In this thesis I describe how to (1) acquire arguments for the construction of the chatbot’s knowledge base with the help of crowdsourcing, (2) how to automatically identify the concerns that arguments address, and (3) how to construct the chatbot’s knowledge base in the form of an argument graph that can be used during persuasive dialogues with users.

I evaluated my methods in four case studies that covered several domains (physical activity, meat consumption, UK University Fees and COVID-19 vaccination). In each case study I implemented a chatbot that engaged in argumentative dialogues with participants and measured the participants’ change of stance before and after engaging in a chat with the bot. In all four case studies the chatbot showed statistically significant success persuading people to either consider changing their behaviour or to change their stance.

Acknowledgements

Undertaking this PhD has been a truly life-changing experience for me and it would not have been possible to do without the support and guidance that I received from many people.

Firstly, I would like to express my highest gratitude to Anthony Hunter - my supervisor and guide throughout my academic life. Without his guidance and constant feedback, this PhD would not have been achievable. The past two years were especially challenging and I would like to thank him for being particularly patient during that time (a pandemic and failed peaceful revolution in your home country do not necessarily contribute to a healthy state of mind). I would also like to thank my secondary supervisor Henry Potts, who was always available when I needed him.

I gratefully acknowledge the funding received for my PhD from the Engineering and Physical Sciences Research Council. Also, a big thanks to the Computer Science Department of UCL for providing me with extra funding for the conferences I attended during the course of my PhD and the Faculty of Engineering, for giving me this invaluable opportunity to pursue a PhD, and for acknowledging me as a good teaching assistant¹.

A wielkie dzięki goes to Sylwia Polberg, my former colleague, whose advice and support were indispensable to my success. I am also extremely grateful for the competent admin support I received throughout my PhD from Sarah Turnbull and Elizabeth Halton.

I would also like to thank my parents (who kindly sponsored all the hardware during my PhD), my brother, and my boyfriend for being hugely supportive throughout my studies. I also want to thank my friends, especially Claudia Schulz, who introduced me to Anthony. Without her, this all would not have happened in the first place.

¹<https://www.ucl.ac.uk/engineering/news/2019/nov/iep-team-celebrates-another-successful-academic-year>

A big thanks also go to all the people I had the pleasure of teaching with, especially Dariush Hosseini, and to all students who I have met while being a teaching assistant at UCL.

Last but not least I want to acknowledge the valuable gift of time I was given during my PhD which allowed me to become a better educated, stronger and healthier human being (although the past two years took their toll on me).
Russia will be free. Glory to Ukraine. Long live Belarus. Россия будет свободной - Слава Україні - Жыве Беларусь!



Impact Statement

Argumentation aims to exchange different viewpoints or opinions, handle conflicting information and make informed decisions. A key role for argumentation is persuasion, and computational persuasion incorporates computational models of argument in software agents that can persuade people. This can be potentially valuable in roles such as behaviour change where the aim is to get the persuadee to make specific changes to the lifestyle (e.g. to eat more fruit, exercise more, commute by cycle, etc.) that can benefit them or those around them. For example, such a software agent (like a chatbot) could persuade people to do more sports by presenting arguments in favour of exercising and countering the arguments given by the user as to why she is not willing to. However, explicit use of argumentation is not central to most current manifestations of persuasion technologies. Yet, presenting explicit arguments in applications that facilitate behaviour change can be highly effective, particularly when someone is lacking some essential information or entertaining misconceptions about a topic.

In this work, I present a framework for developing persuasive chatbots, which includes methods for the acquisition of arguments, ways to structure those arguments in the form of argument graphs which serve as the chatbot's knowledge base, and strategies for the chatbot to select moves that increase the persuasiveness of the chat. Given my research was backed up with case studies with actual participants, I have acquired several small datasets of chatlogs and a big argument graph with over 1200 arguments on the topic of UK university fees. The datasets, as well as the source code for all the chatbots, are open source and can be used by other researchers.

The research presented in this thesis offers many attractive future research opportunities in the fields of natural language processing and computational models of argument like question answering, argument mining and argument detection, dialogue strategies, and argument generation; as well as in domains concerning persuasion and behaviour change, like health psychology. The

chatbots also have the potential to be utilised for public health purposes. In a post-pandemic world, with NHS waiting lists at record highs, and no human resources to spare, it is as important as ever that people live a healthy lifestyle and make decisions that are beneficial for themselves, as well as those around them - for example getting vaccinated. But given we live in a democratic country, we cannot force people to take those decisions against their will and need to use dialogue. Doing so by automatic means with the help of chatbots makes it possible to talk to thousands and even millions of people, address their concerns, and answer their questions from the comfort of their own homes and without discrimination, given that 96% of Britons have internet access. Learning about people's concerns on a certain topic (e.g. why people are reluctant to get a vaccine) would also help policymakers to design appropriate policies to address the issues at hand. The vast amount of data that would be collected would also allow the researchers and developers of these chatbots to improve them over time and to enhance user experience.

Contents

Abstract	5
Acknowledgements	7
1 Introduction	19
1.1 Problem Statement	21
1.2 Proposed Solution	23
1.2.1 Desired Features for my Chatbot	25
1.3 Thesis Narrative	27
1.3.1 Initial Idea	27
1.3.2 Move to Persuasive Chatbots using Concerns	29
1.3.3 Construction of big Argument Graphs and Chatbots with Natural Language Understanding	30
1.3.4 Different Approach to Knowledge Base Construction, depending on Domain	31
1.4 Scope and Structure of Thesis	33
1.5 Thesis Contributions	34
1.6 Publications	36
2 Literature Review	38
2.1 Dung’s Abstract Argumentation Framework	39
2.2 Overview of Argumentation Strategies	42
2.3 On the Notion of Persuasion	45

2.4	The Notion of Concerns in Computational Argumentation & Persuasion	49
2.4.1	Value-Based Argumentation	50
2.4.2	Concerns	53
2.5	Automated Persuasion using Knowledge Graphs and Menu-Based User Input	55
2.6	Argument Acquisition for Knowledge Base Construction	58
2.7	Conceptualising Chatbots	61
2.7.1	History	61
2.7.2	Task-specific chatbots since 2016	62
2.7.3	Examples of Argumentative Chatbots	63
2.8	Conclusion	65
3	Argument Harvesting Using Chatbots	67
3.1	Introduction	68
3.2	Concern Assignment and Suitability of Counterarguments	69
3.3	Hypotheses	71
3.4	Chatbot Design for Argument Harvesting	72
3.5	Experiments	75
3.5.1	Argument Harvesting	76
3.5.2	Experiment I: Argument-Concern Labeling	78
3.5.3	Experiment II: Assessment of Harvested Arguments as Meaningful Arguments	80
3.5.4	Experiment III: Argument-Counterargument Matching	81
3.6	Discussion	86
4	Impact of Argument Type and Concerns in Argumentation with a Chatbot	88
4.1	Introduction	89
4.2	Argument Types and Appeal	90

4.3	Hypotheses	93
4.4	Argument & Counterargument Acquisition	94
4.5	Experiments	97
4.5.1	Evaluation of Argument Types	97
4.5.2	Evaluation of Chatbot	99
4.6	Discussion	105
5	Knowledge Acquisition and Corpus for Argumentation-Based Chatbots	107
5.1	Introduction	107
5.2	Method	109
5.2.1	Argument Processing	109
5.2.2	Argument Acquisition for Next Depth	111
5.3	Case Study and Corpus	112
5.3.1	The Corpus	115
5.4	Evaluation	115
5.5	Discussion and Conclusions	118
6	A Persuasive Chatbot using a Crowdsourced Argument Graph and Concerns	120
6.1	Introduction	121
6.2	Hypotheses	124
6.3	Chatbot Design	125
6.3.1	Argument Graph	125
6.3.2	Default Arguments	126
6.3.3	Concern Labelling and Classification	126
6.4	Evaluation of the Chatbot	128
6.5	Discussion	134
7	Addressing Popular Concerns regarding COVID-19 Vaccination	

with Natural Language Argumentation Dialogues	136
7.1 Introduction	137
7.2 Hypotheses	139
7.3 Chatbot Design	141
7.3.1 Knowledge Base Construction	141
7.3.2 Understanding the user input	144
7.4 Experiments	145
7.5 Evaluation of the Chatbot	147
7.6 Discussion and Conclusion	151
8 Conclusion	155
8.1 Contributions	155
8.2 Limitations	157
8.3 Future Work & Research Areas of Interest	159
8.3.1 Question Answering	160
8.3.2 Argument Mining & Argument Detection	161
8.3.3 Dialogue Strategies	163
8.3.4 Persuasion	164

List of Figures

1.1	Simple argument graph showing a potential dialogue move between a chatbot and a user.	24
2.1	Argument Graph for Example 2.3	52
3.1	Example chat between a participant and the AH1 chatbot. Chatbot arguments are in the dark boxes and user arguments in the light boxes. The chat begins in the left column and continues in the right one.	77
4.1	Example chat between a participant and the AH1 chatbot. Chatbot arguments are in the dark boxes and user arguments in the light boxes. The chat begins in the left column and continues in the right one.	103
5.1	Representation of depths and attack relationships between arguments in my argument graph. Arguments <i>B</i> , <i>C</i> and <i>D</i> are counterarguments to <i>A</i>	108
5.2	Visualisation of the generated argument corpus in graph form .	116
5.3	Scores for each individual dialogue.	117
6.1	Argument graph where child nodes are attacking parent nodes.	122

6.2	<p>Example chat between a participant and the strategic chatbot. Chatbot arguments are in the dark boxes and user arguments in the light boxes. Default arguments are indicated with an *. The chat begins in the left column and continues in the right one.</p>	132
7.1	<p>Part of the chatbot’s knowledge base with concerns <i>a</i> and <i>b</i> representing clusters of arguments in level 1 (concerned about insufficient testing of the vaccine and possible side effects respectively) attacking the chatbot’s goal argument, and arguments in level 2 addressing the concerns.</p>	143
7.2	<p>Example chat between a participant from the fourth batch. Chatbot arguments are in the dark boxes and user arguments in the light boxes. Default arguments are indicated with an *. The chat begins in the left column and continues in the right one. The participant indicated that he or she was <i>somewhat unlikely</i> to get vaccinated before the chat but changed their stance to <i>neither likely nor unlikely</i> after the chat.</p>	151

Chapter 1

Introduction

Computational argumentation has become an important field within Knowledge Representation (KR) which is the field of Artificial Intelligence (AI) dedicated to representing information about the world and automating various forms of argumentation such as joint deliberation, persuasion, negotiation and conflict resolution with the help of multiagent systems [Ontanón and Plaza, 2006]. Argumentation is of enormous importance in human communication and decision-making. Whether about coffee preparation, music taste, or legal cases, arguing has always been the dominant way of rationalising opinions [Boltužić and Šnajder, 2014]. It helps to select one or several alternatives or to explain and justify an already adopted choice [Amgoud and Prade, 2009]. Consequently, when voicing their opinion on a certain issue or explaining why they took a particular decision, people use arguments. It is therefore not surprising that the development of automated agents that can engage in argumentation with human agents has become a hot research topic.

Computational persuasion can be seen as a sub-field of computational argumentation. Persuasion is an activity that involves one party trying to persuade another party to believe or do something. Sales and marketing are heavily

dependent on persuasion; however, physicians, when influencing their patients to take their medications or change their lifestyles, or politicians, when explaining to the public why vaccinations are essential, also rely heavily on persuasion. Agents that can engage in a persuasive dialogue with a user to persuade them to do (or not to do) some action have the potential to be used in domains like behaviour change, politics, or addressing social issues. To do so, the agent aims to use convincing arguments and counterarguments to persuade the user to adopt the agent's stance.

During my research, I have worked on developing methods to acquire and construct knowledge bases for conversational agents that can engage in persuasive dialogues with users. The agent's goal was to persuade the user, using reasonable arguments, to accept its stance. The agents took free-text input from the user and were therefore *chatbots*. I have deployed several chatbots on various topics and tested and evaluated them in experiments with participants. I acknowledge that argumentation is by no means the only way to persuade people to change their stance and especially behaviour. However, in my research, I only focused on the argumentative aspect of persuasion.

To the best of my knowledge, my work on persuasive chatbots is the first one which (1) uses argumentation for persuasion whereby the user inputs free text arguments and the system/chatbot uses natural language processing to reply with an appropriate counterargument, (2) uses crowdsourcing as a means of constructing argument graphs, and (3) uses crowdsourced argument graphs as the chatbot's knowledge base.

In the remainder of the chapter, I will outline the problem statement (Section 1.1), my proposed solution (Section 1.2), present the narrative of my research and show how each subsequent study was built upon previous findings (Section 1.3), and the scope and structure of the thesis (Section 1.4). Further, I

summarise the thesis contributions (Section 1.5) and list the publications that resulted from the research conducted in writing this thesis (Section 1.6).

1.1 Problem Statement

Argumentation aims to exchange different viewpoints or opinions, handle conflicting information and make informed decisions. The importance of argumentation has led to the development of computational models of argument that aim to formalise aspects of argumentation within software. A key role for argumentation is persuasion, and computational persuasion incorporates computational models of argument in software agents that can persuade people. This can be potentially valuable in roles such as behaviour change where the aim is to get the persuadee to make specific changes to the lifestyle (e.g. to eat more fruit, exercise more, commute by cycle, etc.) that can benefit them or those around them [Hunter, 2016]. For example, such a software agent could persuade people to do more sports by presenting arguments in favour of exercising and countering the arguments given by the user as to why she is not willing to. The software agent is thereby engaging in an argumentation dialogue where it acts as the persuader and the user as the persuadee. Such computer-based solutions that automate persuasion, known as persuasion/persuasive technologies, have come out of developments in human-computer interaction research (see, for example, the influential research by BJ Fogg [Fogg, 2002]). However, Fogg ignores arguments - a gap to be filled by computational models of argument.

Currently, persuasion technologies are only common in the behaviour change domain, and explicit use of argumentation is not central to most current manifestations of persuasion technologies. Either arguments are provided implicitly in the persuasion technology (e.g., through the provision of information or game playing), or it is assumed that the user has considered the arguments for

changing behaviour before accessing the persuasion technology (e.g., when using diaries or receiving email reminders). Explicit argumentation with consideration of arguments and counterarguments is not supported by existing persuasion technologies [Hunter, 2014a]. Software agents that can engage in argumentation dialogues in domains outside of behaviour change have been presented in the literature (e.g. [Le et al., 2018, Rakshit et al., 2019]). However, their goal was to keep the chat going, not to persuade.

Yet, presenting explicit arguments in applications that facilitate behaviour change can be highly effective, particularly when someone is lacking some essential information, or entertaining misconceptions about a topic. However, for the agent to be able to engage in persuasive dialogues, it needs to be equipped with arguments from both sides. It needs to be aware of arguments for (not) engaging in a particular behaviour that could potentially be given by the user and arguments that attack the user's arguments (counterarguments), which the agent can use to reply to the user's arguments. This calls for the development of methods for the acquisition of appropriate arguments and counterarguments that can be used by the persuasive agent and an appropriate structure to store them in the agent's central repository, also known as a knowledge base.

To reply with an appropriate counterargument, the agent must know what qualifies as a counterargument to the argument the user gives. If, for example, the user says that they do not have enough time to engage in physical activity due to other obligations, like family and work, the chatbot's counterarguments need to address those issues, for example, by suggesting to do sports together with the whole family. A counterargument that fails to address these issues is less likely to be perceived as persuasive by the user. I, therefore, also need methods to establish argument-counterargument relationships amongst the acquired arguments.

Further, there may be several counterarguments to choose from, and their degree of impact can vary. If a university student says he has no time to do sports because he spends most of his time studying, one could think of several appropriate counterarguments. For example, that university is not just about learning but also about making friends and building relationships, and joining a sports team or club would provide the perfect opportunity. However, the counterarguments that doing sport will improve the student's physical health and help him to study better is also valid. The best choice of counterargument by the agent during a persuasive dialogue, therefore, depends on the values and concerns of the user, as well as the overall persuasiveness of an argument.

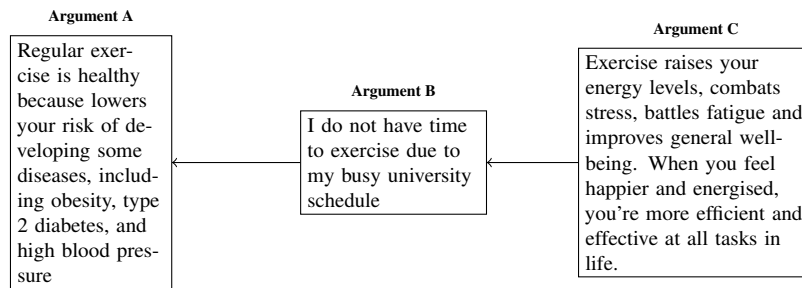
1.2 Proposed Solution

Three things are required for the development of a software agent that can engage in persuasive dialogues with people:

1. A medium for the agent. For example, a virtual agent in the form of an app that is equipped with voice recognition software; or a chatbot that a person can converse with via text.
2. A method to acquire arguments and counterarguments and represent their relationships that can be used for the agent's knowledge base, as well as methods to assess the quality and verify the validity of the arguments used by the agent.
3. A strategy to persuade different people. Arguments that might convince one person might fail to convince another. Therefore, an agent must accommodate that and choose arguments that increase persuasiveness during the dialogue with individual users.

A situation involving argumentation can be represented by a directed graph

Figure 1.1: Simple argument graph showing a potential dialogue move between a chatbot and a user.



that captures the attack relationships between arguments, as proposed by Dung [Dung, 1995]. Each node represents an argument, and each arc denotes an attack by one argument on another. One can then analyse such a graph to determine which arguments are acceptable according to some general criteria [Besnard et al., 2014, Baroni et al., 2011]. A chatbot could utilise such a graph to store arguments in its knowledge base. Figure 1.1 shows a simple argument graph showing a potential dialogue move between chatbots and a user. Argument *A* could be given by the chatbot, the user could counter with argument *B*, and the chatbot would counter the user’s argument with argument *C*.

In my studies, I use a chatbot as a software agent. Chatbots are software systems that can converse with people via text. I propose a method to acquire user arguments and the chatbot’s counterarguments in graph form via crowdsourcing and a hybrid approach (where the chatbot’s arguments are curated by the researcher), depending on the domain. The advantage of crowdsourcing is that it does not require professional research but solely relies on the input of participants and scales easily, which allows obtaining many arguments from different people and thereby create large and comprehensive argument graphs. Crowdsourcing the chatbot’s arguments, however, is less suitable for domains that require more professional knowledge, in which case a domain specialist should manually craft the arguments that the chatbot uses. Developing auto-

matic means for assessing the quality and validity of arguments was out of the scope of my research.

And lastly, I present the notion of *concerns* and a way of automatically identifying them in the users' arguments that the chatbot can use to choose suitable arguments to increase the persuasiveness of the dialogue. This idea is based on the fact that the persuader might present a perfectly valid argument, and the persuadee might not even disagree with it. However, if the argument addresses an issue the persuadee is not concerned about, this argument may have no impact on her stance or behaviour. But if the persuader presents an argument that addresses her concern, it is more likely to change her stance or behaviour. It should be noted that during my studies, I have focused more on the identification of user's concerns than the evaluation of the persuasiveness of individual arguments.

1.2.1 Desired Features for my Chatbot

When I first started to investigate dialogical argumentation systems for persuasion, I noticed that most of the existing theories make at least one of the following two strong assumptions: either that all the arguments in the domain are known and/or that the agent has some knowledge about the user. To accommodate these assumptions, none of the existing persuasive dialogue systems that use computational models of arguments allows free-text input. They are menu-based systems where the user can select his arguments from a pre-defined list. For such a system to be useful in a dynamic, data-driven world, it should allow the user to express themselves freely, given that he or they might have access to more information than the chatbot developers had at the time of development. I believe that the following three features are necessary for an argumentative chatbot to be useful for the general public:

1. **Handle free-text input.** Given that I am not assuming that all arguments

are known in advance, and I am interested in using the chatbot to discover new arguments that can be added to its knowledge base, the chatbot has to handle free-text input from the user. This means that (1) there is no moderation of the chat, and the user is free to give any arguments he or she wants, even if they do not follow any known argumentation protocol, which makes the chat more general, interesting and, most importantly, realistic. And (2) that the chatbot cannot use a global strategy to plan the outcome of the chat but needs to use a local one depending on the user's utterance at the current time step.

2. **Not assume knowledge of the user.** Many existing argumentation frameworks assume some (even if uncertain) knowledge about the persuadee's beliefs. In a real-world scenario, however, it is seldom the case that information about specific users is openly available and hence my chatbot does not rely on any prior knowledge about the user. Any person who has an internet connection should be able to engage in a persuasive dialogue with the chatbot.
3. **Engage in a persuasive dialogue.** As I am concerned with designing a chatbot that may be used in the behaviour-change domain, the chatbot should engage in a one-sided persuasion dialogue, where the chatbot is trying to persuade the user to accept the chatbot's stance (e.g. to take a vaccination) but the user is not trying to convince the chatbot to accept their stance. For the scope of this thesis, the chatbot is not engaging in negotiation, inquiry, or deliberation.

Further, I want to mention two more aspects. First, the majority, if not all, existing dialogue systems that make use of computational models of arguments, assume the presence of a scenario-specific knowledge base created by experts, in the form of an argumentation graph. If one has the resources to acquire

such an expert knowledge base, one should certainly make use of this. I am, however, not making this assumption and presenting a method to quickly and efficiently crowdsource arguments and use them to create the chatbot’s knowledge base.

And secondly, I want my chatbot to be of use in the real world, with people from all backgrounds and classes. I focus on crowdsourced, simple arguments from ordinary people that are nevertheless full of meaning and importance. They are the kind of arguments that have been neglected in the formal and informal argumentation literature.

1.3 Thesis Narrative

In this section, I recapitulate my research and present it as a “narrative” to demonstrate how each subsequent study was built upon previous findings.

1.3.1 Initial Idea

My initial idea was to develop a chatbot that could collect arguments from users, for which I coined the term “argument harvesting”, to use those to construct argument graphs, such as proposed by Dung [Dung, 1995]. The motivation behind this was that, despite argument graphs being a necessary formalism in computational models of argument, the issue of acquiring the graphs tended to be omitted. To construct graphs using *real* arguments, as opposed to theoretical, made-up scenarios, arguments must be acquired from real-life sources.

A common approach to argument acquisition assumes a static resource available on the internet where the topic of interest is/was already discussed. An excellent example of such an approach is presented in [Boltužić and Šnajder,

2014], where the authors compiled a corpus of user comments (manually annotated with arguments) using two debating websites. The authors point out that online user discussions are by far the largest source of opinions, giving (1) comments on newspaper articles, (2) social networks, (3) blogs, and (4) discussion forums as examples (I will summarise all those platforms as *social media*). However, such user-generated content is neither moderated nor structured, making the construction of argument graphs challenging.

I wanted to primarily focus on *behaviour change* and harvest arguments about why people engage in certain unhealthy behaviours or do not engage in healthy ones respectively, and appropriate counterarguments on why they should (not) engage in them. This made the use of social media even more problematic since people rarely discuss why they smoke, do not follow a healthy diet, or do not exercise regularly on public websites.

I, therefore, wanted to create a chatbot that would automatically populate argument graphs on such topics. I made the first attempt in my first study, which is described in Chapter 3, where I describe a model for argument harvesting using a simple chatbot with no domain knowledge. I present ideas on how to structure an argument graph with the harvested arguments, which could be utilised as the knowledge base for a subsequent chatbot to continue expanding the graph. I make use of the notion of *concerns* (which were inspired by the use of *values* in value-based argumentation [Bench-Capon, 2002a]) to label the arguments and show that different people assign the same or a similar concern to a given argument. I then show that similar arguments that address the same concern can be countered with the same counterargument. These two concepts (argument graphs for knowledge bases and the use of concerns) were used and expanded throughout all my future studies.

1.3.2 Move to Persuasive Chatbots using Concerns

During my first study, I defined a counterargument as *suitable* if it had the same concern assigned to it as the argument it attacks. In traditional value-based argumentation (VBA), each agent usually holds the same value during the whole debate. For example, one party might hold the value *life* to justify breaking into a house to steal some food because they were starving, whereas the other party might hold the value *property*. Therefore, in VBA the arguments of one party and the counterarguments of the other are promoting different values. On the other hand, my chatbot wants to convince the user to change her behaviour/stance and not only defend its stance or argue for the sake of arguing. Consequently, the user's argument and the chatbot's counterargument have the same concern assignment, where the argument *raises* a concern and the counterargument *addresses* the concern at hand.

I decided to change the method I used to acquire arguments for the chatbot's knowledge base. Using argument harvesting with the help of a chatbot (1) costs more money and time due to the duration of the chat in comparison to, for example, an online questionnaire, (2) takes longer to develop, and (3) results in unnatural dialogues. A chatbot is a much more suitable medium for persuasive dialogues; hence, I decided to shift my research focus to developing persuasive chatbots instead of using them as a means of argument acquisition.

I used *crowdsourcing* to acquire arguments and counterarguments for the chatbot's knowledge base with the help of Google or Microsoft forms. The crowdsourced arguments were used to automatically identify users' concerns and construct the chatbot's knowledge base. The resulting chatbot was able to provide counterarguments to the users' arguments based on the users' concerns and, therefore, engage in persuasive dialogues. My first study, where I used crowdsourcing to acquire user arguments to identify their concerns and developed a chatbot that tried to persuade the user to change their stance, is

described in Chapter 4.

1.3.3 Construction of big Argument Graphs and Chatbots with Natural Language Understanding

My next move was to construct a sufficiently large argument graph that covered a wide range of possible arguments on a particular topic that the chatbot could utilise as its knowledge base. My first persuasive chatbot only had 24 arguments in its knowledge base, which were presented one after the other during the chat. The chatbot was unaware of the users' arguments and was therefore unable to counter them, but only to give a new one which was not a reply to the user's argument. Hence, the chats only consisted of argument-counterargument pairs. The resulting argument graph would have extensive breadth but not go beyond two levels: the chatbot's arguments and the users' counterarguments.

I chose UK university fees as the topic and used an iterative approach to collect the arguments for each depth of the argument graph using crowdsourcing. The resulting argument graph consisted of 5 levels of depth, starting with the root statement "University fees in the UK should be kept at 9k pounds" (depth 0) and over 1200 arguments. The arguments in each level of depth countered the arguments from the previous level of depth. The acquisition of the argument graph is covered in Chapter 5.

Depths 1-4 of the graph were used as a knowledge base for the chatbot presented in Chapter 6. Depths 1 and 3 contained arguments against keeping university fees. In contrast, depths 2 (attacking depth 1 arguments) and 4 (attacking depth 3 arguments) contained arguments that support the chatbot's stance on keeping university fees. When the user typed in an argument (source argument), the chatbot used a similarity measure to find the closest match of

the user argument in the graph (target argument). If the chatbot found a similar argument in the graph, the chatbot chose one of the counterarguments that attacked the target argument in the graph as a response.

I also further developed my idea of using concerns to make the dialogues between the user and the chatbot more persuasive. I developed two versions of the chatbot, one that did not take the user's concerns into account (baseline) and one that did (strategic). The strategic chatbot was significantly more persuasive than the baseline, which did not consider the users' concerns.

1.3.4 Different Approach to Knowledge Base Construction, depending on Domain

One drawback of the big argument graph was that it included many similar arguments and, hence, also counterarguments. Some participants pointed out that they often received similar counterarguments from the chatbot. It should be noted that my aim when acquiring the large argument graph was not only to use it for a chatbot but also to provide a corpus for fellow researchers. But it is a fair criticism to say that using it as a chatbot knowledge base, clustering similar user arguments and deleting similar counterarguments would have been desirable. However, clustering similar arguments automatically proved to be quite challenging. The argument graph might be significant for an argument graph, but it was still comparatively small for a corpus to apply unsupervised clustering techniques.

Therefore, I needed an approach that would account for this problem during the acquisition stage - instead of retrospectively. I picked COVID-19 vaccination as a case study. I started collecting the arguments for depth 1 of the argument graph (reasons why people are reluctant to get a vaccine once one becomes available). During this acquisition stage, I noticed an interesting pattern:

arguments that raised the same concern were much more similar than in my previous work about UK university fees. In contrast to my earlier works, where I covered established domains, COVID-19 was a novel domain. Due to the novelty of the problem and the lack of information at the time of writing, the domain of potential novel arguments against a COVID-19 vaccine was smaller and less complex than, for example, the domain of arguments for abolishing/maintaining university fees. Therefore, there was no need to collect counterarguments for several different arguments to get good coverage of counterarguments.

This resulted in a different structure of the chatbot's knowledge base. Instead of a traditional argument graph, where each node denotes an argument, and each arc indicates one argument attacking another, the knowledge base consisted of nodes representing the user's concerns in depth 1, and nodes representing arguments that address the concerns in depth 2. The chatbot was, therefore, equipped with a concern classifier to identify the concern raised by the incoming user argument. But there was no need for a similarity measurement since the chatbot replied with one of the counterarguments that addressed the concern raised by the argument. This made it much easier only to include different counterarguments and avoid repetition.

This was the first study described in detail in Chapter 7, where I also used the chatbot to expand the graph with new arguments (concerns in this case) from the user chats. I show a whole pipeline of (1) acquiring the initial knowledge base, (2) training a concern classifier with the crowdsourced arguments, (3) implementing a chatbot that utilises the knowledge base and a concern classifier to engage in persuasive dialogues and (4) analysing the chats to identify new concerns that can be added into the chatbot's knowledge base along with suitable counterarguments.

This completes the circle: my initial idea was to use chatbots to acquire arguments and construct argument graphs with the harvested arguments. However, during my research, I discovered that it is better to use crowdsourcing to build the chatbot's *initial* knowledge base. And then, with the help of persuasive dialogues that the chatbot engages in with the users, discover new arguments and concerns, and add them to the chatbot's knowledge base.

1.4 Scope and Structure of Thesis

The thesis aims to provide a framework for the development of persuasive chatbots, including argument acquisition for the chatbot's knowledge base, methods to structure the knowledge base according to the domain, and suggest a strategy that can be used in order to produce persuasive dialogues, with the help of concerns. In Chapter 2, I present the relevant literature.

During the project, I undertook four studies to investigate the abovementioned points. All case studies included experiments with actual participants to test the ideas and verify the hypotheses. Chapters 3 to 7 describe the four studies. Each chapter starts with an overview and an explanation of how that chapter builds upon my previously conducted research and how it extends my pre-existing work. Chapter 3 describes the very first study where I experimented with *argument harvesting* which I defined as acquiring arguments with the help of a chatbot, and describes the notion of *concerns*, inspired by value-based argumentation [Bench-Capon, 2002a]. In Chapter 4, I move away from argument harvesting and instead use crowdsourcing for argument acquisition of the chatbot's knowledge base and start using chatbots for persuasive dialogues. Both case studies described in Chapters 3 and 4 focus on behaviour change (i.e. the chatbot aims to persuade the user to change their behaviour). Chapter 5 illustrates a method for acquiring a large corpus of arguments in graph form with the help of crowdsourcing that can be used as a chatbot knowledge

base. Chapter 6 describes the implementation of the chatbot that utilises the previously acquired argument graph as its knowledge base and the user's concerns to engage in persuasive dialogues with participants to convince them to change their stance. Chapter 7 describes a different way to build the chatbot's knowledge base and how to update it as more data comes in and presents a chatbot, using a different domain. In Chapter 8, I conclude the thesis by recapping the contributions and by giving future work directions.

The thesis also provides two software solutions for a chatbot: the chatbots can be deployed on Facebook or hosted on a stand-alone website. The code for the chatbots I implemented during the different case studies is available on GitHub [git, a, git, b, git, c, git, d] .

It should be noted that this thesis does not focus on validity and quality assessment of the arguments that were used for the chatbot's knowledge base. The chatbots were also not tested under real-world conditions. All participants were remunerated for participating in the experiments, and the majority followed the given instructions (i.e. to reply with arguments and not to ask the chatbot questions, not to spam the bot, etc.).

1.5 Thesis Contributions

To recapitulate, the following are the central presuppositions that I make and the main issues that I tackle in his thesis.

- I want the chatbot to allow the user to give free-text input, not limiting myself to menu-based chatbots where the user has to select arguments from a menu.
- I believe that it is possible to acquire the user arguments for the chatbot's knowledge base directly from people with the means of crowdsourcing

and not to limit myself to arguments from forums and other social media or websites.

- I acknowledge that crowdsourcing the chatbot's counterarguments poses the risk of acquiring low-quality arguments language and validity wise and therefore have to be excluded from the chatbot's knowledge base (manually if necessary). For specific topics, it is desired to use a hybrid approach where only the user arguments are crowdsourced and the chatbot's arguments manually crafted by domain experts.
- I wish to equip the chatbot with the ability to choose arguments that increase the persuasiveness of the dialogues tailored to the user. Although argumentation is not the only aspect of persuasion, I will only focus on argumentation and not on the user's personality traits and other psychological aspects.
- I note that to measure behaviour change, follow-up studies are necessary. However, I limit myself to measuring the change in the intention or stance by comparing the intention/stance after the chat with the chatbot to the one before the chat to evaluate the chat's persuasive effect.

Having these points in mind, my work and contributions can be summarised as follows:

- In this thesis, I propose a framework for developing persuasive chatbots, which includes methods for the acquisition of arguments, ways to structure the chatbot's knowledge base, and a strategy for the chatbot to select moves that increase the persuasiveness of the chat using concerns.
- Specifically, I focus on crowdsourced, simple arguments from ordinary people that are nevertheless full of meaning and importance. They are the kind of arguments that have been neglected in the formal as well

as informal argumentation literature. There is little literature on how to analyse this sort of argument and even less on how to acquire them. I show how to acquire such arguments and how they can be used to structure the knowledge base for a chatbot that can engage in persuasive dialogues with users.

- I further present the notion of *concerns* which is a simple concept according to which the chatbot can choose which arguments to select in a dialogue to make it more persuasive. I show that concerns addressed by arguments can be automatically detected and that by presenting arguments that address the user's concerns, the persuasiveness of the dialogue increases, and the user is more likely to change his or her stance on the topic in question.

I back up my research with case studies with actual participants to evaluate my chatbots. The UCL Ethics board formally approved all studies.

1.6 Publications

The work presented in this thesis has resulted in several publications:

1. L. A. Chalaguine, A. Hunter, F. Hamilton and H. H. Potts. Argument Harvesting Using Chatbots. In *7th International Conference on Computational Models of Argument*, pages 149-160, 2018. [Chalaguine et al., 2018] (Received best student paper prize)
Covers Chapter 3.
2. L. A. Chalaguine and A. Hunter. Chatbot Design for Argument Harvesting. In *Workshop on Systems and Algorithms for Formal Argumentation at COMMA'18*, pages 457-458, 2018. [Chalaguine and Hunter, 2018a]
Covers Chapter 3.

3. L. A. Chalaguine, A. Hunter, F. Hamilton and H. H. Potts. Impact of Argument Type and Concerns in Argumentation with a Chatbot. *31st International Conference on Tools with Artificial Intelligence*, pages 1557-1562, 2019. [Chalaguine et al., 2019]
Covers Chapter 4.
4. L. A. Chalaguine and A. Hunter. Knowledge Acquisition and Corpus for Argumentation-Based Chatbots. *3rd Workshop on Advances in Argumentation in Artificial Intelligence*, pages 1-14, 2019. [Chalaguine and Hunter, 2019]
Covers Chapter 5.
5. L. A. Chalaguine and A. Hunter. A Persuasive Chatbot using a Crowdsourced Argument Graph and Concerns. *8th International Conference on Computational Models of Argument*, pages 9-20, 2020. [Chalaguine and Hunter, 2020]
Covers Chapter 6.
6. L. A. Chalaguine and A. Hunter. Addressing Popular Concerns regarding COVID-19 Vaccination with a Persuasive Chatbot
Sixteenth European Conference on Symbolic & Quantitative Approaches to Reasoning with Uncertainty, pages 59-73, 2021. [Chalaguine and Hunter, 2021]
Covers Chapter 7.

Additionally, the COVID-19 vaccine chatbot was used by a researcher at the University of Alberta for his research purposes, as well as by a Master's Student from the Computer Science Department at UCL during his Master's Thesis.

Chapter 2

Literature Review

In this chapter, I briefly introduce abstract argumentation, overview the different argumentation strategies presented in the literature, talk about the notion of persuasion in computational argumentation, introduce value-based argumentation and show how values differ from concerns, discuss the drawbacks of menu based persuasion techniques in automated persuasion, discuss argument acquisition techniques, and present the current literature on chatbots that engage in argumentative dialogues.

The empirical approach taken during my research is part of a trend in the field of computational argumentation for studies with participants [Hadoux and Hunter, 2019]. This includes studies that compare human reasoning and argumentation theory. For example, [Rahwan et al., 2010] and more recently [Cramer and Guillaume, 2019] conducted empirical cognitive studies that have tested how humans evaluate sets of arguments depending on the attack relation between them, and multiple other empirical studies have evaluated the correspondence between human reasoning and formalism from abstract, structured and dialogical argumentation [Cerutti et al., 2014, Hunter and Polberg, 2017, Polberg and Hunter, 2018, Rosenfeld and Kraus, 2016a]. It also includes studies comparing

a confrontational approach to argumentation with argumentation based on the appeal to friends, appeals to the group, or appeal to fun [Vargheese et al., 2013, Vargheese et al., 2016] and studies to investigate physiological responses of argumentation [Villata et al., 2017].

2.1 Dung’s Abstract Argumentation Framework

The assessment of a set of given arguments and their conclusions to establish their justification status is addressed in a principled and highly stylised form in the context of the theory of abstract argumentation frameworks [Atkinson et al., 2017]. The first and probably most influential proposal for abstract argumentation was introduced by Dung [Dung, 1995]. The underlying idea of argumentation frameworks (AFs) is to model and evaluate sets of arguments, called *extensions*, by considering how well they can be defended against other arguments that can attack and defeat them. Dung frameworks are directed graphs in which arguments (nodes) are related by binary attacks (arcs). So if there is an arc from A to B, then A attacks B and A is, therefore, a counterargument to B. Here, I recall the following basic concepts that were introduced by Dung in [Dung, 1995]:

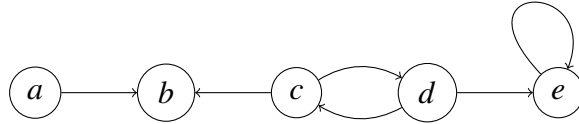
Definition 2.1 A Dung argumentation framework (AF) is a pair $AF = (A, R)$ where A is a *finite set* of arguments and R is a binary relation on A , i.e., $R \subset A \times A$ is an attack relation on the arguments A . For two arguments a, b the statement $(a, b) \in R$ means that a attacks b .

Example 2.1 Let (A, R) be an argumentation framework AF with

$$A = \{ a, b, c, d, e \}$$

$$R = \{ (a, b), (c, b), (c, d), (d, c), (d, e), (e, e) \}$$

This argumentation framework AF is represented by the following graph



A acceptability semantics identifies a set of arguments that can survive the conflicts encoded by the attack relation R . This produces accepted subsets of arguments (extensions) that accord to numerous stances one may take given the available arguments. The result is, therefore, arguments that are accepted, rejected, or neither. The internal structure of the arguments is not considered. Whether an argument is accepted or rejected, hence, is evaluated by considering whether or not it is able to be defended from attack from other arguments in that set.

Definition 2.2 Let (A,R) be an AF and let X be in A . For $S \subseteq A$ I say that:

- for any $S \subseteq A$, X is acceptable w.r.t. S iff for every Y that attacks X , there is a $Z \in S$ that attacks Y (in which case Z is said to defend X)
- S is conflict free iff no two arguments in S attack each other. Then for any conflict free $S \subseteq A$, S is an extension that is:
 - *admissible* if every argument in S is acceptable w.r.t. S ;
 - *complete* if it is admissible and every argument acceptable w.r.t. S is in S ;
 - *preferred* if it is a maximal under set inclusion complete extension;
 - *grounded* if it is the minimal under set inclusion complete extension;

Example 2.2 Continuing the previous example, the AF has the following extensions:

- *admissible*: $\{ \{ \}, \{a\}, \{c\}, \{d\}, \{a,c\}, \{a,d\} \}$. Note, that the empty set is always an admissible extension.
- *complete*: $\{ \{a\}, \{a,c\}, \{a,d\} \}$.

- *preferred*: $\{\{a,c\}, \{a,d\}\}$.
- *grounded*: $\{a\}$.

Abstract argumentation semantics [Dung, 1995] allows us to determine which extensions are able to survive the conflict together and can be regarded as collectively acceptable [Atkinson et al., 2017]. Treating arguments as entirely abstract allows for capturing a wide variety of situations which gives Dung’s general abstract framework a wide range of potential applications, including legal argumentation, medical argumentation, and eGovernment [Atkinson et al., 2017]. Numerous proposals have been made that investigate variants and developments of Dung’s proposal (see [Baroni et al., 2018] for a comprehensive review), one of which is bipolar argumentation frameworks [Cayrol and Lagasquie-Schiex, 2005a]. Bipolar argumentation frameworks can be used to express both positive and negative relations between arguments and hence introduce a support relationship. During my research, I focused on attacks. However, support relationships will be discussed in Chapter 8 as part of my future work.

In this work, I will use the definition of an argument graph to establish attack relationships between arguments. However, I do not use dialectical semantics to establish which arguments win and which arguments lose, given that it requires knowledge of the target audience, which we do not have, as well as some sort of quality assessment to ensure the integrity of the argumentation framework, which is not within the scope of this thesis. In the next section, however, I give an overview of some works who do consider some of these issues.

2.2 Overview of Argumentation Strategies

For a long time, the literature on dialogical argumentation has mainly focused on defining protocols that can capture the complex structure of real-life argumentation-based interactions (e.g. [Amgoud et al., 2000a, Amgoud et al., 2000b, Sadri et al., 2001]). For example, [Hitchcock et al., 2001] used dialogue-game frameworks to allow the participating agents to assert statements in the dialogue and to respond to statements made by other participants. The protocol defines what locutions are possible and the circumstances under which they may be used. Although such argumentation protocols present an agent with a communications language and the syntax for its use, it does not prescribe when specific locutions should be used by an agent [McBurney et al., 2003]. Since a protocol is not sufficient to generate an automatic discussion between agents, more recently, research interest has shifted to the development of argumentation strategies.

And strategies in argumentation-based dialogues are just like strategies in other areas of life and depend on many things, including the goal, the audience, and available knowledge (e.g. assumptions about beliefs or preferences of the audience). There is no strategy that fits all. The development of a strategy also depends on available resources. If, for example, a marketing agency has sold certain customers something before and has access to their records, they will be able to come up with a much better strategy adapted to the individual than if they did not have such information readily available. The same applies to dialogue strategies. If one has access to previous dialogues and knows the outcome of these dialogues, these can be used to train a probabilistic model. Most proposals for strategies using probability theory (for examples, see [Rienstra et al., 2013, Hadjinikolis et al., 2013, Hunter and Thimm, 2016, Hunter, 2014b, Hadoux et al., 2015]) either assume that the participant has a model of its opponents, which may be constructed on the basis of a participant's

accumulated dialogue experience, or assume that the dialogue is highly structured in order to design optimisation techniques. The presence of this data is, however, seldom the case, making the use of such techniques in a free-text setting very complicated, if not impossible (without first discovering artificial general intelligence).

Strategies that deal specifically with persuasion are presented in the works of Elizabeth Black [Black et al., 2014, Black et al., 2017]. In [Black et al., 2014] a planning system is used by the persuader to optimise the choice of arguments based on belief in premises. However, this requires the persuader to have a *model* of the persuadee, which assigns probabilities to sets of beliefs, representing the likelihood that each set is the persuadee's actual beliefs. This assumes a finite set of beliefs, a finite set of arguments, and some (even if uncertain) knowledge about the persuadee's beliefs. Such a strategy could be implemented in a menu-based set-up where arguments are presented to a user in the form of a menu. Still, it would not be feasible in a system that does not have any prior knowledge of the persuadee and allows free-text input that requires natural language understanding.

In [Black et al., 2017] the authors use an automated planning approach for persuasion that accounts for the uncertainty of the persuader's model of the persuadee by finding strategies that have a certain probability of guaranteed success no matter what arguments the persuadee presents. The authors tested their approach by considering the case where there is no knowledge about the likelihood of the possible persuadee models. However, this means that the persuader believes all possible opponent models are equally likely, i.e. the persuader's model of the persuadee contains all arguments available to the persuadee, each assigned with the same probability. Hence, this strategy is also unsuitable for a system that allows free-text input and knowledge of all arguments cannot be assumed. The strategy, however, accounts for the

fact that the persuadee may use knowledge from arguments presented by the persuader to construct new arguments unknown to the persuadee at the start of the dialogue. This is a very important aspect in systems that allow free-text input. One could argue that a chatbot also has some limited knowledge of the arguments the user might present based on the arguments in its knowledge base. The user can come up with new arguments incorporating information from the chatbot's arguments, which the user would not have known (or considered) at the beginning of the chat. However, it would take thousands of chats to collect a substantial amount of arguments (and different ways of rephrasing them) and extensive qualitative analysis of the chats before one could start thinking about this approach. Other works like [Murphy et al., 2016] and [Murphy et al., 2018] also assume that all arguments are known, and in [Hadjinikolis et al., 2013] the authors assume that previous dialogues are available.

The application of machine learning is another promising approach to developing more sophisticated strategies, such as using reinforcement learning [Huang and Lin, 2007, Monteserin and Amandi, 2013, Rosenfeld and Kraus, 2016b, Alahmari et al., 2017, Rach et al., 2018, Katsumi et al., 2018] and transfer learning [Rosenfeld and Kraus, 2016a]. The research by Rosenfeld and Kraus will be discussed in more detail in Chapter 6. I will also use machine learning and natural language processing, however, not to predict strategical moves but to allow the chatbot to understand the free-text user input. Reinforcement learning for the type of dialogues I am interested in, where arguments are not atomic entities, would require a lot of labelled training data in the forms of dialogues in natural language, which is not feasible.

In value-based argumentation proposed by Trevor Bench-Capon [Bench-Capon, 2002a] the audience's values are considered. A value is initially considered as a moral or ethical principle that is promoted by an argument (as seen in the example given in [Bench-Capon, 2002a] - *life vs property*).

However, as considered in [Atkinson, 2006a], the notion of a value promoted by an argument can be more diverse and used to capture the general goals of an agent (as in the example given in [Modgil, 2006] where the values considered were *cost* and *safety*). This led to the proposal of using the term *concern* rather than *value* by certain researchers and utilising the notion of concerns in persuasion systems. For example, the persuasion system presented in [Hadoux and Hunter, 2019] modelled the concerns of the user in order to choose moves. This work, as well as value-based argumentation and the notion of concerns, will be discussed in detail in Chapters 3 and 4.

To sum up, research on protocols and strategies is usually limited to developing formal frameworks that make strict assumptions and hence strongly limit their use in real-world applications. Strategies can be divided into local and global approaches. Global approaches consider *all* possible dialogues between the agents when deciding what move to take. In contrast, local approaches do not model all the possible combinations of decisions and outcomes but only plan one move ahead. It is evident why global strategies are impossible to implement in a real-world setting that allows free user input given that such a strategy is not scalable and assumes that the system is aware of all possible arguments the user might give. Local strategies are much better suited for dynamic systems where the user can give any response they want and will be discussed in more detail below.

2.3 On the Notion of Persuasion

In the scope of this thesis I am using Walton and Krabbe's definition of persuasion given in [Walton and Krabbe, 1995]. Six main dialogue types were identified: persuasion, negotiation, inquiry, deliberation, information seeking and eristic. Dialogue types can be identified by specifying (1) the initial situation, (2) the overall (collective) goal, and (3) the individual aims of

the participants (which may differ or even conflict) [Atkinson et al., 2020]. A summary of these defining characteristics for the six dialogue types is shown in Table 2.1 [Atkinson, 2005].

Negotiation is a dialogue between two or more people or parties intended to reach a beneficial outcome over one or more issues where a conflict exists with respect to at least one of these issues. For *inquiry* dialogues, the initial situation is that both agents are not certain of some proposition and the individual and collective goals are the same: to determine whether or not to follow the proposition. *Deliberation* is generally held to concern actions: initially, both agents are unsure whether or not to take a certain action, and individually and collectively, they wish to come to an agreement as to whether or not to take that action [Atkinson et al., 2012]. *Information seeking* aims at transferring information, and *eristic* is the verbal substitute for a fight. In *persuasion* one agent will argue that a proposition is true or that some action should be done in order that the other agent will come to an agreement. However, with regards to individual goals, persuasion is *asymmetric*: the persuader wishes to convince the persuadee, whereas the persuadee wishes to explore the possibility that its current opinion should be revised in the light of information known to the persuader [Atkinson et al., 2012]. This classification is not meant to be exhaustive and leaves room for dialogues of mixed type [Prakken, 2006]. However, for the scope of this thesis, I would like to emphasise that I will use the notion of persuasion as described above. A different case of persuasion, which I am not considering in my thesis, is a dispute. In this case, the persuadee also wishes to convince the other agent that its own original position is correct so that its individual goal is now that the other should believe that the proposition is not true and that the concerned action should not be done.

The distinction between the different types of argumentative dialogues is impor-

Table 2.1: Description of six argument types as defined by Doug Walton. Table taken from [Atkinson, 2005].

Type	Initial situation	Main goal of dialogue	Participants' aim
Persuasion	Conflicting points of view	Resolution of such conflicts by verbal means	Persuade the other(s)
Negotiation	Conflict of interests and need for cooperation	Making a deal	Get the best out of it for oneself
Inquiry	General ignorance	Growth of knowledge and agreement	Find a proof or destroy one
Info-seeking	Personal ignorance	Spreading knowledge and revealing positions	Gain, pass on, show or hide personal knowledge
Deliberation	Need for action	Reach a decision	Influence the outcome
Eristic dialogue	Conflict and antagonism	Reaching an accommodation in a relationship	Strike the other party and win in the eyes of onlookers

tant, given that each dialogue type requires a different framework. Some works that seem to be focused on persuasion actually focus on deliberation [Black and Atkinson, 2011], practical reasoning [Atkinson et al., 2005, Atkinson, 2006a] and inquiry [Riley et al., 2011] where agents aim to reach an agreement on how to act in order to achieve a particular goal. This is very useful for collective decision-making, where all arguments are known and can be evaluated with the help of an argument graph where each argument has a strength associated with it. However, such strategies are not suitable in an asymmetric setting where agents are not collaborative. Negotiation and deliberation take the views of both parties into account to try to find a solution that satisfies both parties, whereas in “pure” persuasion, the user has to be persuaded to take a certain action (e.g. get a vaccine), and the persuader is not interested in an alternative solution. Some users may ask questions and hence engage in information seeking during the dialogue. Addressing this was not within the scope of the thesis but was addressed by the aforementioned Master’s student who implemented a question-answering (QA) module, which will be further discussed in Chapter 8, as well as some more recent work by [Fazzinga et al., 2021] on the same topic of COVID-19 vaccinations.

Whilst persuasion and deliberation may have some similarities, they are different, as can be seen in the two example scenarios presented in [Atkinson et al., 2013]. In both dialogues, arguments are put forward that take the form of options presented that should be accepted on the basis of having certain features. Whether they have these features is intended to be an objective matter, whereas whether these features make the option desirable is a subjective matter about which agents may rationally agree or disagree. The two dialogues are very similar, using the same speech acts, similar arguments, and the underlying views of the participants do not change between the dialogues. But there is a crucial difference: Harry wants to go to the Thai Palace and is committed

to making a case for it in the persuasion dialogue. He advances all the pro arguments while George and Jane make objections (same as during a chat with a chatbot where the chatbot advances pro arguments and the users raise objections). Their objections indicate the points on which they wish to be convinced. Whereas in the deliberation scenario, no one is committed to any of the proposals, although they may have their *preferences*. Several options emerge neutrally, without any commitment on the part of the person introducing them. Then several criteria emerge together with how the options stand in relation to these criteria. Finally, all the options are assessed according to the criteria that have been chosen in the discussion, and the best fit is chosen by mutual agreement.

2.4 The Notion of Concerns in Computational Argumentation & Persuasion

A concern is something that is important to an agent. It may be something that she wants to maintain (for example, the satisfaction and taste of foods high in cholesterol), it may be something she wants to achieve (for example, losing weight), or it could be something she is worried about or scared of (like potential side-effects of a medication or vaccine). Despite the notion of concerns being mentioned for the first time in [Hadoux and Hunter, 2019], my research was conducted in parallel. In fact, in my first empirical study, described in this chapter and resulted in a publication [Chalaguine et al., 2018], I already used concerns, which at that time I still referred to as *values*, inspired by value-based argumentation frameworks (VAF) [Bench-Capon, 2002a]. However, in that paper, I already explained how my definition of values differs from the traditional definition, and I hence decided to use the term *concern* for my subsequent research. Nevertheless, I will give an overview of VAF, given that they motivated my use of concerns and are an essential

concept in argumentation.

2.4.1 Value-Based Argumentation

As mentioned above, extensive theoretical research has been done on Dung's abstract argumentation theory, and numerous subsequent works have extended the basic frameworks. One example is Value-Based Argumentation Frameworks developed by Bench-Capon. He argues that Dung's framework may be appropriate in the sort of argumentation where I have arguments that another argument can counter in such a way that the first argument has to be rejected. This (standard) argumentation framework is suitable for arguing about matters of fact or logical reasoning. However, as he emphasises in [Bench-Capon, 2002a], in many more interesting situations where humans are called upon to reason, it seems that finding an argument is not enough. There are situations where an argument may attack another without defeating it, and the notion of "validity" is not appropriate. In such cases of disagreement, particularly in practical reasoning, it is impossible to demonstrate conclusively that either party is wrong. The role of an argument then would be to persuade rather than to prove, demonstrate or refute [Bench-Capon, 2003].

According to him, persuasion in cases involving practical reasoning relies on recognising that the strength of an argument depends on the social values that it advances and whether the attack of one argument on another depends on the comparative strength of the values advanced by the arguments concerned. To represent this, he extends the argumentation framework to include the notion of value.

Definition 2.3 A value-based argumentation framework (VAF) is a 5-tuple:

$$VAF = (A, R, V, val, valpref)$$

Where A , and R are as for a standard argumentation framework, V is a non-

empty set of values, val is a function which maps from elements of A to elements of V , and $valpref$ is a preference relation (transitive, irreflexive and asymmetric) on $V \times V$. I say that an argument A relates to value v if accepting A promotes or defends v ; the value in question is given by $val(A)$. For every $A \in AF$, $val(A) \in V$.

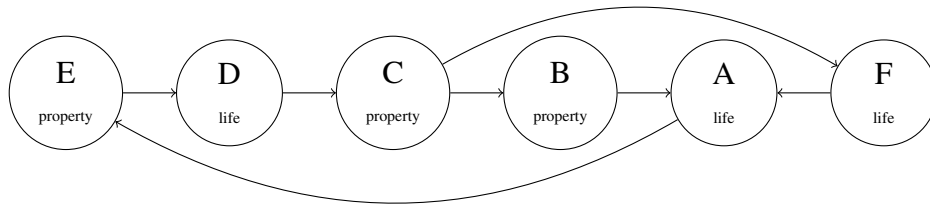
The purpose of extending the AF was to distinguish between one argument attacking another and that attack succeeding so that the attacked argument is defeated. Bench-Capon, therefore, introduced a definition of defeat:

Definition 2.4 Let $(A, R, V, val, valpref)$ be a VAF. An argument $X \in A$ defeats an argument $Y \in A$ iff both $attacks(X, Y)$ and not $valpref(val(Y), val(X))$.

In order to exemplify the notions of acceptance in VAFs, I consider the example from [Bench-Capon, 2002a] which I summarise next.

Example 2.3 Hal, a diabetic, loses his insulin in an accident through no fault of his own. Before collapsing into a coma, he rushes to the house of Carla, another diabetic. She is not at home, but Hal enters her house and uses some of her insulin. Was Hal justified, and does Carla have a right to compensation? The first argument A is that Hal is warranted since a person has the privilege to use the property of others to save their life. The second argument B is that it is wrong to infringe the property rights of another. Argument C denotes that Carla's rights have not been infringed. Argument D stands for the fact that if Hal were too poor to compensate Carla, he should nonetheless be allowed to take the insulin. Argument E denotes that poverty is no defence for theft. The last argument F means that Hal is endangering Carla's life. Arguments A; D and F are based on the value that life is important (life). Whereas arguments B; C and E are based on the value that property owners should be able to enjoy their property (property). The VAF is represented in the graph shown in Figure 2.1.

Figure 2.1: Argument Graph for Example 2.3



If we assume that *life* has a higher value than *property*, some admissible extensions of the VAF are $\{B, D\}$, $\{D, E, F\}$, $\{B, D, F\}$, $\{B, D, E, F\}$. The preferred extension from the VAF is $\{B, D, E, F\}$. If I would value *property* more than *life*, the preferred extension would be $\{A, C, E\}$.

I believe that VAFs give a more realistic picture of real-life argumentation than Dung’s frameworks because they explain *why* people posit certain arguments. In other words, they capture people’s *motivations*. However, as demonstrated in Example 3.1, VAFs assume that the agent sticks to one (or multiple [Kaci and van der Torre, 2008]) value throughout the whole argumentation. In traditional VAFs two opposing parties give arguments that promote different values according to their value preference. In the diabetic example described above, one of the parties gives arguments that promote the value “life” while the other party counters those with the value that promotes “property”. Argument and counterargument in a VAF therefore promote *different* values. VAF also assumes that each agent sticks to *one* value throughout the whole dialogue.

Despite giving a lot of theory on VAFs none of those papers that apply VAFs to specific examples explains where the values come from and according to what rules they should be chosen. Bench-Capons work has been applied mainly in the *AI and Law* domain. Therefore it is not surprising that his choice of values reflects the legal motivations of the opposing parties. In [Atkinson, 2006b]

Atkinson discusses the application of VAF to the domain of eDemocracy. She gives an example of the Government's hypothetical justification for invading Iraq. The values involved in the justification are *world security*, *human rights* and *public trust* which are all taken from the legal domain.

Two further examples of applications of VAFs, which are applied to more realistic real-life situations, are given in [Modgil, 2006] and [Besnard et al., 2008]. In [Modgil, 2006] the example involves choosing the appropriate treatment for a patient threatened by blood clotting. The values considered were *safety* and *cost*. In [Besnard et al., 2008] the study examined a problem facing a department's university heads and reflected the need to balance costs and departmental and individual interests. The values in question were *happiness*, *publication*, *experience* and *esteem*. *Happiness* and *experience* are the values relative to individual students, while the other two are relative to the department. No definition of values was given in [Besnard et al., 2008] or guidance on choosing those values. While the values *happiness*, *experience* and *esteem* (meaning enhancement of reputation) intuitively seem to describe some sort of motivation, *publication* does not seem to be a value. *Fame* or *reputation* seem more suitable in this context, while the publication is what would cause the value to "materialise".

2.4.2 Concerns

Atkinson, to my knowledge, is the only one who discusses values in more detail in [Atkinson and Wyner, 2013]. She describes values as social interests that a person/agent wishes to promote. She stresses that values are often referred to in everyday reading material such as newspapers and that people widely recognise and understand the concept of values. As an example, she gives the French national motto "Liberté, égalité, fraternité" (liberty, equality, fraternity). One could therefore argue that values are of a higher order than concerns.

Let's take liberty as an example - a person might be highly reluctant to take a newly developed vaccine and believe that an obligatory vaccination programme would infringe her liberty. However, her concerns about not taking the vaccine might include the vaccine's fast development and potential (long-term) side effects.

Concerns are more suitable for argumentative dialogues where one party strictly takes the role of the persuader and the other party as the persuadee (for example, in a doctor-patient setting). In these situations, one party argues for or against a certain behaviour while the other defends her actions, without trying to persuade the persuader of the opposite. In contrast to VAFs, the persuader's argument does not have a different value from the argument of the persuadee, but *addresses* the concern of the persuadee's argument. In traditional applications of VAFs, each agent usually holds the same value during the whole debate. For example, one party might hold the value *life* to justify breaking into a house to steal some food because they were starving, whereas the other party might hold the value *property*. The parties are, of course, not limited to just one value, however in VFA the arguments of one party and the counterarguments of the other are always promoting different values. In my research, however, the arguments of both, the persuader and the persuadee, have the same concern assignment.

Arguments can raise or address various concerns for the persuadee that need to be accounted for. A persuader might present a perfectly valid argument, e.g. "Meat consumption has a negative impact on the environment as it causes deforestation as huge tracts of rainforest are burned for pasture". The persuadee might not even disagree with this argument, however, if she is not concerned about the environment, this argument may not impact her intention to change her behaviour. If, however, the persuadee is concerned about her health, then the argument "Some meats are high in saturated fat. Eating a lot of saturated

fat can raise cholesterol levels, which raises your risk of heart disease” is more likely to change her intention to consume less meat. Whilst this is a simple and intuitive idea, there is a lack of a general framework for using concerns in making strategic choices to move in the way suggested by the above example [Hadoux and Hunter, 2019].

The concerns differ from domain to domain, and one of my research questions is whether I can automatically identify the most suitable ones. This will be discussed in the following chapters. The actual concerns one might consider, and their scope and granularity, depends on the domain and the application [Hadoux and Hunter, 2019]. For example, it may be appropriate to group “finer-grained” concerns like *fossil fuel usage, effluent waste, water and land consumption* into the concern *environment*. However, I assume that they are atomic, that the set is sufficient to be able to label all the possible arguments that are included in an argument graph and that various agents, independently of each other, identify similar concerns, i.e. that different people will assign the same or a similar concern to a given argument. When an argument is labelled with a type of concern, it is meant to denote that the argument has an impact on that concern, irrespective of whether that impact may be positive or negative [Hadoux and Hunter, 2019]

2.5 Automated Persuasion using Knowledge Graphs and Menu-Based User Input

Several menu-based persuasion systems were presented in the literature and used in empirical studies which also use argument graphs as knowledge base. In [Hadoux and Hunter, 2019] the authors implemented an automated persuasion system (APS), with which the user could interact by selecting arguments from a menu. The authors distinguish between two types of move: a *posit*

move, which is a set of arguments chosen by the system; and a *menu* move, which is a set of arguments by the user from a set of arguments provided by the system. The menu move is used as a way for the user to give his or her input into the discussion, as the APS cannot accept free-text input from the user.

The APS' knowledge base was constructed in the form of an argument graph which the researchers carefully researched with the help of a web search on the pros and cons of city cycling. From that search, they manually identified a number of arguments and attacks between them. They ended up with a total of 51 arguments which they manually "typed" with different concerns including *health, comfort, and safety*. The authors acknowledge that the difficulty of this step lies in finding a set of concerns that is sufficiently broad to cover all possibilities but not too specific to not type an argument with too many overlapping concerns.

These systems have several disadvantages: Firstly, the user is limited to the choice of possible counterarguments presented by the APS, which might not include the user's preferred choice. This might limit the argumentative dialogue's persuasive effect and deny the APS the opportunity to acquire novel arguments on that topic that were not collected during the acquisition phase of the graph and could have been used to extend the existing argument graph.

Secondly, the manual acquisition of the argument graph introduces the bias and beliefs of the researchers or the sources they used into the knowledge base of the APS. The information on why people engage (or not engage) in a particular behaviour is not always available on the internet or only represents the point of view of one specific demographic. Relying on limited information of the internet hence makes it impossible to accommodate free-text input since the

APS has no information in advance on how the users will potentially reply and might not be able to identify the concern of the incoming user argument and hence to provide a suitable counterargument from its knowledge base.

Another example is presented in [Hunter et al., 2020] where an APS was used to persuade the user to accept the chatbot's stance on university fees in the UK. The argument graphs that were used as the APS' knowledge base were hand-crafted and manually labelled. The APS also did not allow free-text input and was strictly following the arcs of the argument graph.

It should be noted that in both works, the authors stated that despite using the definition of an argument graph, they did not use dialectical semantics because they were not concerned with determining the arguments acceptable according to some normative principles but instead wished to model how persuasion may occur in scenarios where the participants are allowed the freedom of opinion and hence, do not need to adhere to any rationality principles. As mentioned before, several studies showed via empirical studies that humans do not follow the main principles behind these semantics [Polberg and Hunter, 2018, Rosenfeld and Kraus, 2016a]. The same applies to my research.

One advantage of menu-based approaches is the possibility of testing out different dialogue strategies. The aforementioned work used decision trees for that, which was formerly proposed in [Hadoux and Hunter, 2017]. The authors, however, acknowledge that despite being a helpful tool in artificial intelligence, they quickly become unmanageable in applications with a large number of possible outcomes. It should also be noted that compared to the previously described work, this work uses a global strategy when choosing the next dialogue move. As mentioned in previous sections, this is an unsuitable strategy for systems that allow free-text input.

2.6 Argument Acquisition for Knowledge Base Construction

Argument graphs are extensively studied in the computational argumentation literature. Their acquisition, however, tends to be neglected. Nevertheless, much research has been conducted on acquiring arguments from the web (usually not in graph form) and is generally referred to as *argument mining*. Argument mining exploits existing and develops new techniques from Machine Learning (ML) and Natural Language Processing (NLP); re-purposing and extending them to identify argument structures within text [Wells, 2014]. For an extensive overview of the latest research, please refer to [Lawrence and Reed, 2020, Cabrio and Villata, 2018]. Online generated discourse in forums or specific debating websites (e.g. *createdebate*¹ or *reddit*²) has also attracted research on argument mining. Threads from *reddit*, for example, have been used to create argument graphs for highlighting only the relevant arguments involved in a discussion [Pazienza et al., 2017] and assessment of persuasiveness [Tan et al., 2016]. Data from *Createdebate* was used as data source in several studies including [Swanson et al., 2015, Boltužić and Šnajder, 2015a, Hasan and Ng, 2014]. IBM’s *Debater* project [Levy et al., 2018] heavily relies on argument mining techniques and mines the arguments from published sources like Wikipedia.

Using forums for online discussions as a source for the development of automated agents sounds tempting due to the large repositories which contain a great deal of human knowledge on many topics. However, using threads from websites like *reddit* for a chatbot knowledge base raises several problems. Firstly, unless it is a prevalent topic, it can take months to acquire a substantial number of arguments and risk not collecting any. Secondly, debating web-

¹<http://www.createdebate.com/>

²<http://www.reddit.com/>

sites often only attract a small sample of the population, and the arguments, therefore, may not represent a wide variety of views. Thirdly, not all posts contain arguments. Often people share stories, ask or answer questions or make opinionated statements. Unmonitored online forums often contain noise in the form of inappropriate language, statements of opinion and inaccurate information and fake news rather than arguments (I have shown in a previous paper [Chalaguine and Schulz, 2017] that the dataset used by [Habernal and Gurevych, 2016a] was not a good one to study linguistic features of convincing arguments due to the noise in the data). Fourthly, long posts most likely contain several arguments, and individual arguments would have to be extracted with argument mining techniques. And lastly (and probably most importantly), the language of forum posts and short chatbot responses may vary, and there can be many ways of expressing the same thing using different words. It is unlikely that a forum will contain the same argument many times expressed differently.

Mined arguments from non-social media sources, like Wikipedia, introduce new problems. Those arguments are atomic and do not have any predefined relationships, unlike forums like Reddit, where a reply to a post could be seen as a counterargument. Given that I want to use the acquired arguments as a knowledge base for a chatbot that can engage in argumentative dialogues, I require an argument graph that contains a wide variety of arguments that represent both stances of the topic in question (arguments and counterarguments) and which are free of noise and ideally also free of inaccurate information and fake news.

Another approach to collecting arguments and constructing an argument graph, without the use of online discussion forums or extensive research, was conducted using Dialog-Based Online Argumentation (D-BAS) and is described in [Krauthoff et al., 2017]. Their resulting graph contains 265 arguments.

However, it should be noted that the researchers instructed the participants on how to counter previous arguments to obtain high-quality arguments and counterarguments. They also did not allow the repetition of arguments and motivated the participants to flag repetitions, as well as statements that should be revised, were off-topic or irrelevant, or abusive. Due to the highly regulated environment, the resulting argument graph is much better suitable as a chatbot knowledge base but, at the same time, not easily scalable. However, this was the first *crowdsourced* argument graph to my best knowledge. Inspired by this idea, I also relied on crowdsourcing when constructing argument graphs for my chatbots throughout my research.

Most chatbots are implemented using templates: for a specific question, the chatbot provides an answer from a list of possible answers. These are usually hand-coded, and the construction of chatbot knowledge bases is therefore time-consuming and difficult to adapt to new domains. There is limited research on fully automated chatbot knowledge acquisition. A suitable method for my research was proposed in [Huang et al., 2007]. It describes a process of using online discussion forums to extract chatbot knowledge by automatically extracting the titles of threads and their replies, creating <thread-title, reply> pairs. In this way, they constructed a knowledge base for a chatbot. These pairs, however, were not connected in a graph-like structure and the chatbot's purpose was to answer questions and not engage in an argumentative dialogue. Chatbots that use argumentation usually assume an existing knowledge base where the counterarguments can be drawn from or require researching the arguments and manually constructing the knowledge base. Climebot [Toniuc and Groza, 2017] (a conversational agent able to explain issues related to global warming), for example, relied on textual entailment to identify the best answer for a statement given by a human agent. The argumentative corpus from which the chatbot could choose was extracted from three debating sites.

2.7 Conceptualising Chatbots

Conversational agents that only understand text are also called *chatbots*. In my work, I allow the user to input his arguments freely while conversing with the chatbot. Hence, in this section, I will give an overview of chatbots with a focus on argumentative chatbots.

2.7.1 History

The interest in conversational technology has increased considerably since the 1960s and the development of the first chatbot. ELIZA was the first chatbot developed by MIT professor Joseph Weizenbaum in the 1960s [Weizenbaum, 1966], which aimed to act like a Rogerian psychotherapist using pattern matching and substitution of the previous user prompt to simulate a conversation. ELIZA was designed to imitate a therapist who would ask open-ended questions and respond with follow-ups which created an illusion of understanding. PARRY was another historic chatbot, developed in 1972 by Kenneth Colby, at Stanford University, which simulated a person with paranoid schizophrenia. ALICE (artificial linguistic internet computer entity) was developed in 1995 by Richard Wallace, which utilised the XML schema known as AI markup language (AIML), which helps specify conversation rules.

Broadly, chatbots can be classified into two groups: general-purpose and task-specific. The general-purpose chatbot is a multi-tasking agent and plays the role of a personal virtual assistant. In 2001, Microsoft developed the first personal assistant SmarterChild (replaced with Cortana in 2014), after which Siri by Apple followed in 2010, Amazons Alexa in 2015, Google Assistant in 2016, and Samsungs Bixby in 2017.

A task-specific chatbot focuses, as the name tells, only on one task, for example, answering common customer questions regarding a service or allowing

users to book flights. In this thesis, we are also dealing with task-specific chatbots.

2.7.2 Task-specific chatbots since 2016

Early in 2016, social media platforms like Facebook enabled developers to build a chatbot for their trademark or service so that customers could carry out some of their daily actions from inside their messaging platform. It was believed that this would revolutionise chatbot development, given Facebook made chatbots more accessible for businesses. However, according to a 2017 report, 70% of the 100.000+ bots on Facebook Messenger fail to fulfil simple user requests³. There are many reasons why chatbots did not take off in 2016 and analysing them is out of the scope of this thesis. However, chatbots have not entirely fallen short. They are still being used in industries like fintech, health care, mental health, sales and CRM, retail, education, government, and law. Applying advances in AI and NLP (natural language processing), they are making health care more easily accessible, shopping more personalised, and lawyers more efficient.

From the above-mentioned industries, behaviour change, one of the domains tackled in this thesis, is mostly related to health care and mental health. Some examples of chatbots in these domains are:

- **Health Care:** Babylon Health⁴ offers online medical consultation and health service. Florence⁵ acts as a personal nurse that can remind patients to take their pills. One of the most popular symptom assessment solutions on the market is Ada Health⁶ which can assess the user's health based on the indicated symptoms (for an academic pilot study which

³<https://tinyurl.com/tzz6uj6a>

⁴<https://www.babylonhealth.com/en-us/what-we-offer/chatbot>

⁵<https://www.florence.chat/>

⁶<https://ada.com/>

investigated the diagnostic quality of the chatbot, please see [Jungmann et al., 2019]).

- **Mental Health:** popular chatbots like Woebot⁷ and Wysa⁸ provide emotional support for people who suffer from depression or anxiety [Fitzpatrick et al., 2017, Prochaska et al., 2021, Inkster et al., 2018]. Another chatbot worth mentioning, given it has seen a 35% increase in traffic since the start of the pandemic, is Replika⁹. As many people were stuck alone at home during lockdowns, they have turned to AI for companionship. Replika avatars can be designed by users to be their friends, mentors, or even romantic partners.

To my knowledge, there are currently no chatbots publicly available that fall into the behaviour change domain. Interestingly a few years ago, there were a few that now have been disabled (e.g. Forksy¹⁰ a chatbot that could give you nutrition tips). Some recent academic works include a chatbot that helped people achieve healthier eating habits [Prasetyo et al., 2020] and a chatbot that helped to motivate smoking cessation [He et al., 2022].

None of these chatbots, however, use argumentation as a key component.

2.7.3 Examples of Argumentative Chatbots

There are also no publicly available chatbots that explicitly use argumentation. Several chatbots, however, were presented in academia. Some recent examples include:

Plant-Based Eliza [Zaal et al., 2017]: In this study, a computer-based dialogue system was designed to persuade users to switch to a plant-based

⁷<https://woebothealth.com/referral/?ref=website>

⁸<https://www.wysa.io/meet-wysa>

⁹<https://replika.ai/>

¹⁰<https://getforksy.com/>

diet which mainly used strategies from motivational interviewing (MI). Strictly speaking, it was not a chatbot, but one could argue that one could have been easily implemented. However, it did not allow free-text input. The system included an argumentation module that provided individually tailored arguments that addressed the individuals' arguments (beliefs) for not engaging in a plant-based diet. The authors do not provide any information in their paper on how these beliefs were identified.

Climebot [Toniuc and Groza, 2017]: *Climebot* was developed to explain issues related to global warming. It relied on textual entailment and an ontology to identify the best answer for a statement given by a human agent. The chatbot was hosted on API.AI (now part of Google's Dialogflow¹¹) and made use of the platform's pattern matching and machine learning algorithms which allowed the chatbot to accept free-text input. The argumentative corpus from which the chatbot could choose arguments from for the textual entailment module was extracted from three debating sites: *Debatepedia*¹², *Debate*¹³ and *ForAndAgainst*¹⁴.

Debbie [Rakshit et al., 2019] *Debbie* was developed to simply keep a good argument going. The aim of Debbie was not to persuade but to continue the argumentative dialogue as long as possible. Debbie's knowledge base consists of a subset of the qualitatively best arguments from the corpus created by Swanson et al [Swanson et al., 2015] which is a combination of online political debates, Internet Argument Corpus (IAC), [Walker et al., 2012] and dialogues from online debate forums and covers several topics (gay marriage, gun control and death penalty). Hence, Debbie can only talk about these three topics. A similarity score is used to retrieve the most appropriate counterargument.

¹¹<https://cloud.google.com/dialogflow>

¹²www.debatepedia.org

¹³www.debate.org

¹⁴www.forandagainst.com

Dave [Le et al., 2018] Similar to Debbie, *Dave*'s aim was also to keep the argumentative dialogue going. The main difference is that the authors also present a generative approach using a hierarchical recurrent (RNN) encoder-decoder architecture alongside a retrieval-based method. A newer version of the IAC corpus, a collection of 65k posts on 5.4k debate topics retrieved from `convinceme.net` [Abbott et al., 2016] was used as Dave's knowledge base for the retrieval-based system.

Neither Dave, nor Debbie are appropriate benchmarks for the chatbot I am proposing given their aim was to keep a conversation going, whereas mine is to convince people to change their stance.

Some earlier theoretical works on dialogical argumentative persuasion systems include the research by Floriana Grasso on dialogue games using health promotion [Grasso, 1998, Cawsey et al., 1999, Grasso et al., 2000, Grasso, 2003], dialogue management for persuasion [Andrews et al., 2008], reinforcement learning of negotiation dialogue policies [Georgila and Traum, 2011], and the use of persuasion systems in the healthy-eating domain [Mazzotta et al., 2007].

2.8 Conclusion

The literature presented in this chapter fails to deliver all the components required for an argumentative chatbot which allows free-text input the aim of which is to change the user's stance. Argumentation strategies usually assume some knowledge about the user, or that all arguments in the domain are known. Also, there is little work on what I called above "pure persuasion" where one party (the persuadee) has to be persuaded to take a certain action, and the persuader is not interested in an alternative solution and cannot be persuaded by the persuadee to accept their stance. Given this, VBF are not suitable

given both parties try to persuade each other and hence promote different values. A chatbot which aim is to persuade the user, however, wants to address the concerns of the user and hence the user's arguments and the chatbot's counterarguments have the same concern assignment.

I outlined above why existing approaches on using arguments from forums for online discussions as a source for the construction of argument graphs which can be used as the chatbot's knowledge base are not appropriate. The main reasons being that a certain topic might have never been discussed in anywhere online, and because the language of forum posts and short chatbot responses may vary. Existing implementation of argumentative chatbots are either menu-based or do not intend to persuade the user but rather just to keep the conversation going for as long as possible.

To address the issues outlined above, I will present my research in the following chapters.

Chapter 3

Argument Harvesting Using Chatbots

In this chapter, I give some background theory on concerns in computational argumentation and present an empirical study I conducted to investigate the perception of concerns amongst participants to motivate their use in computational argumentation. I also introduce the first chatbot that I implemented in order to acquire arguments about why people engage in certain (unhealthy) behaviour. At the beginning of my research, I focused on behaviour change applications with the goal of harvesting arguments that could then be used to construct an argument graph that a chatbot could potentially utilise as a knowledge base to persuade people to change their behaviour. The rest of the chapter is structured as follows: In Section 3.1, I introduce my research study; Section 3.2 presents some definitions; Section 3.3 presents the aim of the research and the hypotheses; Section 3.4 describes the chatbot architecture that was used for argument harvesting; Section 3.5 describes the experiments that were conducted throughout the study including their methodology and results, and in Section 3.6 I discuss and conclude my findings.

3.1 Introduction

In order to construct graphs using *real* arguments as opposed to theoretical, made-up scenarios, arguments must be acquired from real-life sources. A common approach to argument acquisition assumes a static resource available on the internet where the topic of interest is/was already discussed. This, however, raises several problems: firstly, what if no discussion platform for a particular topic exists? Secondly, even if it exists, what if not enough representative people contribute to the opinion exchange? Thirdly, such platforms do not consider the attributes of the individuals who posited the arguments. This is a drawback of other systems such as D-BAS [Krauthoff et al., 2016] that are more suitable for public argumentation or collective decision making where *all* existing arguments on a particular topic are of interest. In this specific study, I focused on behaviour change and therefore required a more individual approach. One solution is to use questionnaires or to interview people directly. That, however, may be a labour-intensive and expensive undertaking. To address these issues, I believe that it is possible to automate the process of argument acquisition using chatbots. As a proof of concept, in this chapter, I present a method focused on argument acquisition for behaviour change applications but which could be adapted to other application domains as well.

Human agents respond and converse with artificial agents in ways that, to some extent, mirror emotional and social discourse dynamics when discussing behavioural health [Bickmore et al., 2005]. Therefore, there is literature to suggest that using a chatbot to acquire user arguments on certain behaviour and address the problems of traditional argument acquisition is possible. As already shown by Weizenbaum, a chatbot that uses only generic questions can indeed encourage the user to talk about himself [Weizenbaum, 1966]. His chatbot *Eliza* simulated conversation by using pattern matching and pronoun substitution, giving users an illusion of understanding even though it had no

built-in knowledge. This indicates that generic questions may enable a chatbot to harvest arguments in diverse domains.

Prior to my work, no attempts to use a chatbot for argument acquisition have been made in the computational argumentation domain. In this chapter, I investigate the approach at *argument harvesting* which I define as acquiring arguments with the help of a chatbot. I further performed three experiments with crowdsourced participants to analyse the arguments' meaningfulness, concerns and relationships. The contribution of my work was threefold: first, I described a model for argument harvesting using a simple chatbot with little or no domain knowledge. Second, I showed that people who give the same or a semantically similar argument are most likely motivated by the same concern when positing it. I demonstrated that it is, therefore, possible to train a concern classifier to predict the concern of an argument. And third, I presented a method to cluster the harvested arguments by concern and semantic similarity in order to create several possibilities to counter a given argument automatically.

3.2 Concern Assignment and Suitability of Counterarguments

Let us consider as an example a woman who does not want to engage in physical exercise. I am concerned with the notions of *concern* of an argument and the *suitability* of a counterargument. The following example illustrates the two concepts: given the concerns for not exercising $V = \{family, comfort, dignity, wealth\}$, suppose a woman (the persuader) is trying to convince her female friend (the persuadee) to do more sports and gives the following argument:

Example 3.1

A1: “*Physical activity is healthy and you should therefore go to the gym more often.*” The persuadee, assuming she is rational, will not try to counter the fact that physical activity is healthy and will most likely *accept* that fact. She may, however, counter the conclusion (which action to take) with an argument that reflects her concerns for not engaging in physical activity. She might say: **A2:** “*I have no time because I have to look after my kids.*” In this case, the argument promotes the concern *family*.

To generalise this idea, I give the following definition for concerns which delineates how I can assign a concern to an argument.

Definition 3.1 A **concern assignment** by an agent to an argument *A* is a concern for the agent if the agent were to posit *A*.

In the above definition, I use the phrase “if the agent were to posit *A*” because I will investigate how individuals assign concerns to arguments independently of a specific dialogue in Section 3.4.2.

I believe a counterargument should have the same concern assignment as the argument it attacks to increase the probability of it successfully defeating it. Continuing with the example above, the persuader would respect the concern *family* and give a counterargument **A3** that attacks **A2** but respects the concern *family*. For example **A3:** “*You could incorporate your children into your exercise routine. Like going rollerblading in the park or swimming.*” So **A3** attacks **A2** while respecting the same concern and still pursues the initial intention of persuading the persuadee to do more sports. This does not mean that the persuadee has to agree with the given counterargument. It merely means that the counterargument can be given as a *suitable* counterargument to the previously posited argument. I define the notion of *suitability* of a counterargument next:

Definition 3.2 Let A be an argument and let CA be a counterargument that attacks A . CA is a **suitable** counterargument to A iff A has a concern assignment C and CA has a concern assignment C' such that $C = C'$.

Given these notions of concern and suitability, I wanted to test a hypothesis using an empirical study described in the next section.

3.3 Hypotheses

In this study, I made the first step toward argument harvesting. I chose *attitudes of women to participate in sport* as a case study. I have developed a chatbot that harvests arguments and concerns from women on why they do not engage in (more) physical activity. The chatbot also asked them to provide suitable counterarguments to their arguments (more on the dialogue protocol in the next section). Each argument, therefore, has a concern and a counterargument. Given this, I wanted to test three issues: first, whether different people were motivated by the same concern if giving the same or semantically similar argument. Second, whether my chatbot was capable of harvesting meaningful arguments, i.e. those considered to be appropriate arguments by sufficiently many participants from the people group the argument was harvested from. Third, whether I could automatically match an argument with more suitable counterarguments and therefore create more possibilities to counter a certain argument. I summarise these points in the following three hypotheses:

- H1** The majority of people that are exposed to, but do not necessarily posit, the same argument assign to it the same or similar concern, therefore making it possible to predict the concern of an argument.
- H2** A domain-neutral chatbot, with little or no domain-specific knowledge, and by giving general responses, can acquire arguments that are perceived as meaningful by the people group the arguments were harvested

from.

H3 Given arguments semantically similar in meaning with the same concern, counterarguments are interchangeable, making it possible to use the counterargument of one argument as a counterargument to another argument.

In the remainder of this chapter, I describe the design of my chatbot that was used for argument harvesting and explain the experiments conducted with the harvested arguments in order to test my hypotheses.

3.4 Chatbot Design for Argument Harvesting

Messaging has become the most widely used communication layer on mobile platforms during the last few years, with Facebook Messenger (FM) being the most popular messaging application¹. FM is a free instant messaging service and software application that lets Facebook users chat with other users (or chatbots) on the main website and the mobile app. For building chatbots, the Messenger Send API gives the ability to send and receive messages. Due to the popularity of FM and the free API that Facebook provides, I decided to use FM as the platform to deploy my chatbot.

I created an application called *ArgHealthBot* to which users can send messages. The application is linked to a Facebook page with a *Send Message* button. The page also displayed a link to a website that contained the terms and conditions of the chatbot and stated that I received ethical approval for my study and a short description of the current experiment. For the screenshots of the website and the application, see Appendices G and H [app, a]. When users clicked on *Send Message*, an FM window popped up, allowing them to send private messages to the application to which the chatbot is connected. The chatbot

¹1.3 billion active users as of December 2017

code was written in the Python programming language and consisted of a Flask server and the text-processing code. The server code communicated with the Send API, and the text-processing code processes the incoming messages from users and sends appropriate responses.

The dialogue protocol was the following: after the participant initiated the chat and consented to continue with the experiment, the chatbot asked to provide a reason for why she was not engaging in (more) physical activity, to which the participant answered with an argument (A1). If the chatbot considered the answer too short (less than 12 words), it asked to expand on the given argument. The chatbot queried the participant to expand on the argument only once. The expansion of the argument (if there was one) was added to the initial argument, and the complete, harvested argument was added to the argument database. The pseudo-code and a description of the algorithm for query generation (asking to expand on a given answer) are shown in Algorithm 1.

To assign concerns to the arguments, I needed a set of concerns to choose from. I used the list of personal concerns from Scott Jeffrey² as a reference and pragmatically chose concerns that I found suitable. The concerns were: *responsibility, comfort, dignity, satisfaction, relaxation, family, friendship, professionalism, productivity, wealth, knowledge, fun, recreation, ambition* and *safety*. The chatbot presented the user with the list of concerns after she provided an argument and asked to choose the one she most associated with her argument.

The chatbot then asked what the user would recommend to a friend with the same problem. This was the counterargument to the previously given argument (CA1). The chatbot picked up on that and asked why the user was not following her own advice. The user answered with another argument (A2).

²<https://scottjeffrey.com/core-concerns-list/>

Algorithm 1 Chatbot question generation

```
1: procedure GENERATEQUESTION(user_response, stage)
2:   if length(user_response) < 13 and stage[-1] ≥ 0 then generate-
   Query(user_response)
3:   else analyse_stage(stage) ▷ returns appropriate chatbot answer according to
   stage in chat
   return chatbot_answer
4:
5: procedure GENERATEQUERY(user_response)
6:   time_words = ['time', 'busy']
7:   negations = neg = ['not', 'dont', 'arent', 'cant', 'wouldnt', 'isnt', 'don't',
   'aren't', 'can't', 'wouldn't', 'isn't']
8:   explanations = ['that is', 'because', 'due to', 'since', 'that's', 'thats', 'hence',
   'therefore']
9:
10:  ▷ The questions are examples. They were randomly chosen from a list with
   several similar possibilities
11:  if length(user_response) is 1 then
12:  chatbot_answer = "How does this hinder you doing exercise?"
13:  else if user_response contains word in time_words then
14:  chatbot_answer = "I understand you have no time, but can you expand on that
   please. Why not?"
15:  else if user_response contains word in negations then
16:  chatbot_answer = "Why not?"
17:  else if user_response contains word in explanations then
18:  chatbot_answer = "Could you go into more detail? Why do you say that?"
19:  else if length(user_response) ≤ 5 then
20:  chatbot_answer = "Why?"
21:  else if length(user_response) > 5 then
22:  chatbot_answer = "Elaborate please...how does this hinder you from doing more
   exercise?"
   return chatbot_answer
```

The chatbot asked again what she would advise a friend with the same problem (CA2). After harvesting two argument-counterargument pairs, the chatbot asked the participant whether she wanted to continue or end the chat. My chatbot, therefore, harvested a minimum of two argument-counterargument pairs $\{(A1, CA1), (A2, CA2)\}$.

3.5 Experiments

In this section, I describe how I collected the arguments concerning women's participation in sports via argument harvesting (AH) and the experiments conducted with the harvested arguments. For each experiment, I give the purpose, the methods used, the results and the conclusion of my findings. The participants for all experiments were recruited via *Prolific*³, which is an online recruiting platform for scientific research studies. For each experiment, I recruited from three disjoint groups: students (aged 18-25 and no children), women with children (aged 18-40 and not students) and women without children (aged 18-40 and not students), in the following referred to as the *student*, *kids* and *nokids* groups respectively. I opted for this division to get a wider spectrum of different arguments from different people groups, or *audiences*. For each experiment, I evaluate how the audience perceived the arguments it is meant for, based on the assumption that a particular argument is addressed to a specific audience [Bench-Capon, 2002b]. The general prerequisites for participating in my study were being female, over 18 and engaging in less than 150 minutes of physical exercise per week. For the argument harvesting, I required the participants to have a Facebook account to chat with the chatbot. For the experiments, Google Forms were used.

³<https://www.prolific.ac/>

3.5.1 Argument Harvesting

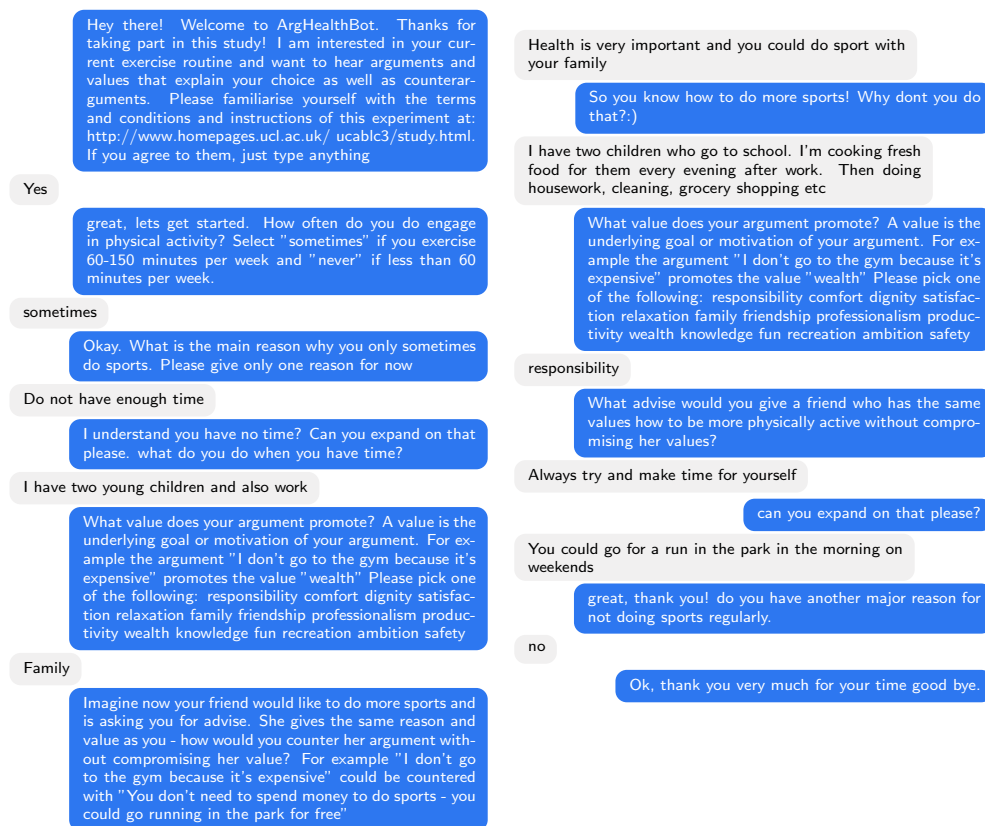
I conducted two rounds of argument harvesting (referred to as AH1 and AH2). In AH1, I used my chatbot to harvest arguments and their associated concerns and counterarguments from the three participant groups. In AH2, I harvested arguments and counterarguments without their concerns. The harvested arguments were used for making observations about the concerns, for training the concern-classifier described in Experiment I, and for evaluating the argument clustering algorithm in Experiment III.

For AH1, I recruited 30 participants for the *student* group, 30 for the *kids* group and 50 for the *nokids* group. The women who participated in the study and agreed to chat with the chatbot initiated the conversations, and the chat followed the dialogue protocol described in the previous section. An example of a chat between a participant and chatbot is shown in Figure 3.2.

Dialogues where participants described certain medical conditions like social anxiety, depression and scoliosis were removed from the data (10 dialogues in total). I decided that those require professional consultation and should not be included in this study. I also narrowed down the set of concerns by disregarding concerns that appeared in the whole data less than five times. The dialogues where at least one of the arguments had a deleted concern were removed (18 dialogues in total). The concerns used for the following experiments were: *responsibility, family, productivity, dignity, wealth, comfort, relaxation, and fun.*

For AH2, 20 participants from each group were recruited and asked to chat with the chatbot. This time I included more prerequisites during the recruitment, namely no chronic diseases, no long-term health conditions/disabilities or ongoing mental illnesses. In this round, the chatbot did not ask the participants to assign concerns to their arguments. I harvested 40 arguments for each

Figure 3.1: Example chat between a participant and the AH1 chatbot. Chatbot arguments are in the dark boxes and user arguments in the light boxes. The chat begins in the left column and continues in the right one.



participant group in AH2 (no dialogues were deleted). After the two rounds of argument harvesting, the total number of argument-counterargument pairs was 284 and can be found in Appendix A [app, a].

After AH1, I made the following three observations. Firstly, some concerns were chosen more often than others, and a smaller set of concerns, therefore, sufficed to cover most of the arguments. Secondly, my simple chatbot was capable of harvesting a significant number of arguments. And lastly, I observed that many participants gave similar arguments or even the same argument, using different words. This opens the possibility of grouping arguments using clustering techniques. The experiments I conducted with the harvested arguments to test my hypotheses are described in the following subsections.

Table 3.1: Average agreement (AGT) for concerns (C) and parent-concerns (PC) for arguments harvested in AH2 and the corresponding κ scores (κ).

Group	S	K	NK
C AGT	68%	63%	67%
C κ	0.4	0.27	0.4
PC AGT	81%	86%	81%
PC κ	0.52	0.42	0.49

Table 3.2: Accuracy (AC) of the classifier-predicted concerns (C) and parent-concerns (PC) and the corresponding $F1$ scores.

Group	C AC	C F1	PC AC	PC F1
S	50%	0.47	78%	0.77
K	55%	0.55	83%	0.84
NK	43%	0.45	70%	0.69
Avg	49%	0.49	77%	0.77

3.5.2 Experiment I: Argument-Concern Labeling

The purpose of the experiment was to test whether different people assign the same (or similar) concerns to the same arguments that they have not posited themselves and whether it was possible to *predict* the concerns of arguments by training a classifier and therefore verify Hypothesis I.

The methods used in the experiment were the following. Twenty participants for each group were recruited using the same prerequisites as for the argument harvesting, apart from the Facebook account, as no chatting with the chatbot was required. I used Google Forms for this task. Since I was interested in how the same group of people judged the arguments, I asked members of the *student* group to assign concerns to the arguments given by the students (respectively for the kids and nokids groups). The participants were presented the 40 arguments from their group harvested in AH2 and given a choice of 8 concerns. They were asked to “read the argument for not engaging in physical activity and pick the concern that they associated with the given argument”. The concern that received the highest vote amongst the participant (concern agreement) was chosen as the corresponding concern for that particular argument. For example,

if for argument A1, 16 out of 20 participants chose the concern *family*, then *family* was assigned to A1, and the concern agreement is 80%.

I observed that certain concerns are interchangeable: for example, the concern *responsibility* was equivalent to *family* in the kids group and *productivity* in the student group. I therefore grouped six out of the eight concerns into the following two groups, calling these *parent-concerns*: **CRF**: *comfort, relaxation and fun*. **FRP**: *family, productivity and responsibility*. The remaining two concerns *wealth* and *dignity* had no parent-concern⁴. Parent-concern agreement for the individual arguments was calculated by adding the agreement rates for the individual concerns in that parent-concern group. The agreement ratios for the individual groups are shown in Table 3.1. I also calculated *Fleiss Kappa* scores to assess the reliability of agreement between the participants of each group.

I used the concerns assigned by the participants that received the highest concern agreement (participant concerns) to score the concern-classifier. The arguments and concerns from AH1 were used for training, while the arguments from AH2 and the concerns assigned by the participants were used for testing. Using the bag-of-words model, I trained a Support Vector Machine with a linear kernel. I scored the classifier by comparing the classifier-predicted concerns to the concerns assigned by the participants. The results are shown in Table 3.2. Accuracy is defined as the number of arguments where the concern predicted by the classifier was the same as the concern assigned by the participant. There was a choice of 8 concerns and three parent-concerns. Random classification would therefore be 12.5% and 33.33%, respectively. My classifier had an average accuracy of 49.9% for the concerns and 76.7% for the parent-concerns. Table 3.2 also includes the weighted *F1* scores for each participant group.

⁴They were grouped as a parent-concern during the classification to create a bigger group for the classifier as the two concerns on their own had too few examples.

Table 3.3: Distribution of arguments (Args) with parent-concerns *FRP*, *CRF*, and concerns *Dignity* and *Wealth*. S, K, and NK stand for the student, kids and nokids groups respectively.

Group	No. of Args	FRP	CRF	Dignity	Wealth
S	80	31.25%	60%	1.25%	7.5%
K	92	72.83%	25%	0%	1.09%
NK	112	25.89%	67.86%	1.79%	3.57%

The accuracy of prediction for the nokids group was lower than the other two groups due to the more diverse arguments compared to the other two groups. Table 3.3 shows how many arguments in each group were assigned with a specific parent-concern. In the kids group, 72.83% of the arguments were assigned the concerns *family* or *responsibility*. These arguments often contained the words *children*, *baby* and *kids*. For the nokids group the majority of the arguments (67.86%) addressed the concerns *comfort*, *relaxation* and *fun*. Those arguments were much more diverse and did not have as many keywords in common, making classification more difficult.

It can be concluded that even though people might disagree on nuances like whether a certain argument promotes the concern *family* or *responsibility* in the kids dataset or cannot decide whether an argument given by a person is better associated with *relaxation* or *comfort*, the majority of people agree on the parent-concern for a given argument. The results, therefore, support my Hypothesis I, that most people independently assign the same or similar concerns to an argument they have not posited themselves in the given domain.

3.5.3 Experiment II: Assessment of Harvested Arguments as Meaningful Arguments

In this experiment, I wanted to assess whether a chatbot can be used as a tool for harvesting meaningful arguments and therefore verify Hypothesis II.

The methods used in the experiment were the following. I recruited 10 par-

Table 3.4: Meaningful arguments (Args) in each group when the threshold is set to 70% annotator agreement and above

Group	No. of Args	No. of meaningful Args
S	40	28 (70%)
K	40	33 (82.5%)
NK	40	33 (82.5%)

Participants for each group (like in the previous experiment, participants were representatives of the groups, e.g. students judging the arguments given by students). The prerequisites were the same as in Experiment I. Participants were presented with all 40 arguments harvested from the corresponding group in AH2 in a Google Form. I told the participants that the arguments were crowdsourced reasons for not exercising and asked them whether they “considered the given arguments as reasons they could give appropriate advice”. I also asked them not to judge the quality of the reason but rather just the completeness of it. After each argument, they selected *yes* or *no*.

The results of the experiment are summarised in Table 3.4. I explain how I derived the results as follows: I set the threshold for considering a statement as an argument at 70% annotator agreement (approval rate). This means that if a minimum of 7 out of the 10 participants answered the question of whether a given statement is an argument positively, I labelled it as *meaningful*.

From the results, it can be concluded that a chatbot can indeed harvest meaningful arguments using no or very little domain knowledge, which supports my Hypothesis II. In total, over 78% of the arguments that were harvested in AH2 were considered meaningful.

3.5.4 Experiment III: Argument-Counterargument Matching



The purpose of the experiment was to test Hypothesis III i.e. to evaluate whether the counterarguments of semantically similar arguments are inter-

changeable, making it possible to use the counterargument of one argument to counter another similar argument.

The methods used in the experiment were the following. In order to cluster similar arguments, I needed a clustering algorithm. My dataset was too small to apply general-purpose unsupervised clustering algorithms, so I developed a specialised clustering algorithm that could take advantage of domain-specific knowledge. I describe the algorithm below and the pseudo-code can be found in Algorithm 2.

Algorithm 2 Clustering similar arguments

```

1: procedure CREATECLUSTERS(arguments, group)
2:   synonym_list = create_synonym_list(arguments)
3:   concern_list                                     ▷ list of parent-concerns
4:   arguments                                       ▷ stored with corresponding id and concern
5:   normalise_arguments(arguments) ▷ delete stopwords, punctuation, set caption
   to lower
6:   arg_nouns = []
7:   for arg in arguments do                       ▷ iterate through arguments
8:     arg = extract_nounphrases(arg)
9:     arg_nouns.append(arg.tokenize()) ▷ arguments now reduced to wordlists
10:  for arg in arg_nouns do                         ▷ iterate through argument wordlists
11:    replace_synonyms(arg) ▷ words looked up in synonym dictionary and
   replaced
12:    stem_args(arg_nouns) ▷ words in argument wordlists stemmed
13:    args_same_concern ▷ done for each concern in concern_list
14:    arg_dic = [] ▷ dictionary to store a list of similar arguments for each argument
15:    for i in range (args_same_concern) do
16:      compare all arguments with each other using compare function below
17:      if TRUE then arg_dic[arg1].append(arg2) ▷ if arguments similar, arg2 is
   added to the list of similar argument for arg1
18:    clusters = arg_dic.concerns()
19:    delete_sublists(clusters)
20:     clusters
21:
22: procedure COMPARE(arg1, arg2)
23:  if intersection of common words  $\geq 0.5$  then  TRUE

```

First, I created a synonym list using WordNet [Miller, 1995]. This list contained lists of all the words in a given corpus that were synonyms of each other. Then stopwords and punctuation were deleted and the casing was set to low. I

also deleted exercise and time-related words (*exercise/s, sport/s, day/s, week/s, hour/s, thing/s, reason/s, main, lot*) because many people repeated the chatbot's question in their answer (e.g. "*The main reason I don't exercise is [...]*"). So I did not want to consider those in my similarity measurements. I also disregarded words that were used to describe how often they did or did not engage in a particular activity. Finally, for each argument, the noun phrases were extracted and stored as separate words and the synonyms were replaced with the first word in the corresponding synonym list. The arguments are stemmed in order to avoid treating different forms of a word as different words. After preprocessing the arguments, all arguments with the same concern were clustered by comparing them to each other and clustering those together that shared more than 50% of the words. This resulted in clusters where each argument shared over 50% of words with every other argument. An argument could occur in more than one cluster.

I applied the algorithm separately to the arguments of each participant group. Every argument had an original counterargument as given by the same participant during the chat with the chatbot. Each argument that appeared in a cluster (was 'clustered') was matched with all the counterarguments from the other arguments in that cluster, apart from its original one. For example, if the arguments A1, A2, and A3 formed a cluster, then A1 would be matched with counterarguments of the other two arguments CA2 and CA3.

I evaluated the suitability of the counterarguments as follows: 10 participants for each group were recruited, with the same prerequisites as in Experiments I and II. I again used a Google Form where each argument was presented with its matched counterarguments, and the participants were asked to choose which ones they believed were a suitable counterargument for the argument given. They were told that the arguments, as well as the counterarguments, were collected via crowdsourcing and that they should not judge the quality of

Table 3.5: Total number of arguments (Args) in each group, number (percentage) of arguments clustered, the average number of counterarguments (CAs) per clustered argument and the number of argument clusters generated in each group.

Group	Args	Clustered total (%)	Clustered AH1 (%)	Clustered AH2 (%)	Avg CAs	Clusters
S	80	40 (50%)	18 (45%)	22 (55%)	3.65	19
K	92	49 (53.26%)	23 (44%)	26 (65%)	7.39	22
NK	112	42 (37.5%)	24 (33%)	16 (40%)	6.62	14

the arguments and counterarguments but rather whether the counterargument is an appropriate response to the given argument.

The results of the experiment are summarised in Tables 3.5-3.7. Table 3.5 shows how many arguments were clustered in the individual groups and the two rounds of harvesting. We can see that in the nokids group fewer arguments were clustered than in the other two groups. This is due to the higher diversity in arguments and more complex synonyms.

The counterarguments of each argument received a certain approval rate, showing how often a participant selected a given counterargument. Table 3.6 (column 3) shows the average approval rates of the counterarguments for each argument in that group. For example, if an argument had three counterarguments and their approval rates were 20%, 70% and 90%, the average approval rate of the counterarguments for that argument would be 60%. For more examples, see Appendix D [app, a].

I considered the average number of suitable counterarguments per argument by using an approval rate threshold of 50%. If, for instance, an argument had three counterarguments with the approval rates of 40%, 50% and 60% respectively, the second and third would be considered suitable and the number of suitable counterarguments would be 66.7% (2/3). The results are shown in Table 3.6 (column 4). The reason for the lower threshold is the high variance of quality amongst counterarguments. Some counterarguments scored poorly because

Table 3.6: Average approval rate (AR) of counterarguments (CAs) per argument and the average number of suitable CAs per argument with approval threshold of 50%.

Group	No. of Args	Avg. CA AR	Avg. No. suitable CAs
S	40	70.37%	80.66%
K	49	69.04%	84.41%
NK	42	60.10%	78.89%

they give inappropriate advice (see Example 3.3).

I also analysed the approval rate that the individual counterarguments received, averaging all the approval rates that a counterargument received for all the arguments it was matched with. This way, I wanted to identify inadequate counterarguments and wrongly clustered arguments. For example, if counterargument CA4 was matched with three arguments A1, A2, and A3 and received an approval rate of 40% for A1, 50% for A2 and 80% for A3, the average approval rate for CA4 would be 56.7%. The results are shown in Table 3.7. The following is an example of an inappropriate counterargument:

Example 3.3 The argument A4 and counterargument CA4 were given by the same participant.

A4: *“I only sometimes do sports because I am too busy and tired from my uni work”.*

CA4: *“You could join a sports team with a friend or find a gym buddy”.*

A4 was clustered with similar arguments (a total of 6), and therefore, CA4 was

⁵There are only 46 counterarguments for the 49 clustered arguments because in three cases, the participants answered *“I don’t know”* instead of giving a counterargument.

Table 3.7: The average approval rate (AR) of individual counterarguments (CAs) when matched with the corresponding arguments in their cluster.

Group	No. of CAs	Avg. AR
S	40	69.18%
K	46 ⁵	72.01%
NK	42	58.82%

matched with all the arguments of that cluster. It was, however, never approved as a suitable counterargument and had the lowest average approval rate in the *student* dataset (17.5%). It is not surprising that this counterargument was not considered a good one. It does not advise on how to manage your time better or emphasise the benefits of physical exercise and hence does not fit my definition of suitability because $C' \neq C$. In the chat, when the chatbot asked why the person was not following her own advice, the participant indeed answered: “*like I said, I am often too busy to do so. I mostly study or try to catch up on sleep*”. A counterargument that can be countered with “*like I said...*” is unlikely to be an appropriate counterargument.

From the results in this section, it can be seen that counterarguments of similar arguments are interchangeable as long as they give appropriate advice, which supported my Hypothesis III about the interchangeability of counterarguments of semantically similar arguments. With the current data, participants perceive a counterargument from a similar argument as suitable about 80% of the time when I set the threshold for suitability at 50% approval rate. Regarding the clustering algorithm, only 131 out of the 284 arguments were clustered. This was due to several factors, including the wrong classification by the concern-classifier, more complex synonyms and lost negations during the preprocessing of the arguments, specific explanations for a common reason, implicit meanings and specific arguments that did not repeat within the data. In the next section, I discuss the results of my experiments.

3.6 Discussion

My contribution in this study was threefold. Firstly, I have shown that a simple chatbot with little or no domain knowledge can acquire meaningful arguments. I have focused on the behaviour change domain, ordinary people give simple arguments that are nevertheless full of meaning and importance. They are

the kind of arguments that have been neglected in the formal and informal argumentation literature. There is little literature on how to analyse this sort of argument and even less on how to acquire them.

Secondly, I have shown that most people assign the same or a similar concern to given arguments, which makes it possible to predict arguments' concerns with a classifier's help. Given this observation, it can be concluded that given an argument, most people will be motivated by the same concern if positing it. I also made a first attempt to find a suitable set of concerns for a specific topic by letting the participants assign the concerns to their arguments themselves.

Thirdly, I presented a method to cluster arguments by concerns and similarity to create several possibilities to counter a given argument and evaluated whether those counterarguments are interchangeable. The results show that this is the case, given the counterargument itself was appropriate.

However, despite the promising results, I decided not to use argument harvesting in my subsequent studies. Developing a chatbot is more time-consuming than setting up a well-designed online questionnaire, and chats with participants require more time than filling out a form which results in higher study costs. I, therefore, opted for Google and Microsoft forms for argument acquisition. In the following chapters, I will present chatbots that engaged in persuasive dialogues with users.

Chapter 4

Impact of Argument Type and Concerns in Argumentation with a Chatbot

In this chapter, I present my first persuasive chatbot, which I used to conduct an empirical study to investigate how *concerns* of the persuadee and the *type* of counterargument impact the effectiveness of arguments in persuasion dialogues. The rest of the chapter is structured as follows: In Section 4.1, I introduce my research study; In Section 4.2, I discuss relevant literature; Section 4.3 gives some background theory on the notions of *appeal*, different types of counterarguments and presents the ones I investigated in my empirical study; Section 4.4 presents the aim of the study and the hypotheses; Section 4.5 describes the argument counterargument acquisition process; Section 4.6 describes the experiments that were conducted with the acquired data, namely the evaluation of the counterargument types and the chatbot that was used for the persuasion dialogue; and in Section 4.7 I discuss and conclude my findings.

4.1 Introduction

In persuasion, the way an argument is communicated is just as important as its message. A persuader who wants to convince a persuadee to do more exercise can present his argument in many different ways. He can, for example, point out the advantages of regular exercise: “*Regular exercise will strengthen your bones, muscles, and joints*”. However, he could also phrase it in a negative way: “*Lack of regular exercise leads to weakening of your bones, muscles, and joints*”. This notion of *framing* is well studied in psychology, and health care [O’keefe and Jensen, 2006, Plous, 1993, Tversky and Kahneman, 1985, Ogden, 2012]. Other persuasion techniques such as referral to authority and social proof [Cialdini and Cialdini, 2007] have also been used in psychology. I refer to the style of persuasion used in an argument, as *argument type*.

I wanted to investigate some common argument types used in persuasive dialogues in the behaviour change domain. Despite the extensive psychology literature on the topic of message framing and persuasion techniques, the notion of argument type is underdeveloped in the computational argumentation field. Walton’s *argumentation schemes* [Walton et al., 2008] could be viewed as a non-exhaustive summary of argument types. However, that leaves some important types for behaviour change unconsidered. Also, to the best of my knowledge, no empirical studies with participants were undertaken to test the effectiveness of argument types in persuasion.

Furthermore, I wanted to investigate how *concerns* of the persuadee impact the effectiveness of arguments in persuasion dialogues. The results in [Hadoux and Hunter, 2019] showed that taking persuadee’s concerns into account improved the persuasiveness of a dialogue. The persuader might present a valid argument the persuadee does not disagree with (e.g. that regular exercise is good for her health), but which has no impact on the persuadee because she is not concerned

with her health at that moment. However, she might be very concerned with her academic performance. The argument “*Regular exercise not only helps your physical health but it will help you study better*” might have a bigger impact on the persuadee because it addresses her concern.

In this chapter, I investigate different *types* of counterarguments and their preference with the persuadee based on the persuadee’s *concerns*. I propose a method for crowdsourcing arguments and counterarguments and assess a typology of counterarguments and concern assignments to be used by a chatbot. I used meat consumption as a case study. To verify my approach, I developed a strategic chatbot that takes the concern of the user into account and, during an argumentation dialogue with the user, presents only those types of counterarguments that address their concern. For comparison purposes, I also developed a baseline chatbot that does not address the user’s concerns. My results show that the strategic chatbot outperforms the baseline one and has a more significant impact on the user’s intention to reduce their meat consumption in the future.

In contrast to my previous study, in this study, the chatbot can engage in a fully-fledged argumentation dialogue with the user and thus act as a persuader, and instead of argument harvesting, I used Google Forms for crowdsourcing the arguments used by the chatbot.

4.2 Argument Types and Appeal

Apart from concerns, I also used the notions of argument types and argument appeal. I chose the topic *meat consumption* and was interested in different argument types in favour of reducing meat consumption which I could present to meat-eaters as counterarguments to their arguments in favour of eating meat. I acquired the arguments by crowdsourcing. Crowdsourcing offers a

fast and efficient way to gather a large number of diverse arguments without introducing the researcher's bias into the selection of arguments if gathering them by hand. I opted for this method because there exists no central repository of all possible counterarguments on the topic.

Argument Types: A persuader who wants to convince a persuadee to do more exercise can present his argument in many different ways. I refer to the style of persuasion used in an argument as *argument type*. I investigate six argument types in total. The most common type of counterargument used in the computational argumentation literature is the *negation* of an argument. Hence, I included a kind I call *direct counterargument* which negates the given argument.

In my previous study, described in the previous chapter, when people were asked to provide a counterargument to their given argument, people mostly gave arguments in the form of suggestions. Suggestions are often enthymemes and do not explicitly negate the argument. They, however, imply that changing the behaviour is advantageous (therefore attacking the argument) and provide a solution for achieving that. For this reason, I include suggestions or *suggestion-based arguments* in my argument types assessment.

As mentioned above, arguments can be framed positively or negatively, either referring to a gain or a loss (i.e. positive or negative consequence for the persuadee). I, therefore, included *positive* and *negative consequences* into my list of argument types. Further, a certain behaviour often has consequences not just for the person engaging in that behaviour but for others as well, which I call *personal* and *impersonal* consequences, respectively. Smoking, for example, is not just bad for the smoker but also imposes a burden on the health care system if the smoker becomes sick due to his behaviour. Therefore, I end up with six argument types I want to investigate. I give the definition and an example of

Table 4.1: Definitions of six investigated argument types with examples. The argument countered in the examples is “*I eat meat because it tastes good*”.

Argument Type	Definition	Example
Direct Counter-argument	This argument is counterargument that directly negates a previously given argument by referring to it.	<i>A raw, unprepared chunk of meat doesn't taste good. It's about the way of preparation and seasoning.</i>
Suggestion-based Argument	This argument gives a suggestion that may implicitly refer to a previously given argument and suggests how to change the behaviour in question	<i>You could introduce one day a week where you don't eat meat. Overtime you can increase the number of days.</i>
Positive Personal Consequence	This argument gives a positive consequence for the persuadee personally, if he or she continues the behaviour in question.	<i>Eating less meat will decrease your cholesterol level which will ultimately lower your risk of stroke and heart disease.</i>
Positive Impersonal Consequence	This argument gives a positive consequence for someone/ something apart from the persuadee, if he or she continues the behaviour in question.	<i>Eating less meat will lead to the reduction of the water footprint on the earth.</i>
Negative Personal Consequence	This argument gives a negative consequence for the persuadee personally, if he or she continues the behaviour in question.	<i>Most processed meats are loaded with artificial chemicals, including flavourants, colourants and preservatives that might be bad for your body.</i>
Negative Impersonal Consequence	This argument gives a negative consequence for someone/ something apart from the persuadee, if he or she continues the behaviour in question.	<i>Much land is needed to raise cattle for which forests have to be cut down, therefore causing deforestation.</i>

each type in Table 4.1.

Concerns: In a dialogue, the chatbot needs to choose which arguments to present. Various criteria can be used for selecting arguments. Since I aim to maximise the persuasive effect, the chatbot should preferably select the most persuasive argument. I consider how the chatbot can select arguments based on concerns that the user has. Given a set of possible counterarguments to the posited user argument, the chatbot should select an argument that addresses the same concern as the user argument.

From a preliminary investigation which involved researching the most common arguments against meat consumption on the internet, I discovered that most arguments revolve around two major *concerns*: *Health* and *Environment*. Note, I can view the health concern as a *personal* concern and the environment concern as an *impersonal* concern.

Appeal: In this study, I was interested in the appeal of the argument *type*, not the argument itself. As pointed out by [Hunter and Polberg, 2017] an appealing argument might not necessarily be convincing. The argument that education should be free might be very appealing. Still, at the same time, I can acknowledge that universities need resources to function and, therefore, not be very convincing. I am, however, interested in the appeal of the *type* of the argument and not in the appeal or convincingness of the *message* of the argument. I believe that one type of argument cannot be more convincing than another type per se, but one type can indeed be more appealing to the persuadee than another.

4.3 Hypotheses

In this chapter, I show how the persuader's choice of argument type and concern influences the persuadee's intention to change their behaviour. Firstly,

I explored different *types* of counterarguments and evaluated their appeal to the participants in the behaviour change domain. Secondly, I investigated whether the persuadee's concerns impact the argument types the persuadee found most appealing. Thirdly, I used a chatbot to test whether presenting only those counterarguments that address the persuadee's concern is more likely to change his or her intention positively than presenting counterarguments that address other concerns in the domain. I summarise these points in the following three hypotheses:

- H1** When a person is presented with counterarguments of various types, some types are perceived as more appealing than others.
- H2** When a person is presented with counterarguments that address different concerns in that domain, people find those counterarguments more appealing that address the concern that they perceive as more important.
- H3** A chatbot with no natural language understanding, just by presenting the type of counterarguments that take the user's concern into consideration, is more likely to have a positive impact on changing the user's attitude than a chatbot that presents the type of counterarguments that ignore the user's concern.

In the remainder of this chapter, I describe the methods for acquiring the main arguments why people eat meat, different types of counterarguments for meat consumption, and explain the experiments conducted with them to test my hypotheses and present the results.

4.4 Argument & Counterargument Acquisition

My study consisted of two parts: the argument and counterargument acquisition, described in this section, and the experiments (described in the next

section), which I conducted with the acquired data to test my hypotheses. The participants for all surveys and experiments were recruited via *Prolific* (www.prolific.ac). For each survey, I recruited from either one of two disjoint groups: meat-eaters and vegetarians. I opted for this division to obtain counterarguments from people who do not engage in the behaviour in question (in this case, meat consumption), which can then be used by the chatbot that tries to persuade people who do eat meat to change their behaviour. The general prerequisites for taking part in my study were being over 18 and fluent in English. I used Google Forms for all surveys.

Argument Acquisition and Clustering To find the most popular arguments for eating meat amongst the participants, I recruited 40 meat eaters and asked them in a Google Form to give their main reasons for eating meat. This way I acquired 111 arguments which can be found in Appendix I [app, b]. The average length of an argument was seven words with a standard deviation of 5.

I used the algorithm described in Chapter 3 (Algorithm 2) to preprocess and cluster the arguments automatically. Arguments were clustered by similarity to identify the most popular arguments for eating meat. As a representative argument for each cluster, I randomly picked one of those that contained the highest number of *most common words* in that cluster. In the rest of the chapter, those will be referred to as “the most popular arguments” for eating meat. The

Table 4.2: Summary of the arguments acquired in Step 1. Cluster name, number of arguments (A) in that cluster and representative argument for that cluster.

Cluster	No of A.	Representative Argument
Nutrition	15	<i>For its nutritional value and source of protein</i>
Filling	6	<i>It's filling</i>
Taste/Like	40	<i>It tastes good</i>
Easy	6	<i>Quick and easy to prepare</i>
Health	11	<i>It's healthy and contributes to a balanced diet</i>
Variety	3	<i>It offers more variety to my meals</i>

size of the clusters and the representative argument of each are given in Table 4.2. The name of each cluster is the most common word found in that cluster (excluding stopwords).

Counterargument Acquisition After identifying the most popular arguments for eating meat, I started with the direct counterargument acquisition. For each of the six most popular arguments, I created one survey. I recruited ten vegetarians per survey and asked them to counter the given argument by giving a single argument. This way, I acquired ten direct counterarguments for each of the most popular arguments.

I was interested only in the “best” counterarguments and, therefore, created six surveys (one for each of the six most popular arguments) and recruited 20 participants for each survey, who identified themselves as meat-eaters. The participants were presented the argument and the 10 acquired counterarguments for that particular argument. Since I was not interested in the message of the counterargument (e.g. its believability or convincingness) but still wanted clear, understandable and appropriate representatives of each counterargument type, I asked the participants to select those counterarguments that they found best at communicating their message. I counted the number of times each counterargument was voted for and ordered them by the number of votes. All counterarguments can be found in Appendix II a [app, b].

For the remaining counterargument types, I created one survey for which I recruited ten vegetarians and asked them to provide one counterargument of each type. They were given the same examples of these argument types as given in Table 4.2. Using the same approach as for the direct counterargument acquisition, I again received a ranking that allowed us to identify the best counterarguments (those ranked the highest).

4.5 Experiments

The experiments were split into two parts: the first was concerned with the evaluation of the different argument types according to their appeal and their concern assignment. I wanted to test whether there is a correlation between the concerns of the participants and their preferred argument type. In the second part, I used a chatbot to test whether presenting counterarguments that take the user's concerns into account was more likely to change the user's attitude positively compared to a chatbot presenting counterarguments that ignored the user's concerns.

I used the three counterarguments that were ranked the highest by meat-eaters. This resulted in 18 direct counterarguments (three for each of the six most popular arguments for eating meat) and the top three ranked of the remaining five types. These counterarguments can be found in Appendix II b [app, b].

4.5.1 Evaluation of Argument Types

In this part of the experiment, I evaluated the six different argument types according to their appeal to the participants and showed the correlation between the preferred argument type and the concerns of the participants in order to investigate hypotheses H1 and H2.

Methods

I created a survey where the participants were asked what their main reason for eating meat was. There was a choice of the six most popular arguments (Table 4.2) and the option "other". Then they were presented with the three highest ranked counterarguments of each type. They were asked to pick all the counterarguments that appealed to them. Note that if they selected the option "other" in the previous step, no direct counterarguments were presented. In the

Table 4.3: Results from the evaluation of argument types. *DIR* = *Direct Argument*, *SUG* = *Suggestion*, *NIC* = *Negative Impersonal Consequence*, *NPC* = *Negative Personal Consequence*, *PIC* = *Positive Impersonal Consequence*, *PPC* = *Positive Personal Consequence*.

Arg Type	NPC	PPC	NIC	PPIC	SUG	DIR
Votes	55%	51%	50%	47%	33%	6%

end, they were asked to provide a short explanation of why they chose those counterarguments. I recruited 100 meat-eaters.

Results

I was interested in two things: Firstly, whether there is a difference in the popularity of argument types. And secondly, whether there is a correlation between the preferred argument type of the participants and any of the information that they provided in their explanation which the chatbot could take into consideration when presenting the arguments during an argumentation dialogue. Table 4.3 shows how many of the shown argument types were selected overall by all participants, i.e. there were three arguments of each type since 33% of the *Suggestion-based arguments* (SUG) were selected, which means that on average, each participant selected one SUG. One can see that *Direct Counterarguments* are much less popular compared to the others¹. The four types of consequential arguments were the most popular. The results supported my H1 that different types of counterarguments differ in their appeal.

I observed that the explanations of most participants raised concerns about their health or the environment, or both, which further supported my choice of concerns for this domain. To investigate this, I automatically assigned concerns to the explanations. Every explanation that contained the word *health* was assigned the concern *Health* and those that contained the words *animal*, *environment*, *planet* were assigned the concern *Environment*. Explanations that contained words from both concerns were labelled *Both*. I observed a high

¹only 6 participants chose the option “other” and were therefore not presented any DIR

statistical correlation between the participants' concerns and their preferred argument type. Participants who gave an explanation that was labeled with the concern *Health* preferred the *personal* consequences, whereas those with *Environment* preferred the *impersonal* consequences. Participants who gave an explanation that was labelled with both concerns preferred all consequential counterarguments equally. I used the Chi-Square test to calculate statistical significance by comparing the numbers of the available counterarguments with the number of the selected ones². The p-values for all three groups were below 0.001. The results are summarised in Table 4.4. The results supported my hypothesis H2 that people strongly prefer argument types that relate to their concerns.

4.5.2 Evaluation of Chatbot

I developed two versions of the chatbot, one that took the user's concern into account when presenting counterarguments (strategic chatbot) and one that did not (baseline chatbot). The purpose of the chatbot was twofold: firstly, to test whether presenting counterarguments that address the user's concern is more likely to change the user's attitude positively than presenting counterarguments that ignore the user's concern. Secondly, to test whether a chatbot that has no natural language understanding can engage in an argumentation dialogue and influence the user's attitude about the discussed topic. By natural language understanding, I mean that the chatbot does not "understand" what the user writes, i.e. no keyword matching or machine learning. So this experiment was to investigate hypothesis H3.

²For example, since 28 participants were concerned about health only, there were 168 (28 x 6) consequential personal arguments to select from and 336 (28 x 12) of the remaining types. 120/168 out of the consequential personal arguments were selected in contrast to only 70/336 of the remaining types.

Table 4.4: Total number of selected counterarguments per concern and number of participants (p). See caption for Table 4.3 for acronyms.

Concern	Argument Type						No of p.
	DIRECT	SUG	NIC	NPC	PIC	PPC	
Health	7	26	22	65	15	55	28
Env.	2	25	59	31	59	22	31
Both	3	17	39	40	39	42	18

Methods

The chatbot was deployed on Facebook via the Messenger Send/Receive API. For more on the implementation of such a chatbot, see [Chalaguine and Hunter, 2018b]. For each chatbot, I recruited 50 participants who to the question whether they *definitely wouldn't*, *probably wouldn't*, *might*, *probably would* and *definitely would* reduce their meat consumption did not reply with *definitely would*. The dialogue protocol is described in dialogue steps DS1 to DS8:

DS1 The participant was asked at the beginning of the chat to select whether they would consider reducing their meat consumption. The choices were: *definitely wouldn't*, *probably wouldn't*, *might*, *probably would* and *definitely would*.

DS2 Then, they were asked what they were more concerned about: the impact that meat consumption had on their health or the impact it had on the environment and animals. They were given two options to select: *health* and *environment/animals*.

DS3 Then, they were asked to select their main argument for eating meat (see Table 4.2) and the option “other”.

DS4 Then, the chatbot presented its first counterargument. The strategic chatbot would present either six positive and six negative personal arguments (if the participant selected *health*), or six positive and six negative impersonal arguments (if the participant selected *environment*) during the course of the chat. The baseline chatbot did not take the concern into account and presented three counterarguments of each type. For the chatbot, I only used the four consequential types of counterarguments since they scored the highest during the first part of the study, described in section 4.

DS5 After each counterargument that the chatbot presented, the participant had the choice to select *agree* or *disagree*.

DS6 If the participant agreed, the response depended on the chatbot’s variant. I implemented two slightly different variations of each chatbot.

Variant I Chatbot presented the next counterargument.

Variant II Chatbot asked “*Why do you eat meat then?*”.

DS7 If the participant disagreed, the chatbot asked “*Why?*”. The participant gave an argument, and depending on the length, the chatbot either asked the participant to expand or accepted it and presented the next counter-argument, to which the participant agreed or disagreed and so on. The query algorithm is explained in my previous work [Chalaguine et al., 2018] and can be found in Appendix VI [app, b].

DS8 At the end of the chat, the chatbot asked the participant again to select whether they *definitely wouldn't/probably wouldn't/might/probably would/definitely would* consider reducing their meat consumption.

Examples of chats with all four chatbots can be found in Appendix IV [app, b] and one with the strategic chatbot of Variant I can be found below in Figure 4.1.

I divided the 50 participants for both variations of the chatbot into two groups depending on which concern they selected. For each concern group, I calculated the change in intention. The change in intention is the final choice of intention minus the original choice of intention. I call the units of this measure *intention points* (IP). For example, if one participant changed her intention from “probably wouldn’t” to “might” after chatting with the bot, this counts as 1 IP, whereas changing from “might” to “probably wouldn’t” counts as -1 IP. Table 4.5 shows the number of participants in each concern group and their average intention change within the group and in total. One can see that the total number of IPs for strategic Variant I is over three times higher than for the baseline of Variant I, and for strategic Variant II the number of IPs is over four times higher than for the baseline Variant II.

Interestingly, the total average number of arguments participants disagreed with while chatting with the baseline chatbot remained the same as the strategic chatbot. On average, participants disagreed with 3.5 arguments out of 12 and 4 out of 12, respectively. From this, it follows that participants do not necessarily

Figure 4.1: Example chat between a participant and the AH1 chatbot. Chatbot arguments are in the dark boxes and user arguments in the light boxes. The chat begins in the left column and continues in the right one.

Welcome. A few things before we start: Please type your answers into one message (don't send several messages). Please do NOT send 'likes' or gifs and always select options when given. The chat will naturally come to an end. At the end of the chat you will get the prolific completion code. If the chatbot for some reason stops replying, please just send your prolific ID and end the chat. This chat does not work on the Messenger Lite app. All good?

Yes

Great! I will present you with reasons why you should consider reducing your meat consumption. You can either agree or disagree. If you disagree, I am interested why you don't agree with them. Are you ready?

yes

Awesome. Please tell me what applies most to you: 1: I definitely wouldn't 2: I probably wouldn't 3: I might 4: I probably would 5: I definitely would consider reducing my meat consumption

3

Please tell me what you are more concerned about: the impact that meat consumption has on your health, or the impact it has on the environment and animals?

my health

Okay. What is the main reason you eat meat? Select one of the following: 1: I eat meat because of its nutritional value and source of protein 2: I eat meat because it's filling 3: I eat meat because it tastes good! 4: I eat meat because it's quick and easy to prepare 5: I eat meat because it's healthy and contributes to a balanced diet. 6: I eat meat because it offers more variety to my meals 7: Other

3

But your reduction of meat will leave you feeling better physically and mentally. There is less risk of health complications, including food poisoning and an increased benefit of adding plant based foods to your diet to make it healthier and more varied

I disagree

Why?

Although there may be some benefits to reducing my meat consumption, I enjoy the taste and sensation of eating meat. And life is for living, for experiencing things you enjoy.

However, reducing the consumption of (processed) meat and substituting it with mock meat and better with vegetables, grains and mushrooms will increase the diversity in your diet, provide you with a higher variety of nutrients and reduce your intake of hormones and saturated fat.

I agree

eating processed meat several times a week (or even day) can lead to serious health issues due to the high consumption of saturated fat, hormones and medicine (antibiotics) in the meat

I agree

you will reduce your risk for cancer, heart disease and diabetes if you cut down your meat consumption

I disagree

Why?

Whilst I agree that those points are valid. I don't currently consume lots of meat and I hardly eat any processed meat at all. I may have some salami or bacon once every 2 weeks, but that's all.

All right. However, eating less meat results in better digestion and less constipation

I disagree

Why?

Well it depends what you mean by using the statement less. A varied diet is a healthy diet and provided you ensure you eat anything in moderation it can't be so bad

However, eating less meat improves your general health, enables your digestive system to process foods more easily and gives you more energy, leaving you feeling less sluggish and tired

I agree

it may make you antibiotic resistant since the antibiotics that we depend on to treat human illnesses are now used to promote growth in animals and to keep them alive in horrific living conditions that would otherwise kill them

I agree

great, what do you think about this reason: excessive consumption of meat can lead to many health problems, including heart disease, cancer and high blood pressure

I agree

red meats have been proven to have negative effects on the human digestive system, causing harm with potential illness.

I disagree

Why?

This would only apply when eating excessive amounts. Provided you eat meat as part of a balanced diet and don't consume too much, you should be fine health wise

But meat consumption was linked to cancer, heart disease and diabetes.

I disagree

Why?

I agree with the point, but again it depends on the amounts you eat. I have never eaten excessive amounts of meat. It is just another part of my balanced diet

Ok, I see. But eating too much meat is linked to health complications such as an increased likelihood of a stroke or heart attack.

I agree

if you eat less meat you will lose weight.

I disagree

Why?

Again it depends upon how much meat you were eating in the first place. It also depends upon how you are trying to lose weight. Meat is a great source of protein and if you are turning fat into muscle, a balance level of meat in your diet can be beneficial

I ran out of arguments :) let's end the chat here. One more question. Please tell me what applies most to you: 1: I definitely wouldn't 2: I probably wouldn't 3: I might 4: I probably would 5: I definitely would consider reducing my meat consumption

3

Table 4.5: Results for Variant I & II grouped by Baseline/Strategic and the concerns Health/Environment and their totals/averages.

Chatbot	Baseline			Strategic		
	Health	Env.	total/avg	Health	Env.	total/avg
Variant I						
No of p.	27	23	50	26	24	50
Sum of IP	6	4	10 (0.2)	20	12	32 (0.64)
Variant II						
No of p.	29	21	50	28	22	50
Sum of IP	-1	6	5 (0.1)	12	10	22 (0.44)

disagree with counterarguments that do not address their concerns. But despite that, those counterarguments do not impact their intention. It is not surprising that fewer people changed their intention positively when chatting with Variant II, due to its “annoying” nature. Many people were irritated by the repetitive question of “Why do you eat meat then?” after they agreed with an argument. For an example, see Appendix IV [app, b].

Table 4.6 shows how many participants changed their intention for the worse and the better, disregarding the number of intention points for the baseline and the strategic chatbot. I consider a positive change as a change from negative to neutral, negative to positive and neutral to positive. Changes from *definitely wouldn't* to *probably wouldn't*, and changes from *probably would* to *definitely would* were not considered. I used the number of participants who changed their intention for the better to calculate the statistical significance of the difference between the control group that chatted with the baseline chatbot and the group that chatted with the strategic chatbot using the Chi-Square

Table 4.6: Number of participants that changed their intention to the worse (W) and to the better (B) for all four chatbots.

Participant Group	Baseline				Strategic			
	Health		Env.		Health		Env.	
Change of intention	W	B	W	B	W	B	W	B
Variant I	1	5	2	7	0	17	0	11
Variant II	4	3	3	5	0	10	0	12

test. Due to the small numbers, I did not distinguish between variants in the calculation of the p-value. The p-value of 0.002 supported my hypothesis H3 that presenting arguments that address the user's concern is more likely to positively change the user's attitude than presenting arguments that ignore the user's concern. All the chat data can be found in Appendix V [app, b].

4.6 Discussion

My contribution to this chapter is fourfold. Firstly, I have shown that some types of arguments are considered more appealing than others in the behaviour change domain. Direct counterarguments and suggestions were the least popular in my study. Suggestions might not necessarily be unappealing but simply not tailored to the specific argument of the persuadee and therefore not relevant. On the other hand, direct counterarguments might trigger negative feelings from the persuadee, who, as a result, might get defensive. There was no significant difference in the remaining argument types in their appeal. In a user study on the persuasiveness of healthy eating messages [Josekutty Thomas et al., 2017], positively framed messages were shown to be more persuasive than negatively framed messages.

Secondly, I have shown that people prefer arguments that address the concern they perceive as more important. This is not surprising. However, *concerns* are often ignored when judging the effectiveness of arguments or choosing a strategy. Some studies use different personality traits of the user attributes to evaluate what sort of argument might be more effective for this particular person (for examples, see [Lukin et al., 2017, Ding and Pan, 2016, Villata et al., 2017, Santos et al., 2009, Mazzotta et al., 2007]). Computational argumentation largely focuses on sentimental [Choi and Cardie, 2008], rhetorical [Habernal and Gurevych, 2016b] and structural [Boltužić and Šnajder, 2014] attributes of the argument, rather than attributes about the user. I have shown that without

knowing anything about the user's personality, and by simply asking them what they are more concerned about, I can present arguments that positively impact their intention to change their behaviour. This leads to my third contribution. I have shown that presenting arguments that address the user's concern is more likely to positively change the user's attitude than presenting arguments that ignore the user's concern.

And lastly, I considered how a chatbot with no natural language understanding could engage in an argumentation dialogue and influence the user's attitude towards a certain topic. My approach of crowdsourcing some of the main arguments on why people engage in a certain behaviour and the corresponding counterarguments of various types that are then used by a chatbot to engage in persuasion dialogues is a novel approach in the behaviour change domain. Using crowdsourced arguments does not require professional research but solely relies on participants' input. Using only the highest-ranked counterarguments assures that no inappropriate arguments are chosen for the chatbot. There are, however, also potential risks to consider. For example, the spread of invalid arguments that, despite being popular, might contain wrong information. In subsequent chapters, I will present two more chatbots used in two different domains: a chatbot where the whole knowledge base consists of crowdsourced arguments and a chatbot where the chatbot's arguments were hand-crafted.

The insights gained from this work shaped the methodology of my future work. In the following I only used crowdsourcing with the help of online surveys for argument acquisition, instead of using a chatbot (see previous chapter). And during the knowledge base construction I excluded suggestions and direct counterarguments. In the next chapter, I present a method to crowdsource a large argument graph with several levels of depth instead of the argument-counterargument pairs presented in this chapter.

Chapter 5

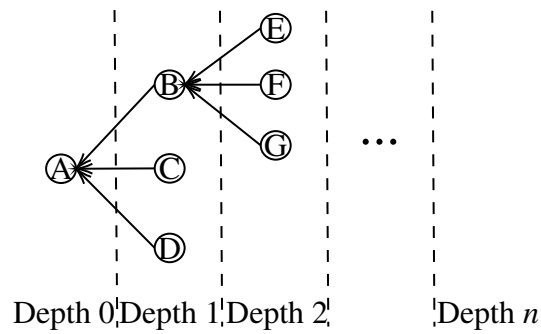
Knowledge Acquisition and Corpus for Argumentation-Based Chatbots

In this chapter, I present a method to acquire a large number of arguments in a graph structure using crowdsourcing. I evaluate this method in a study with participants and present a corpus that can be used for further research in computational argumentation and chatbot technologies for argumentation. The rest of the chapter is structured as follows: In Section 5.1, I describe existing approaches to argument acquisition and present my proposed solution; In Section 5.2, I explain my method for acquiring an argument graph; In Section 5.3, I present my case study and the resulting corpus; In Section 5.4 I evaluate the corpus, and in Section 5.5 I discuss and conclude my findings.

5.1 Introduction

In the previous chapter's work, the chatbot's arguments were crowdsourced. The chatbot, however, was unaware of the users' counterarguments and was

Figure 5.1: Representation of depths and attack relationships between arguments in my argument graph. Arguments *B*, *C* and *D* are counterarguments to *A*.



therefore unable to counter them, but only to present a new one which was not an attack on the user’s argument. Hence, the chatbot was only able to acquire argument-counterargument pairs. The resulting argument graph would have extensive breadth but not go beyond two levels: the chatbot’s arguments and the user’s counterarguments.

In this chapter, I propose a method to acquire a large number of arguments in a graph structure using crowdsourcing and present a corpus that can be used for further research in the computational argumentation domain. I aimed to generate a corpus of arguments in a graph-like structure which I could use as a chatbot knowledge base in my further research where the chatbot would engage in an argumentation dialogue with actual participants. Apart from a minimum and maximum length, participants had no constraints when submitting arguments in order to create a big graph of natural language arguments.

In the rest of the chapter, I describe my method to create an argument corpus on university fees in the UK and evaluate the quality of the obtained arguments in an experiment with crowdsourced participants.

5.2 Method

The depth of a graph is defined as the maximum number of arcs one can follow starting from the root. I created a graph of depth 5, the root argument being depth 0. Starting from the root and following any path, one will end up with a maximum of 5 arguments (excluding the root argument). The arguments in depth 1 attack the root argument and are therefore *against* keeping the university fees, the arguments in depth 2 attack the arguments in depth 1 and are therefore *for* keeping the fees and so on. Figure 5.2 shows a schematic representation of depths in my argument graph.

In the following, I first present my method of acquiring an argument graph and then describe the acquisition of my argument graph on UK university fees using my method.

5.2.1 Argument Processing

To address the problems above, I opted for using *crowdsourcing* as a means to obtain the arguments for the argument graph. For the first level (i.e. depth 1), participants were crowdsourced and presented with the root argument in a survey and asked to counter it with several arguments. The resulting collection of arguments in depth 1 was all counterarguments to the root argument.

In the following, I describe a pipeline that automatically extracts the best arguments from the crowdsourced ones in each depth to include them in the graph and collect their counterarguments in the next level.

1. Argument Length

I wanted a potential chatbot to give counterarguments that are neither too short nor too long. Therefore, I removed all arguments below 15 and above 50 words in length. I would not want a potential chatbot to give a short statement as a

counterargument to the user's argument. I did not include arguments longer than 50 words because they likely contained several arguments. I also did not consider them suitable for a chatbot knowledge base (imagine a chatbot replying with a whole paragraph).

2. Choice of topic words

I then extracted the most common words from the data (excluding stop words and words that do not add value in the given domain). The definition of *most common* depends on the size and nature of the data and is therefore up to the researcher to decide.

From the most common words, I then selected *topic words* which are words that I considered meaningful in the given context. These words can be typed or grouped into *concerns*. The choice of suitable topic words depends entirely on the domain, and their choice is also left to the researchers' discretion. For example, in a set of arguments on university fees, the word *money* appeared many times. They are, however, not very meaningful, whereas the words *debt* and *affordable* tell us more about the topic of the arguments. So by inspecting the frequently occurring words, the researcher can apply their knowledge of the domain to decide which would be good topic words. All arguments that contained at least one topic word were kept, and the rest were removed. One should note that the list of topic words increases with each depth. The threshold of how often a word has to appear to be considered "common" also rises since the number of arguments increases with each depth.

3. Spell-check

I kept all arguments that contained no spelling mistakes. This can be checked by using *Grammarly*¹. I deleted all arguments where Grammarly highlights a typo to avoid including arguments into the chatbot knowledge base that

¹<https://app.grammarly.com/>

contain spelling mistakes since this could influence the persuasive power of the argument. However, I did not consider incorrect punctuation or missing capitalisation as spelling mistakes, given the informality of the setting. Unfortunately, there was no Grammarly API at the time of writing, and I, therefore, had to copy-paste the arguments into the Grammarly app.

4. Final Selection of arguments for current depth

The arguments that were left after steps 1-3 were presented to crowdsourced participants who were instructed to select those arguments that they found communicated their message the best. I opted for this wording since I was not interested in the message of the arguments (e.g. its believability or convincingness) but still wanted to include clear, understandable and appropriate arguments in my graph. The highest-ranked arguments were then included in depth 1 of the argument graph.

Subsequent levels of depth

To minimise the need for crowdsourcing in Step 4 and in subsequent levels of depth, I only kept arguments that covered (i.e. contained) the highest number of topic words. I only presented arguments to crowdsourced participants for ranking, where the topic words were the same, and a selection had to be made. This reduced the need for participants in Step 4 after depth 1. The idea behind this method was to include arguments in the argument graph that addressed the maximum number of issues represented by the topic words.

5.2.2 Argument Acquisition for Next Depth

The arguments for all subsequent levels were collected by presenting the arguments from the previous level to crowdsourced participants who were asked to counter them. Steps 1-3 were then applied to the collected arguments for that level. The last two arguments in the graph were presented to the participants

since presenting only the last may be confusing without the attacked one as a reference. For example, during the acquisition of arguments in depth 4, participants were shown an argument from depth 2, one of its counterarguments in depth 3 and asked to assume the stance of the argument in depth 2 and counter the argument in depth 3.

5.3 Case Study and Corpus

In the UK, the situation at the time of writing was that home students (students from the EU, including the UK) pay around 9000£ tuition fees per year for a Bachelor's degree². This is a controversial situation, with supporters and contestants on both sides. I, therefore, chose this as a suitable topic for my task and selected "*Universities in the UK should continue charging students the 9k tuition fee per year*" as the root topic for my graph. In the following, I describe how I acquired my argument graph corpus on university fees in the UK by applying the method described above.

Participants were recruited via *Prolific*³. I used Google Forms for my study. The prerequisites for taking part in the study were to be over 18, fluent in the English language and a current resident of the UK (to minimise the risk of recruiting participants who do not know anything about the university fee situation in the UK).

For depth 1, I recruited 91 participants who were asked to provide three different reasons in a Google Form on why they think the 9k tuition fees in the UK were inappropriate and should be abolished. I, therefore, collected 273 (3 x 91) arguments at depth 1.

Many responses consisted of short statements like "*It is too expensive*" or

²The study was conducted pre-Brexit.

³<https://www.prolific.ac/>

“*students are poor people*” which I would not want a potential chatbot to give as counterarguments to the user’s argument (assuming the chatbots argues against fees). During the argument acquisition in future depths, I instructed the participants to provide arguments at least 15 words in length as I was only left with 97 arguments after this step in depth 1⁴.

I then extracted the most common, meaningful words from the data. Common words which did not add value in the given domain were words like *education, university, fee, abolish, students, degree* and *tuition*. I extracted all words that came up at least five times in the dataset of 97 arguments. From the most common words, I selected the words *job, debt, afford/affordable, access/accessible* and *free* as topic words for depth 1. Other common words included *study, high, amount, money, pay* and *work*, which I believed were too general. I mentioned above that the list of topic words grows with each depth: In depth 2, for example, the words *loan, tax, government* and *scholarship* were added to the list of topic words.

After steps 1-3, I was left with 48 arguments out of the 273 at depth 1. In depth 1 I decided to include three arguments for each topic word in the graph. I created five surveys (one for each topic word) which presented all arguments that included the topic word in question. I crowdsourced 20 participants per survey and instructed them that the arguments might be very similar and all touch on a particular aspect but that the individual arguments differ in quality. I asked them to select the best arguments they found to communicate their message. I then used the three arguments that were ranked the highest in each group. Some arguments contained two topic words. Therefore, some topic words are represented by more than three arguments.

I aimed to create a graph where each argument after depth 1 has three coun-

⁴When the study took place, Google Forms did not support response validation. Since July 2019, a minimum character count can be specified.

terarguments (on average) to avoid making the graph too big due to limited funding. In subsequent depths, I only kept arguments that covered the highest number of topic words. If several arguments had the same topic words and a selection had to be made, those arguments were presented to crowdsourced participants for ranking.

For example, consider an argument in depth 1 that had six counterarguments in depth 2 after applying Steps 1-3. The counterarguments (CA) contained the following topic words: (CA1) loan, debt; (CA2) loan, debt, scholarship; (CA3) loan, government; (CA4) government; (CA5) loan; (CA6) loan, government.

CA2 was selected for the next depth because it contained the highest number of topic words, and CA1, CA3 and CA6 were presented in a survey to participants to choose the “better” one for the graph.

Depth 1 consisted of 16 arguments. I created three surveys (containing five, five and six arguments, respectively) and recruited 10 participants per survey to counter the given arguments. I split the arguments into three smaller surveys to avoid presenting similar arguments and reduce the risk of participants giving the same counterargument to several arguments. For each subsequent level of depth, the arguments from the previous depth were divided into surveys of 5-6 arguments, and 10 participants were recruited per survey. I, therefore, acquired ten counterarguments per argument in each depth. The last two arguments in the graph were presented to participants. For example, during the acquisition of arguments in depth 4, participants were shown the argument from depth 2 (against fees), one of its counterarguments in depth 3 (pro fees) and asked to assume the position of being against fees and counter the argument in depth 3. One should note that for depth 5, I only recruited 5 participants to counter the arguments of depth 4.

5.3.1 The Corpus

My graph contained 1288 arguments, with each argument on average having three counterarguments and consisting of 5 depths, making it the most extensive corpus of this kind. The overall corpus of acquired arguments contained over 4000 arguments.

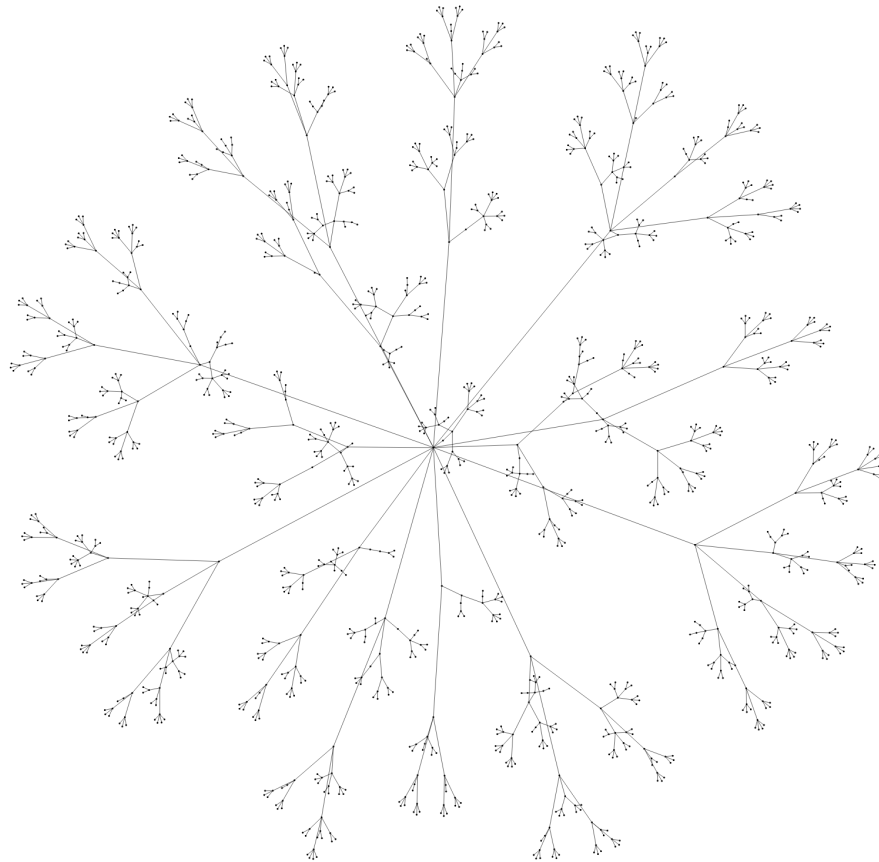
The generated corpus can be found on github [cor,]. It consists of two data sets. One data set contains the raw arguments acquired for each depth. The second dataset contains the arguments that were used in the generation of the argument graph. Each argument contains a unique ID and the ID of the attacked argument in the previous depth. For example, an argument in depth 2 may have the id *depth2_6* and the ID of the attacked argument *depth1_34* which means argument *depth2_6* attacks argument *depth1_34*.

The GitHub repository also contains the python code to generate a visual network graph using the `pyvis` library. The resulting visualisation displays the arguments when hovering over the nodes and is shown in Figure 5.3 (a higher resolution picture is available in the GitHub repository) [cor,].

5.4 Evaluation

I evaluated my generated argument corpus by randomly creating 24 dialogues by following the arcs of the graph, starting from the root and following each of the 16 arcs from the root to the argument in depth 1 at least once. This way, I ensured to create at least 16 completely distinct dialogues. I divided the 24 dialogues into four surveys using Google Forms and recruited 20 participants for each survey to judge the six given dialogues. An example dialogue is given below.

Figure 5.2: Visualisation of the generated argument corpus in graph form



PERSON A: *Universities in the UK should continue charging students the 9k tuition fee per year.*

PERSON B: *Education should be available for everyone, not for only ones who can afford it.*

PERSON A: *People who can't afford to have government help. Government can't afford free education for all unless they increase the taxes, and people won't like it.*

PERSON B: *The government are still paying for the loans and probably won't see the money back when the loans are written off in 30 years' time. Cheaper education and higher taxes are more sustainable than relying on students to pay back the loans, which they won't.*

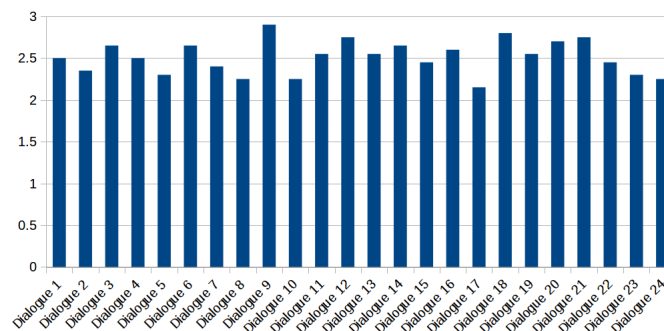
PERSON A: *The government should step out then and leave it to the banks to take the risk. Anyways, with higher taxes and cheap education, there would be plenty of educated unemployed to pay*

I informed the participants that the study involved judging transcripts and that the given dialogues involved two parties arguing whether tuition fees in the UK should be kept at 9000£. Party A believed they should be kept, and Party B believed they should be abolished. I instructed them to judge six transcripts plus an additional one playing the role of an attention check to ensure the honesty/quality of the provided responses. I asked them to score the transcripts in respect of each party staying to the point and defending their point of view. I asked them to not judge the dialogues by whether they believed the presented arguments since I was only interested in the overall quality of the dialogue (whether they make sense and parties sticking to their point of view). The participants were given a choice of three:

1. *Both parties don't stick to the point and don't defend their point of view*
2. *Both parties somewhat stick to the point and somewhat defend their point of view*
3. *Both parties do stick to the point and do defend their point of view*

On average, each dialogue scored 61% for option 3 (both parties sticking to the point and defending their point of view), 29% for option 2, and only 10% for option 1. Figure 5.4 shows the score for each dialogue, option 1 (*don't*) receiving score 1, option 2 (*somewhat*) receiving score 2 and option 3 (*do*) receiving score 3. The average score per dialogue was 2.51, which shows that the dialogues were of good quality and that if following a path in the graph,

Figure 5.3: Scores for each individual dialogue.



the resulting dialogue makes sense despite the individual arguments being collected from different people.

5.5 Discussion and Conclusions

In this chapter, I introduced a methodology to acquire a corpus of arguments for dialogues and present a corpus for research for computational argumentation, natural language processing, and chatbot knowledge base construction. Apart from checking for spelling mistakes, I have not conducted any further quality assessment of the arguments and have not checked for duplicate arguments in the argument graph. This gives researchers the possibility to use my corpus for research in methods like:

- Argument similarity assessment [Boltužić and Šnajder, 2015b, Misra et al., 2016]: many arguments in the graph support the same idea and are fairly similar. However, one can say the same thing in completely different ways, and clustering arguments by their similarity is a challenging but potentially valuable task.
- Argument quality assessment [Wachsmuth et al., 2018, Wachsmuth et al., 2017a, Wachsmuth et al., 2017b]: After clustering similar arguments together, one could apply some quality assessment to decide which argument in the cluster is the “best” according to some criteria (e.g. convincingness [Habernal and Gurevych, 2016a]).
- Establishing more attack (and support) relationships between arguments in the graph [Cocarascu and Toni, 2017, Chalaguine et al., 2018]: After identifying similar arguments, one could establish more attack relationships in the graph. For example, if arguments A and B are the same (just differently phrased), the counterarguments of A also attack B and vice versa.

By applying the methods above, high-grade chatbot knowledge bases could be created that contain only arguments of the highest quality (however one chooses to assess that) and many possible arguments for that domain. I also evaluated my corpus's quality and believe publishing it will give researchers a resource to explore the topics mentioned above, which will facilitate further research in these areas.

In the next chapter, I will present a chatbot that uses the argument graph described in this chapter as a knowledge base.

Chapter 6

A Persuasive Chatbot using a Crowdsourced Argument Graph and Concerns

In this chapter, I present a free-text chatbot that can engage in an argumentative dialogue to persuade the user to accept the chatbot's stance. The chatbot is equipped with a crowdsourced argument graph with automatically assigned concerns to each argument and a concern classifier that can assign concerns to the user arguments during the chat. With the help of this chatbot, I show that it is not necessary to follow the arcs of a graph during each dialogue move (i.e. give the user only the choice to provide arguments to the chatbot's arguments that are explicitly labeled with attack relationships in the graph) to create reasonable and relevant dialogues and that concerns can be automatically detected and used to choose appropriate counterarguments to increase the persuasiveness of the dialogue.

The rest of the chapter is structured as follows: Section 6.1 briefly recaps my previous work that this chapter study builds upon; Section 6.2 gives the aim of

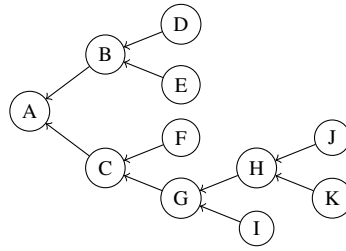
the chapter and the hypotheses; Section 6.3 describes the chatbot architecture that was used for the experiments; Section 6.44 describes the experiments that were conducted with the chatbot including their results, and in Section 6.5 I discuss and conclude my findings.

6.1 Introduction

In the previous chapter, I presented a method and evaluation for acquiring a large argument graph with over 1200 arguments via crowdsourcing. In this chapter, I offer a chatbot that uses that graph to persuade a human agent to accept the chatbot’s stance by presenting arguments from the graph that support its stance and counter user arguments that do not. One way to utilise such a graph is by using a *menu-based* approach where the chatbot, after presenting an argument, gives the user a choice of counterarguments that the user can select from a menu [Hadoux and Hunter, 2019]. Taking the argument graph shown in Figure 6.1 as an example, the chatbot would give argument A and then give the user arguments B and C to choose from. Suppose the user prefers argument C and selects that one. The chatbot selects a counterargument based on some criteria (or randomly) and replies with argument G and gives the user arguments H and I as countering choices, and so on. This way, the chatbot and the user would follow the arcs of the graph until (depending on the type of graph) all the arguments are used, or the user chooses an argument that has no counterarguments in the graph.

The menu-based approach’s drawback is that the user is limited to the choice of possible counterarguments presented by the chatbot, which might not include the user’s preferred choice. This might limit the persuasive effect of the argumentative dialogue and deny the chatbot the opportunity to acquire novel arguments on that topic that were not collected during the acquisition phase of the graph. The user arguments from the chats could then be used to extend the

Figure 6.1: Argument graph where child nodes are attacking parent nodes.



existing argument graph.

An alternative to the menu-based approach would be a *free-text* system that allows using a similarity measure to find an argument in the graph similar to the user argument. If an argument similar to the one used by the user is present in the graph, the chatbot could simply reply with a counterargument from the graph. Taking the graph from Figure 6.1 again as an example, the chatbot would present argument A and allow the user to reply via free-text input. The user would counter with an argument similar to H. Suppose the chatbot counters it with K, and the user replies with an argument similar to B. The chatbot could counter it with D or E and so on. In this case, the chatbot can jump around the graph rather than just following a single branch.

However, this poses two questions for the free-text approach: firstly, how to deal with a user argument that is not present in the graph. Not finding a match to the user's argument can be expected to be a common phenomenon, given that all arguments on that topic are contained in the graph cannot be assumed. The versatility of natural language, with its seemingly infinite number of ways to rephrase something, is also likely to limit the ability of the chatbot to find a similar argument in the graph. And secondly, even if the user's argument is present in the graph, which counterargument to choose to increase the persuasive effect of the dialogue.

A potential answer to the first question is for the chatbot to present an argument that is not necessarily a counterargument to the user's argument. This way,

the dialogue would resemble argumentation as it would happen in real life between two people: if two human agents engage in an argumentative dialogue, just because one presents an argument the other cannot counter, the dialogue is not necessarily ended prematurely. The other agent might switch topics and introduce a new argument they believe in without referencing and directly countering the previous argument. Another example would be product reviews, where reviewers present a range of pro and con arguments. The judgement is not about whether all counterarguments were answered or not, but whether the pro arguments outweigh the con arguments.

An answer to the second question could come from taking the *concerns* of the user into account [Chalaguine et al., 2019, Hunter et al., 2020, Hadoux and Hunter, 2019], a concern being a matter of interest or importance to the user. During the chat with a chatbot that allows free-text input, however, the concerns addressed by the user arguments need to be classified during the chat to choose a suitable counterargument accordingly. The chatbot, hence, needs to be equipped with a concern classifier.

Rosenfeld and Kraus presented methods for designing automated agents for human persuasion through argumentative dialogues without assuming a pre-defined protocol [Rosenfeld and Kraus, 2016b, Rosenfeld and Kraus, 2016a] which partially resembles my work but heavily relies on manual labour instead of using NLP. In both works, they rely on an expert human annotator and on a corpus of chats on which a machine learning model could be trained to obtain a probability distribution of the possible persuadee's argumentation frameworks. During the experiments in both works, they identify arguments used by the students by using a "Wizard of Oz" methodology, where during the chat, a human expert mapped each of the persuadee's sentences into an argument from the knowledge base, hence again assuming that all user arguments can be mapped to a finite set of arguments. Also, despite criticising the current state

of the art of persuasion strategies for adhering to strict protocols and relying on unrealistic assumptions, they assume in [Rosenfeld and Kraus, 2016b] that the persuader is aware of all the existing arguments the persuadee might give, justifying it by claiming that this asymmetrical situation is common when the persuader is an expert in the discussed issue, and the persuadee is not.

In this work, I want to address the shortcoming of the work discussed above and tackle the problem of computational persuasion *without* any prior knowledge of the persuadee and *without* the need for manual annotation that requires a human expert.

6.2 Hypotheses

In this study, I chose UK university Fees as a case study. I developed a chatbot that utilised a crowdsourced argument graph described in the previous chapter as the knowledge base. The chatbot used concerns to make strategic moves to engage in argumentative dialogues with users to persuade them to accept the chatbot's stance (that university fees should be kept).

Given this setting, I wanted to test two questions: Firstly, whether the crowdsourced argument graph can be used as a chatbot knowledge base that allows free-text input. This means that the graph contained at least some common arguments that the user might use, and the resulting dialogues were therefore of an appropriate length and quality, and the users perceived the chatbot's arguments as relevant. And secondly, whether the chatbot can automatically identify the concerns addressed by the user argument and whether replying with counterarguments that address the same concern, increases the persuasiveness of the chat. I summarise these points in the following two hypotheses:

- H1** A crowdsourced argument graph can be used as a knowledge base for a persuasive chatbot allowing free text input by the users. The resulting

chats are of appropriate length and quality, and the chatbot arguments are perceived as relevant by the users.

H2 A concern raised or addressed by a given user argument can be automatically identified to give appropriate counterarguments that address the same concern and thereby increase the persuasiveness of the dialogue.

In the remainder of this chapter, I describe the design of my chatbot that was used for the argumentative dialogues and explain the experiments conducted with the chatbot to test my hypotheses.

6.3 Chatbot Design

I developed two versions of my chatbot to test my hypotheses: one that classified the concern of the user argument and took it into account when presenting counterarguments (strategic), and one that did not (baseline).

6.3.1 Argument Graph

The argument graph described in the previous chapter is used as the chatbot's knowledge base. I only used depths 1-4, since depth 5 does not have any counterarguments. Depths 1 and 3 contain arguments against keeping university fees, while depths 2 (attacking depth 1 arguments) and 4 (attacking depth 3 arguments) have arguments that support the stance of keeping university fees.

When the user typed in an argument (source argument), the chatbot used a similarity measure to find the closest match of the user argument in the graph (target argument). The argument embeddings were created using GloVe word embeddings [Pennington et al., 2014], and I used cosine-similarity as a similarity measure [Singhal, 2001]. Cosine similarity is a metric used to measure how similar the vector representation of two texts is. It calculates the

cosine of the angle between two vectors. The smaller the angle, the higher the cosine similarity. I used a threshold of 0.9 for measuring the similarity of the two arguments. If the chatbot found an argument in the graph with a similarity of 0.9 or above compared with the source argument, the chatbot chose one of the counterarguments that attacked the target argument in the graph as a response. This happened at every dialogue turn, meaning that the target argument could be either in depth 1 or depth 3 of the graph.

6.3.2 Default Arguments

In case no target argument was found, I also acquired arguments for keeping university fees, where the root statement is the opposite of my main argument graph “*University fees in the UK should be abolished*”. It is, therefore, a very shallow graph with only one level of depth where the arguments that attack the root argument are for keeping the fees. I also used crowdsourcing for the acquisition and voting to select the best arguments. The best seven arguments were used as *default* arguments, which the chatbot could use if no match was found. These arguments were therefore not counterarguments in the traditional sense, as they did not refer to or address the source argument but instead “changed topic” and presented a new issue in the debate. I also added phrases like “*Ok but*”, “*I still think*” and “*Don’t you think that*” to the beginning of the default arguments to indicate a deviation from the topic occurs.

6.3.3 Concern Labelling and Classification

The baseline chatbot used the argument graph and default arguments during the chat with the user and did not make use of concerns. The strategic chatbot, however, classified the concern of the source argument and chose one of the attackers of the target argument that addressed the same concern.

Table 6.1: Types of concern for the topic of charging university tuition fees

Concern	Description of what concern deals with
Student Finance	Finances of students, including tuition fees, student debts, life costs etc.
Government Finance	Government finances, including general taxation, government spending etc.
Employment	Careers and employability of students and the general job market.
Free Education	Whether higher education is a human right and should be free or not.
Fairness	Whether something is fair or not (using a general understanding of fairness), including equal and just treatment of individuals.

During the acquisition of the argument graph described in the previous chapter, only arguments were included in the graph that contained *topic words*. These were words that I considered meaningful in the given context. The choice of suitable topic words depends entirely on the domain, and their choice is left to the researchers' discretion and their knowledge of the domain. The topic words in the argument graph were: *loan, debt, job, tax, free, accessible, affordable, government, scholarship, interest, career* and *background*. I grouped topic words that addressed the same or similar issues into 5 concerns: **Student Finance** (loan, debt, scholarship, interest), **Government Finance** (government, tax), **Employment** (job, career), **Free Education** (free) and **Fairness** (affordable, accessible, background). Apart from the concern *free*, the concerns were taken from [Hunter et al., 2020]. The definitions are given in Table 6.1.

I took the arguments from the argument graph and the user arguments from the chats with the baseline chatbot that contained any of the topic words to train a concern classifier using the Python Scikit-learn library¹. The classifier used logistic regression and a tf-idf feature representation to predict the concern of the incoming user argument. I extracted the top two concern predictions. If the top prediction was over 0.7, the argument was labelled with one concern;

¹<https://scikit-learn.org>

otherwise, with two. If a target argument in the graph was found, the chatbot chose one of the attackers of the target argument that addressed the same concern as the counterargument. If a user argument was labelled with two concerns, an attacker was chosen that addressed one of the concerns, with priority given to the concern with the higher predicted value. If no attacker existed that addressed the same concern or that argument was previously used up, an attacker was chosen that addressed the lower-ranked concern.

It could be argued that since the arguments in the graph were labelled with concerns, the source argument addressed the same concerns as the target argument in the graph. Hence, no classifier was needed as one could take the concerns of the target argument. However, the concerns of the target argument were not necessarily the same as the user's free-text argument, despite being similar. For example, the target argument in the graph "*Universities should be accessible to all, not just those that can afford it, or are not scared away from the high debt after their studies*" would be labelled with both concerns *fairness* and *student finance*. A similar source argument "*Universities should be accessible to everyone who wants a higher education, not just those that can afford it*" does not address the concern *student finance* and would be labelled with *fairness* only by the classifier.

If no match in the graph was found or none of the counterarguments of the target argument addressed the same concern, the chatbot replied with a default argument.

6.4 Evaluation of the Chatbot

The chatbots were deployed on Facebook via the Messenger Send/Receive API. For more on the implementation of such chatbots, see [Chalaguine and

Hunter, 2018b]. For each chatbot, I recruited 50 participants via Prolific². Before the chat, the users were directed to a Google Form and asked whether they *strongly disagreed*, *disagreed*, *neutral*, *agreed* or *strongly agreed* that university fees should be kept³.

After submitting their answers, they were redirected to the Facebook page where they could begin the chat. The chatbot started the chat by asking why the user believed that university fees should be abolished. The user, therefore, presented their first argument. The chatbot then replied with either a counterargument from the argument graph or a default counterargument, depending on whether a similar argument was found in the graph or not. If a similar match was found, the baseline chatbot replied with a randomly selected counterargument from the direct attackers of the target argument in the graph. The strategic chatbot, however, selected an attacker from the graph that addressed the same concern as the user argument (if such an argument existed). If no match was found, both chatbots replied with a default argument.

If the user's response was shorter than 6 words, the chatbot queried the user to expand on their answer. However, if the user agreed with an argument the chatbot gave, for example, by sending "*I agree*", the chatbot would not ask to expand despite the message being shorter than six words, and instead replied with a default argument.

The chatbot would eventually end the chat as soon as all default arguments were used up and no match in the graph was found. The users were, however, advised that they could end the chat anytime by sending the word "*stop*". At the end of the chat the chatbot presented the user with a link that redirected them to a second Google Form where they were asked a series of questions:

²<https://prolific.co>

³For the baseline chatbot only two people selected agree and none for the strategic one. 98% of participants therefore did not share the chatbot's stance before the chat

Table 6.2: Answers to first three questions for baseline and strategic groups: whether participants felt understood (Q1), whether the chatbot’s arguments were relevant (Q2), and whether participants felt like all their points were addressed (Q3).

Chatbot	Q1			Q2			Q3		
	Yes	No	Sometimes	Yes	No	Some	Yes	No	Some
Baseline	16	4	30	21	3	26	13	15	22
Strategic	15	6	29	31	1	18	10	14	26

1. Did you feel understood by the chatbot? (Yes/No/Sometimes)
2. Did you feel that the chatbot’s arguments were relevant? (Yes/No/Sometimes)
3. Do you feel like all your points were addressed? (Yes/No/Some of them)
4. How much do you agree that fees in the UK should be kept as they are? (Strongly disagree - strongly agree)

Questions 1-3 were used to test my first hypothesis and judge the chats’ relevance, length and quality, and question 4 was to test my second hypothesis and compare the persuasiveness of the baseline chatbot to the strategic chatbot. Table 6.2 shows the results for the first three questions for the baseline and the strategic groups. One can see that the majority of the participants considered the chatbot’s arguments as relevant in most cases and answered the first three questions with either *yes* or *sometimes*. Interestingly there is a 50% increase in the perception of relevance for the strategic chatbot, while the numbers for questions 1 and 3 remained almost the same. This is a statistically significant difference with a p-value of 0.045 using Chi-Square. Using concerns, therefore, makes the arguments more *relevant*.

Regarding questions 1 and 3, given that the chatbot does not use natural language generation and only relies on finding matches in an argument graph, it cannot address novel arguments or expand on existing ones by giving more information. It is therefore not surprising that the numbers for these questions remained approximately the same. The results, however, are promising, given

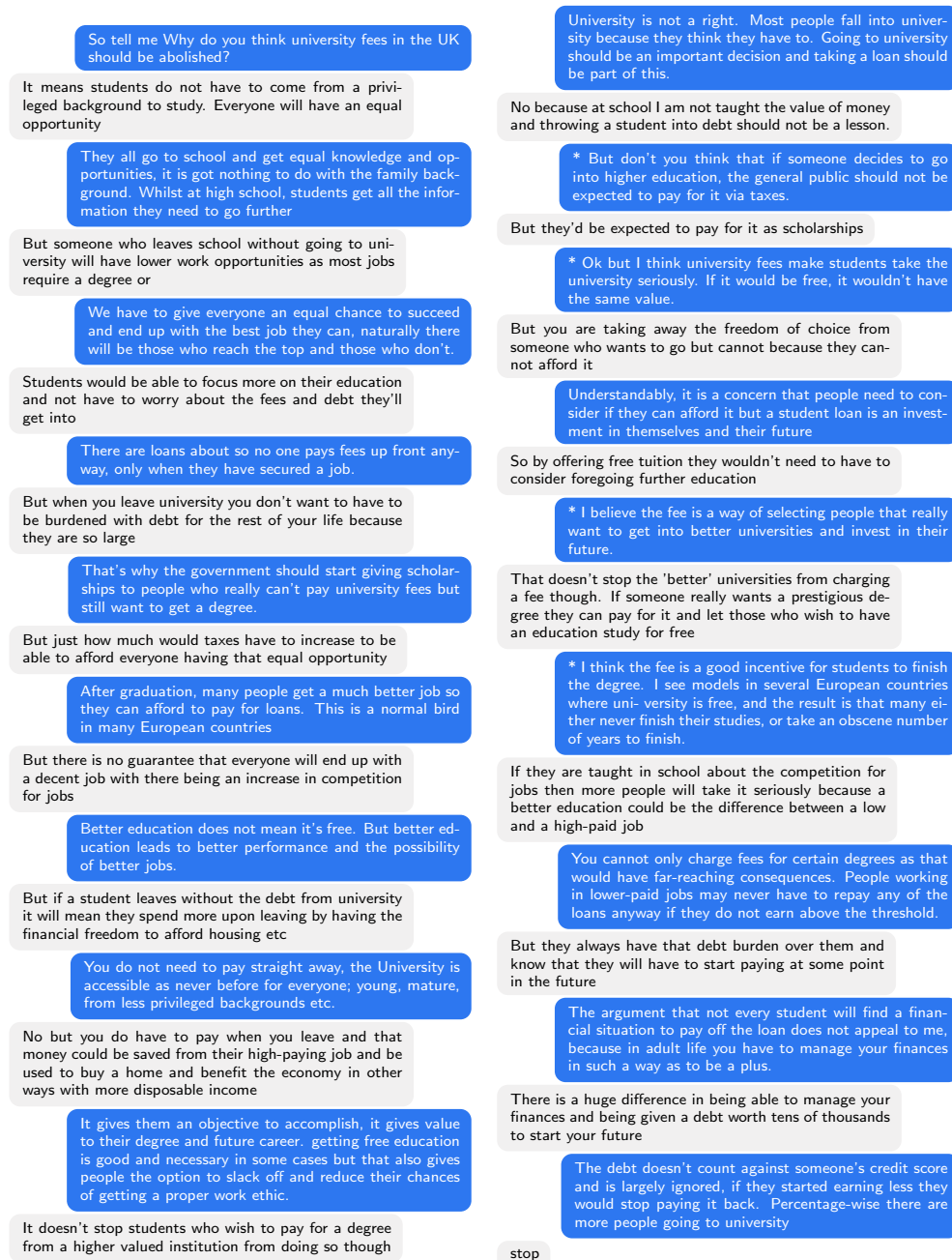
that the chatbot solely relies on a similarity measure to pick counterarguments from a crowdsourced graph or to pick default arguments in case no match is found.

Regarding the length of the chats, they, on average lasted 24 turns (disregarding the chatbot querying the user to expand). This means that the chatbot, on average, gave 12 arguments, 7 of which were default arguments and five from the graph. This supports my first hypothesis that a crowdsourced argument graph can be used as a chatbot knowledge base and that the resulting argumentation dialogues are of good length and quality, with the perceived relevance of the arguments being 50% higher during chats with the strategic chatbot. An example of a chat can be seen in Figure 6.3. All chatlogs and the code for the chatbot can be found in [git, c].

Using the replies for question 4, I calculated the change in stance for both groups to evaluate the persuasiveness of the strategic chatbot. The persuasion goal of the chatbot was that UK university fees should be kept. Since participants were given the choice of 5 stances (“strongly disagree” to “strongly agree”) before and after the chat, they could either change their stance for the better, for the worse or not at all. The change in stance is the final stance minus the original stance. I call the units of this measure *change in stance* (CS) points. For example, “strongly disagree” being 1 and “strongly agree” being 5, if a participant changed her stance from “disagree” (2) to “neutral” (3) after chatting with the bot this counts as 1 CS point, whereas changing from “neutral” to “disagree” counts as -1 CS point. Persuasion means that a person who was against keeping university fees before the chat changed her stance to the positive and that her CS points score is positive.

Table 6.3 shows the number of participants who changed their stance to the worse (negative), to the better (positive), and who did not change their stance

Figure 6.2: Example chat between a participant and the strategic chatbot. Chatbot arguments are in the dark boxes and user arguments in the light boxes. Default arguments are indicated with an *. The chat begins in the left column and continues in the right one.



at all (no change) for both chatbots, as well as the number of total CS points. One can see that 23 people changed their view for the better when chatting with the strategic chatbot with a total of 32 CS points, meaning that some participants changed their stance by more than 1 CS point (e.g. from *disagree*

Table 6.3: Change of stance (Negative (Neg), No Change (NC), and Positive (Pos)) measured by the number of participants and CS points.)

Chatbot	Baseline			Strategic		
	Neg	NC	Pos	Neg	NC	Pos
Change in stance						
No. of participants	5	41	4	1	26	23
Change in CS points	-5	0	5	-1	0	32

to agree). If counting the total number of CS points, including the participants who changed their stance for the worse, the strategic chatbot achieved a total change of 31 CS points, whereas for the baseline, the total number of CS points is 0.

It could be argued that a change from *strongly disagree* to *disagree* is not a remarkable change in stance despite resulting in the change of 1 CS point, whereas changing someone’s stance from *disagree* to *neutral* or even better, *agree* is a much stronger shift in stance. However, for the strategic chatbot, only 2 participants changed their stance from *strongly disagree* to *disagree*, while the remaining 21 participants changed their stance from disagreement (strongly or not) to neutral (16 participants), from neutral to agreement (3 participants) and from disagreement to agreement (2 participants).

I used the number of participants who positively changed their stance to calculate the statistical significance of the difference between the control group that chatted with the baseline chatbot and the group that chatted with the strategic chatbot using the Chi-Square test. All results were statistically significant, with a p-value of 0.00017. The results support my second hypothesis that concerns can be automatically classified based on the use of topic keywords which can be seen as a good indicator of the concerns being addressed or raised by the arguments. Presenting arguments that address the user’s concern is more likely to have a positive impact on their stance than presenting arguments that ignore the user’s concern.

6.5 Discussion

My contribution in this chapter is twofold. Firstly, I have shown that a crowdsourced argument graph can be utilised as a knowledge base for a chatbot that engages in argumentative dialogues. The resulting chats were of good length and quality and were perceived as relevant by the users. And secondly, I have shown that concerns can be automatically identified to give suitable counterarguments that address the same concern and thereby significantly increase the persuasiveness of the dialogue. Additionally, I have shown that the chatbot can jump around in the graph without systematically following each arc and only use arguments that are connected via an attack relationship.

To date, at least two arguing chatbots have been presented in the literature: a chatbot Debbie that uses a similarity algorithm to retrieve counterarguments [Rakshit et al., 2019] and Dave that uses retrieval- and generative-based models [Le et al., 2018]. My chatbot, however, differed in several ways: firstly, my knowledge base consisted of a previously crowdsourced argument graph. And secondly, Dave/Debbie aimed to keep the conversation going, whereas I was interested in persuading the user to accept my chatbot’s stance.

This study can be seen as a partial extension of the work in [Hunter et al., 2020] where a chatbot was used to persuade the user to accept the chatbot’s stance on university fees in the UK. The argument graphs used as the chatbot’s knowledge base were hand-crafted and manually labelled. The chatbot also did not allow free-text input and strictly followed the argument graph’s arcs. The chatbot presented in this chapter allows free-text input, uses a similarity measure to extract similar arguments from the graph, and does not restrict the selection of arguments to a single path in the graph. If a match is not found, the chatbot replies with an argument that is not contained in the original graph. My evaluation showed that this approach performed well and showed that it is not

necessary to and, in fact, often impossible to establish all possible relationships in a big argument graph. Therefore, instead of following a single path through the graph and only allowing the user to choose arguments present in the graph, one can search for a similar argument at each dialogue step without relying on a connecting arc between the new user argument and the previously given chatbot argument. And to avoid ending the chat prematurely if no similar user argument is found, default arguments can be introduced to keep the chat going.

I faced the additional challenge of having to automatically identify the concern of the user arguments during the chat. I showed that by grouping the most common meaningful words of the argument graph (topic words) into concerns, one could train a concern classifier on the graph arguments that can be used by the chatbot to improve its persuasive effect.

The advantage of using a crowdsourced argument graph as a knowledge base is that it does not require professional research but solely relies on the input of participants and can be acquired quickly. This method also scales easily, which allows obtaining many arguments from different people, thereby creating large and comprehensive argument graphs. There are, however, also potential risks to consider. For example, the spread of invalid arguments that, despite being popular, might contain wrong information. In the next chapter, I present a chatbot for a domain that requires strict assessment of the argument included in its knowledge base. Crowdsourcing the chatbot's arguments was unsuitable for the chosen domain and needed the arguments to be hand-crafted. I, therefore, present a hybrid approach where only the user arguments were crowdsourced, but the chatbot's arguments were researched and hand-crafted.

Chapter 7

Addressing Popular Concerns regarding COVID-19 Vaccination with Natural Language Argumentation Dialogues

In this chapter, I present a chatbot that engaged in persuasive dialogues with users who were reluctant to get a COVID-19 vaccine. I show that given a novel domain, like the COVID-19 pandemic and the associated vaccine development, it requires a relatively small repository of counterarguments to address the majority of possible arguments people might have for not getting the vaccine. This chapter is structured as follows: In Section 7.1 I give some background on the COVID-19 pandemic and introduce the chatbot used for this study; Section 7.2 gives the aim of the study and the hypotheses; Section 7.3 describes the chatbot architecture that was used for the experiments; Section 7.4 describes the experiments that were conducted with the chatbot, Section 7.5 presents the results, and in Section 7.6 I discuss and conclude my findings.

7.1 Introduction

During events like the global COVID-19 pandemic, it is even more important than usual that people are well informed and make conscious decisions that benefit themselves and society. One such example is the willingness to get a vaccine. Vaccines have historically proven to be highly successful and cost-effective public health tools for disease prevention [Rémy et al., 2015]. But the effectiveness of a vaccine in controlling the spread of COVID-19 depends on the willingness to get vaccinated in the general population. A sufficiently high vaccine coverage may generate herd immunity, which will protect everyone, including those particularly susceptible to the virus [Fine et al., 2011]. However, a barrier to reaching herd immunity is the prevalence of people who refuse or are hesitant to take vaccines [MacDonald et al., 2015, Sherman et al., 2021]. For example, the most recent numbers from YouGov surveys on vaccine hesitancy from late March 2021 showed that whereas the numbers in the UK are quite high (around 86%), the numbers in, for example, neighbouring France are much lower, at only 49%. In the USA they were a bit higher, at around 59% [you,]. Interviewing all those people who refuse to vaccinate in person and trying to convince them to get the vaccine, would be an impossible task.

This problem can be tackled as an argumentation problem: Arguments can be used to provide information and overturn misconceptions [Hunter, 2018]. They are an essential part of sensible discussions on controversial and problematic topics. However, despite an increasing body of literature on computational models of argument, there is still a lack of practical applications. Conversational agents, also known as chatbots, have the potential of being used as dialogical argumentation systems for behaviour change applications by applying computational models of argument. A chatbot could engage in an argumentative dialogue with people over the internet from the comfort and

safety of their own home, trying to persuade them to get a COVID-19 vaccine. Chatbots could thereby offer a cost-effective and scalable alternative to in-person consultations with health professionals. In order to represent arguments and concerns in the chatbot's knowledge base, I introduced a variant of an argument graph called a *concern-argument graph*.

The problem with controversial topics like the COVID-19 vaccine is that people block out information they disagree with: by creating social media echo chambers, reading partisan news, or only surrounding themselves with like-minded people. A recent paper [Meppelink et al., 2019] found that people select more belief-consistent information and perceive belief-confirming information as more credible, useful, and convincing when searching for online health information, also known as *confirmation bias*. Therefore, a dialogue with someone (or something - like a chatbot) could expose people to new information and potentially have a positive effect on their decision-making process.

However, different people worry about different things and hence arguments for not getting a COVID-19 vaccine will vary in a population. One person might be worried about the potential side effects of a newly developed vaccine, whereas someone else might think that he or she does not need a vaccine because they are young and healthy. A chatbot could address those different concerns by providing counterarguments tailored to the different user arguments and during the course of an argumentative dialogue, try to persuade the user to change their stance about getting vaccinated.

In this study, I used a hybrid approach where I crowdsourced the arguments that people have for not taking a COVID-19 vaccine and hand-crafted the counterarguments for the chatbot ourselves to avoid including invalid or emotionally-loaded arguments in the argument graph. Crowdsourcing arguments provides

insight into the reasons for people’s behaviour, and an indication of what sort of free natural language input the chatbot needs to handle.

I refrained from crowdsourcing counterarguments because COVID-19 is a serious health issue with global impacts. It is a new domain and research on this disease contains many uncertainties, and scientists still do not know many crucial aspects regarding the virus and its transmission characteristics. Moreover, it is the first time that humans are being injected with vaccines based on mRNA technology. All these form a highly uncertain information landscape and I believe that a carefully curated knowledge base is better than a crowdsourced one in such a critical communication framework.

7.2 Hypotheses

In this chapter, I present a chatbot that utilises a set of arguments for taking a COVID-19 vaccine as a knowledge base. The chatbot uses concerns to make strategic choices of moves in order to engage in argumentative dialogues with users to persuade them to get the vaccine.

Given this setting, I want to address the following three questions: Firstly, whether a small and shallow argument graph is enough to counter the majority of arguments people might have for not getting the vaccine and thereby create persuasive dialogues. Secondly, whether it is possible by only identifying the concern of a user argument to give a suitable counterargument. And finally, whether an interactive chatbot is more persuasive than a static web page that presents arguments for getting the vaccine. I summarise these points in the following three hypotheses:

- H1** Given a novel domain (i.e. a domain which is relatively new to the user of the chatbot and for which their knowledge and opinions might be limited), a small set of arguments (between 30-50 arguments) can be

used to represent most of the possible arguments that a set of normal users would know and appropriate counterarguments and can be utilised by a chatbot to create persuasive dialogues, meaning that the stance of the user changes after the chat.

H2 Given a novel domain, the arguments that address the same concern are sufficiently similar to allow for the provision of suitable counterarguments just by identifying the concern of the arguments.

H3 An argumentative dialogue with an interactive chatbot has a higher persuasive effect than presenting the same arguments on a static web page for people to read.

Further, I was also interested in whether during the chats, new concerns could be identified which were not raised in the crowdsourced arguments that were used to construct the chatbot's knowledge base.

I would like to note that I am neither psychologist, nor a health professional and that in this work I am (1) not taking any personality traits of the user attributes into account in order to evaluate what sort of argument might be more effective for this particular person, (2) do not compare the persuasiveness of the chatbot's arguments to other potential arguments, and (3) do not incorporate any other methods of persuasion apart from argumentation. My aim is to present a prototype chatbot that can be used to convince people to get a COVID-19 vaccine using argumentation and leave the aforementioned issues for future work.

In the remainder of this chapter, I describe the design of my chatbot that was used for the argumentative dialogues and explain the experiments conducted with the chatbot in order to test my hypotheses.

7.3 Chatbot Design

In this section, I describe the acquisition of arguments used to construct the chatbot’s knowledge base and the concern classifier used by the chatbot to identify the concerns of the incoming user arguments.

7.3.1 Knowledge Base Construction

To construct the chatbot’s knowledge base which consists of a concern-argument graph, which I define below, I recruited 100 participants via Prolific¹ and asked them to provide three arguments against getting a COVID-19 vaccine. I identified 7 concerns that were raised by the majority of the 300 crowdsourced arguments by inspecting the most common, meaningful words, namely: **short-term side effects** of the vaccine, **long-term side effects** of the vaccine, its **fast development**, the **mutation** of the virus, the **safety** of the vaccine, comparison of COVID-19 to the **flu** and downplaying its danger, and **young** people believing they do not need a vaccine.

The arguments given by users for each of the identified concerns were quite similar as the following example demonstrates:

Example 7.1 Three similar arguments that raise the concern about side effects of the vaccine.

- *I will not get a COVID-19 vaccine because of its potential side effects.*
- *It’s a new vaccine, so I don’t yet know what the side effects are.*
- *The vaccine may have a lot of side effects that could be more dangerous than COVID-19 itself.*

The arguments regarding side effects were very similar, and could all be countered with the same argument, for example, that there is high scrutiny

¹<https://prolific.co>

over the research on those particular vaccines and nobody would allow giving it to the public if it was unsafe. Whereas, drafting individual counterarguments would make the graph unnecessarily big and might result in the inclusion of many similar counterarguments. I, therefore, structured the knowledge base as a *concern-argument graph* which I define as follows:

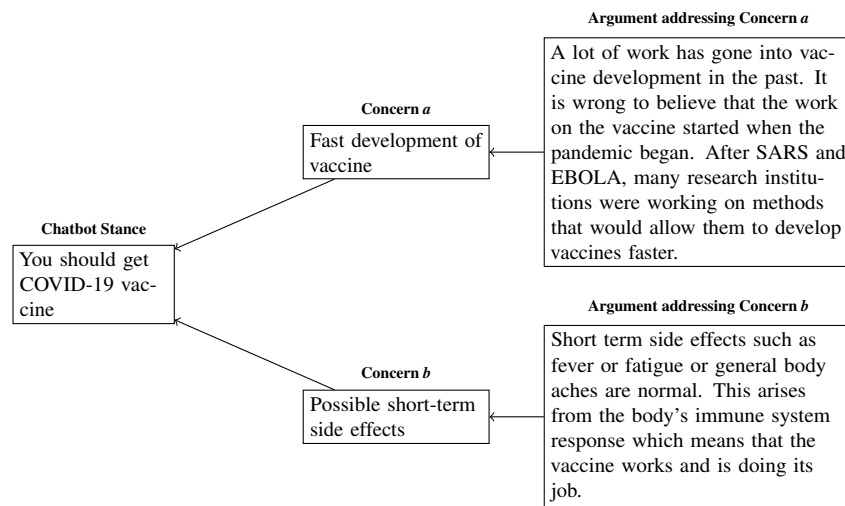
Definition 7.1 A **concern-argument graph** is an directed acyclic graph (N, E, L) where N is a set of nodes, E is a set of edges, and L is a labelling function such that

- N can be partitioned into a set of arguments A and a set of concerns C (i.e. $N = A \cup C$ and $A \cap C = \emptyset$);
- the root of the tree, denoted ρ , is an argument in A ;
- for each argument $\alpha \in A \setminus \{\rho\}$, the labelling function assigns a set of concerns to α (i.e. $L(\alpha) \subseteq C$);
- E is the smallest set satisfying the following conditions:
 - for each concern $\sigma \in C$, there is an edge $(\sigma, \rho) \in E$
 - for each argument $\alpha \in A \setminus \{\rho\}$, if $\sigma \in L(\alpha)$, there is an edge (α, σ)

So a concern-argument graph can be regarded as a compacted version of a three-level acyclic argument graph (i.e. an argument graph with an argument at the root, counterarguments to the root argument, and counter-counterarguments to the counterarguments). Furthermore, in the concern-argument graph, instead of counterarguments, I use concerns as place-holders for a set of similar arguments raising that concern. This information about concerns I will use in the dialogue strategy.

The resulting graph is, therefore, much smaller and shallower than the traditional argument graph that includes both arguments and counterarguments. Figure 7.1 shows an example of such a graph, where the goal argument is

Figure 7.1: Part of the chatbot’s knowledge base with concerns *a* and *b* representing clusters of arguments in level 1 (concerned about insufficient testing of the vaccine and possible side effects respectively) attacking the chatbot’s goal argument, and arguments in level 2 addressing the concerns.



being attacked by two concerns and each concern is being attacked by one counterargument. One could, of course, extend the concern-argument graph and include more concerns that attack specific counterarguments in level 4, and counterarguments that attack these in level 5. But I aimed to investigate whether a simple 3-layer model would be sufficient in this particular domain where people are more likely to present a new argument, instead of countering the chatbot’s counterargument.

The counterarguments for level 3 in the concern-argument graph were carefully researched using reliable sources. Some arguments included links to the NHS website that lists potential short-term side effects from the vaccine², or videos that explain how the vaccine was developed so quickly without compromising its safety³. As already stated above, I did not individually evaluate the persuasiveness of the arguments, as this is out of the scope of this study.

²<https://www.nhs.uk/conditions/coronavirus-covid-19/coronavirus-vaccination/coronavirus-vaccine/>

³All arguments were fact-checked by a medical doctor trainee and a machine learning consultant who works in the pharmaceutical industry

The chatbot also needed default arguments for getting a COVID-19 vaccine that it could use in case the concern of the user argument could not be identified. These arguments are therefore not counterarguments in the traditional sense as they do not directly counter the user argument but instead “change topic” and present a new, important issue in the debate. I also added phrases like “*Ok but*”, “*Have you considered that*” and “*Nevertheless*” to the beginning of the default arguments to indicate that a deviation from the topic occurs. This way the dialogue would resemble argumentation as it would happen between two people: if two human agents engage in an argumentative dialogue, just because one presents an argument the other cannot counter, the dialogue does not necessarily end at that point. The other agent might switch topics and present a new argument he or she believes in, without referencing and directly countering the previous argument. Two out of the three default arguments stated the importance of the vaccine in order to reach *herd immunity*.

My initial concern-argument graph consisted of 7 concerns, each with 2-4 counterarguments, with 19 leaves total. Additionally, the chatbot was equipped with three default arguments it could use in case no concern could be identified.

7.3.2 Understanding the user input

The initial move was the root argument presented by the chatbot. Then the user gave a counterargument which was analysed to determine the concern. The chatbot identified the concern of the user argument using a multinomial logistic regression and a binary feature representation of the arguments (one-hot encoded vectors)⁴. For the initial classifier, the crowdsourced arguments were used for training. If the prediction was over 40% in confidence, the

⁴Due to the small amount of data and the use of one-hot encoded vectors, there was no considerable difference when evaluating different classifiers. More sophisticated methods, like pre-trained language models, could not be used due to the lack of data that could be used for pretraining on the given topic.

argument was labelled with the identified concern, which was used by the chatbot to pick a leaf argument from the concern-argument graph that was a counterargument to that concern. If no concern could be identified, or all the leaves of the identified concern had been used, a default argument was presented to the user.

7.4 Experiments

The purpose of the chatbot was to test all three of my hypotheses. Prior to recruiting participants for the study, I ran a survey where I asked people to choose on a scale of 1-5 whether they would get a COVID-19 vaccine. The options were *very unlikely*, *somewhat unlikely*, *neither likely nor unlikely*, *somewhat likely* and *very likely*. I recruited 300 participants from those that chose *very unlikely*, *somewhat unlikely* and *neither likely nor unlikely*, i.e. those with a negative or neutral stance. 240 participants chatted with the chatbot, and a static web page was presented to 60 presenting the 10 most common arguments used by the chatbot .

Before the chat, the users were directed to a Microsoft Form and asked again how likely they would get a COVID-19 vaccine. After submitting their answer, they were redirected to a web page where they could begin the chat. The chatbot was composed of a front-end I coded in Javascript and a Python back-end using the Flask web server library. The chatbot started the chat by instructing the user that they could end the chat anytime by sending the word “*quit*” and then asking why the user would not get a COVID-19 vaccine once it became available to him/her. The user then presented his/her first argument. The chatbot replied with either a counterargument from the concern-argument graph or a default argument, depending on whether it could identify the concern of the user argument. The counterarguments were stored in a Python dictionary with the concerns as the keys and the list of counterarguments that addressed

that concern as the values. If the concern could be identified, the chatbot replied with the first counterargument in the list. If the message of the user was less than seven words in length and contained a negation, the chatbot queried *Why?* or *Why not?* to force the user to expand. This process was repeated with each subsequent argument given by the user. The chatbot would end the chat as soon as all default arguments were used up and no concern could be identified, or all counterarguments that addressed the concern were also used up. At the end of the chat, the chatbot presented the user with a link that redirected them to another Microsoft Form, where they were asked a series of questions:

1. Did you feel understood by the chatbot? (Yes/No/Sometimes)
2. Did you feel that the chatbot's arguments were relevant? (Yes/No/Some of them)
3. Do you feel like all your concerns were addressed? (The majority of them/None of them/Some of them)
4. How likely would you get a COVID-19 vaccine once one becomes available to you? (Very unlikely - very likely)

Questions 1-3 were used to test my second hypothesis and judge the chats' relevance, length and quality, and question 4 was to test my first and third hypotheses and compare the stances of the participants before and after the chat with the chatbot in order to judge persuasiveness. In order to test my third hypothesis, 60 out of the 300⁵ participants did not chat with the chatbot but instead were presented with the chatbot's ten most commonly used counterarguments in *persuasive* chats on a static web page.

To test whether new concerns could be identified during the chats that were not identified in the crowdsourced arguments, I analysed the chats after every batch of 60 participants (I also recruited participants in batches of 60). By inspecting

⁵Due to limited funding, I did not want to split the participants in half but rather collect more data in the form of chat logs since a web page does not provide data for further research.

Table 7.1: Breakdown of the 240 participants’ stance for getting a COVID-19 vaccine before and after chatting with the chatbot.

	Very unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Very likely
Before	30%	42%	28%	0%	0%
After	22.5%	37.5%	29%	9.5%	1.5%

common, meaningful words, I could identify new concerns after each batch of 60 participants, re-train the classifier with enough examples of the new concern, and add suitable counterarguments to the concern-argument graph. I only added a new concern to the chatbot’s concern-argument graph if I could automatically identify at least 10 arguments that addressed that concern. These arguments were added to the training set with the new label. I also identified new arguments for existing concerns. For example, in the crowdsourced arguments, many people used the word *mutation* (i.e. mutation of the virus) whereas in the chats, the word *strain* (i.e. new strain) was prevalent. These arguments were also added to the training set with the existing concern label. The concern classifier was then retrained on the updated training set.

This way 8 additional concerns could be identified: **death**, that the vaccine does not prevent you from **getting** and **spreading** COVID-19, people claiming they already **had** COVID-19, that COVID-19 has a too high of a **survival rate** to be worried about it, that the vaccine may impact **fertility**, that the vaccine might not be **effective**, that the **ingredients** of the vaccine are unknown, and that herd immunity can be created **naturally** by catching the virus and hence no vaccine is needed.

7.5 Evaluation of the Chatbot

Table 7.1 shows the stance of the 240 participants who chatted with the chatbot, before and after the chat. I divided the change in stance into three categories: a change from negative to neutral (from *very unlikely*/ *somewhat unlikely* to

Table 7.2: Percentage of the 240 participants who changed their stance after chatting with the chatbot.

Negative to Neutral	Neutral to Positive	Negative to Positive	Total
9%	7.5%	3.5%	20%

neither likely nor unlikely); a change from neutral to positive (from *neither likely nor unlikely* to *somewhat likely/very likely*); and a change from negative to positive (from *very unlikely/somewhat unlikely* to *somewhat likely/very likely*). I do not consider a change from *very unlikely* to *somewhat unlikely*. Table 7.2 shows the percentage of the 240 participants who changed their stance by engaging in an argumentative dialogue with the chatbot. 20% of the participants (48 out of 240) had a positive change in stance. This verifies my first hypothesis - that a chatbot can utilise a small, shallow concern-argument graph to create persuasive dialogues.

Given that the chatbot did not use natural language generation and could not address novel arguments or expand on existing ones by giving more information and relying solely on correct concern classification, the results are promising. The length of the chats was, on average, 12 alternating turns. This means that the chatbot, on average, gave six arguments, three of which were default arguments and three from the graph. Table 7.3 shows the results for the first three questions. 35% of the participants felt understood by the chatbot, and further 41% felt sometimes understood. 32% perceived the chatbot’s arguments as relevant, and further 55% perceived them as sometimes relevant. 23% felt that the majority of their concerns were addressed, and a further 54% felt that some of their concerns were addressed. This supports my second hypothesis that

Table 7.3: Answers to the first three questions by the 240 participants who chatted with the chatbot.

Felt understood			Relevance			Concern addressed		
Yes	S-times	No	Yes	Some	No	Majority	Some	None
35%	41%	24%	32%	55%	13%	23%	54%	23%

Table 7.4: Breakdown of stance for getting a COVID-19 vaccine of the group of 60 participants before chatting with the chatbot, and the group of 60 participants who was presented with a static web page (ul = unlikely).

	Very ul	Somewhat ul	Neither likely nor ul
Chatbot	60%	33%	7%
Web page	52%	28%	20%

Table 7.5: Change of stance for the group of 60 participants who chatted with the chatbot, and the group of 60 participants who was presented with a static web page.

	Negative to Neutral	Neutral to Positive	Negative to Positive	Total (no of participants)
Chatbot	12%	5%	2%	18% (11)
Web Page	5%	0%	0%	5% (3)

only by identifying the concern of an argument suitable counterarguments can be presented and that the resulting argumentation dialogues are of satisfactory length and quality. An example of a chat can be seen in Figure 7.2. All chatlogs, the data for the concern-argument graph, the concerns and their descriptions, and the code for the chatbot can be found on GitHub [git, d].

As mentioned in the previous section, I recruited the participants in batches of 60. To evaluate my third hypothesis, the chatbot’s persuasive effect compared to a static web page, I compared the results of the 60 participants who were presented with the static web page with a batch of 60 participants who chatted with the chatbot with similar starting distributions of their stance. The fourth (and last) batch of participants had a similar distribution to the batch recruited to read the arguments on a static web page. The starting distributions are shown in Table 7.4, and the change of stance for both groups is shown in Table 7.5.

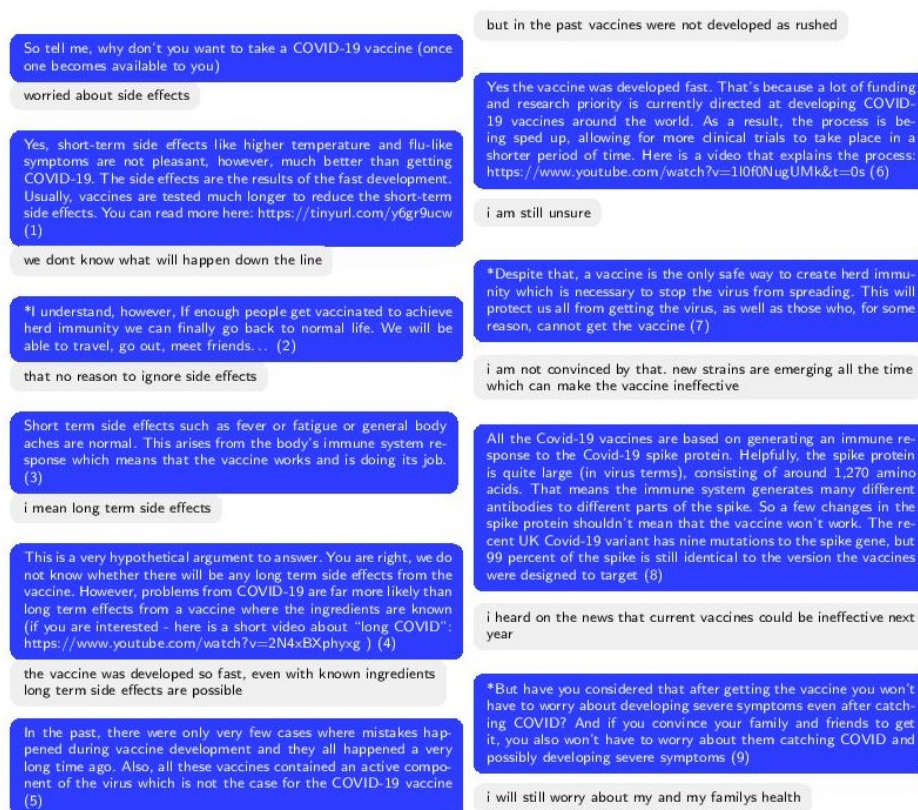
I used a Chi-Square test to compare the number of participants who changed their stance after chatting with the chatbot with the number of participants who changed their stance after reading the 10 most common arguments used by the chatbot (this means that the participants who only saw the static web page, on

average saw 4 more arguments than those who chatted with the chatbot). The results were statistically significant, with a p-value of .023 at $p < .05$. The results, therefore, supported my hypothesis that an interactive chatbot is more persuasive than a static web page.

The newly identified concerns mentioned in the previous section were also raised by some arguments that were collected during the initial argument collection described in Section 7.3.1. So claiming that new concerns could be identified during the chats would be incorrect in regards to this domain. Given a large enough sample of crowdsourced arguments, more concerns could have been identified and included (together with appropriate counterarguments) in the concern-argument graph that was used as the chatbot's knowledge base. However, this is not the case for all domains. Most chats were peculiar in that people were not countering the chatbot's arguments but coming up with new arguments (excuses to not get a vaccine) at each turn. In Figure 7.2 one can see that the user presents new arguments as a response to the chatbot's arguments (3), (4) and (7) (replies to arguments (5) and (8) are implied questions: "*Why were vaccines not developed as rushed in the past?*" and "*Is it true that the current vaccines could be ineffective next year?*"). This is not surprising given the novelty of the domain at that time, where the public was not as well-informed as on more established domains. For those domains, conversations with a chatbot have the potential to reveal new concerns, which are unlikely to be discovered during crowdsourcing for depth 1 of a concern-argument or traditional argument graph, as described in Chapter 5.

Side effects (short and long-term) are the most widespread concern, with 45% of user arguments given during the chats (where a concern could be identified) raising it. This is coherent with previous studies which analysed vaccine hesitancy in France [Ward et al., 2020], the US [Marco-Franco et al., 2020], and the EU [Neumann-Böhme et al., 2020]. The second most prevalent

Figure 7.2: Example chat between a participant from the fourth batch. Chatbot arguments are in the dark boxes and user arguments in the light boxes. Default arguments are indicated with an *. The chat begins in the left column and continues in the right one. The participant indicated that he or she was *somewhat unlikely* to get vaccinated before the chat but changed their stance to *neither likely nor unlikely* after the chat.



concern was about the vaccine's safety in general (28%), and in the third place, the vaccine's fast development and young people believing they do not need one (both 10%).

7.6 Discussion and Conclusion

This study aimed to present a prototype chatbot that can engage in persuasive dialogues with people opposed to the COVID-19 vaccine using computational models of argument. My contribution in this chapter is threefold. Firstly, I have shown that for a new domain, where there exists a lot of uncertainty, a small argument graph can be used to represent most of the possible arguments

in this domain. A chatbot can utilise this argument graph to create persuasive dialogues, and I presented a method how for acquiring and structuring such a graph in the form of a concern-argument graph. In my previous work, described in the previous chapter, the chatbot's knowledge base consisted of an argument graph that included both arguments and counterarguments. The chatbot matched the incoming user argument with a similar argument in the graph (target argument) using cosine similarity of the vector representations of the two arguments (the argument embeddings were created using GloVe word embeddings [Pennington et al., 2014]). The graph was therefore much bigger (containing over 1200 arguments) than the one presented in this chapter.

Secondly, I have demonstrated that no sophisticated natural language understanding of the user arguments is needed to provide suitable counterarguments that address the majority of the concerns of the users. And thirdly, I have shown that an interactive chatbot has a higher persuasive effect than a static web page.

Further, I have shown that in this domain, a concern-argument graph (a three-level acyclic graph) where after the initial move, the chatbot only picks a leaf argument at every turn or uses a default argument, is enough to generate persuasive dialogues. Using a concern-argument graph, as described in this chapter, and not constraint the chatbot to use a larger argument graph that may involve long paths has two main benefits: firstly, the graph can be constructed with less data than a larger argument graph; secondly, this allows the chatbot to counter user arguments that are not direct counterarguments to the previously given chatbot argument. This is important because, during the chats, people often ignored the chatbot's counterargument but gave a new argument on why not to take the vaccine. In some cases, it got to the point where people alleged they *might* be allergic to the vaccine without having a history of allergic reactions to vaccines.

And lastly, my method of analysing the incoming user arguments scales easily and allows obtaining many arguments from different people. A chatbot can address millions of people simultaneously in the comfort of their own homes and collect a vast amount of data in a very short time. The more data comes in, the easier it gets to identify patterns, discover new concerns, acquire arguments that address these concerns and update the chatbot's concern-argument graph accordingly. This allows us to identify common misconceptions and address the lack of information and potentially even fake news.

More data would also allow me to use more sophisticated natural language processing methods to process the user's input. Currently, the chatbot can only reply to well-phrased arguments that raise common concerns that it can identify. These sorts of replies contributed to only 50% of the users' replies. Other types of responses included novel arguments, statements like "*I don't care*" or "*I will take my chances*", emotional accusations (about the government not caring or the chatbot being stupid) and questions about the vaccine. A question-answering module for this chatbot is discussed in the next chapter.

Another reasonable extension to the current framework would be *bipolar argumentation* [Cayrol and Lagasquie-Schiex, 2005b, Amgoud et al., 2008a]. In Figure 7.2, the fourth argument that the chatbot presented addressed the concern of long-term side effects. The user replied with an argument that again raises long-term side effects. Hence, the fifth argument by the chatbot also addresses that concern. Argument 5 can, therefore, be seen as a supporting argument to argument 4. A potential dialogue strategy for the chatbot could be to use arguments that support the previously given argument by the chatbot if a concern cannot be identified instead of giving default arguments. For example, after the user said they were still unsure, the chatbot could have provided a supporting argument to argument (6) and presented the user with another

argument that addresses the fast development of the vaccine.

In the next and final chapter, I will give an overview of some more potentially interesting research areas which can be explored during future work.

Chapter 8

Conclusion

In this chapter, I summarise my contributions, outline some limitations of the presented work, and give future work directions.

8.1 Contributions

I have presented a framework for developing persuasive chatbots, which includes methods for the acquisition of arguments, ways to structure those arguments in the form of an argument and a concern-argument graph which serves as the chatbot's knowledge base, and local strategies for the chatbot to select moves that increase the persuasiveness of the chat using concerns.

A critical feature of the arguments which are acquired via crowdsourcing is their accurate reflection of ordinary people's concerns on the given topic. They are simple arguments that are nevertheless full of meaning and importance, and the kind of arguments that have been neglected in the formal as well as informal argumentation literature. There is little literature on how to analyse these sorts of arguments and even less on how to acquire them. In this thesis, I have shown how to acquire such arguments and how they can be used to structure the knowledge base for a chatbot that can engage in persuasive dialogues with

users.

I have presented two ways to structure the chatbot's knowledge base: (1) By using a traditional argument graph where the root node is the chatbot's stance, the attacking nodes in the first level of depth attack the root node, and the nodes in each subsequent level of depth attack the arguments from the previous level. The chatbot uses a similarity function to find an argument similar to the user's argument in the graph and replies with a counterargument that attacks that argument in the argument graph. (2) By using a concern-argument graph where the root node is attacked by concerns and concerns are attacked by counterarguments. The chatbot uses a classifier to identify the concern of the user's argument and replies with a counterargument that addresses the concern.

Another significant contribution of this work is the notion of concerns used by the chatbot to choose which argument to select in a dialogue to make it more persuasive. I have shown that concerns addressed by arguments can be automatically detected and that by presenting arguments that address the user's concerns, the persuasiveness of the dialogue increases, and the user is more likely to change their stance on the topic in question. I have tested that in three case studies with actual participants on the issues of meat consumption, UK university fees, and the COVID-19 vaccine. In all three studies, the chatbot that used concerns in his argumentation strategy was compared to a baseline where no concerns were used, and in all these studies, the baseline chatbot performed significantly worse.

Given that my research was backed up with case studies with actual participants, I have acquired several small datasets of chatlogs and a big argument graph with over 1200 arguments on the topic of UK university fees. These datasets, as well as the source code for the chatbots, are open source and can be used by

other researchers and hence amount to another contribution of this thesis.

And finally, in my last study, described in the previous chapter, I demonstrated how to acquire new arguments and concerns with the help of the chatbot, which can then be used to extend the argument-concern graph. Using crowdsourcing might not reveal all arguments/concerns that people might have (e.g. for not getting a COVID-19 vaccine). More arguments are collected during the chats, and more concerns are in the users' counterarguments to the chatbot's arguments. Hence, the chatbot has the potential to collect data that health care professionals and politicians can use to address these concerns and fight fake news. An overview of all three chatbots can be found in Table 8.1. The persuasion rate indicates how many participants changed their stance from either negative to neutral, neutral to positive, or negative to positive. Given we used a 5-point Likert scale for all our studies, the persuasion rate does not include a change from *very unlikely* to *somewhat unlikely*.

8.2 Limitations

Some limitations of the work presented in this thesis are given below.

- Quality assessment of arguments: An extensive quality assessment of the arguments acquired during my research was not within the scope of this study. However, should a chatbot like presented in this thesis be used in the real world, extensive quality assessment of the chatbot's arguments is needed to avoid the spread of inaccurate information. For a review of different features to assess argumentation quality, please refer to [Wachsmuth et al., 2017a].
- Not tested with unpaid participants: Although one can assume that the majority of people are reasonable and would reasonably engage with the chatbot in a real world setting, it is possible that the paid participants

Table 8.1: Overview of the three persuasive chatbots described in this thesis.

Topic	Meat Consumption	UK University Fees	Covid-19 Vaccine
No. of participants	39 ^a	50	240
Persuasion rate	46.15%	46%	20%
Strategy	The concern of the user was established prior to the chat and the chatbot only presented arguments that addressed that particular concern.	A similarity score was used to find arguments in the chatbot's knowledge base in the form of an argument graph that were similar to the users' arguments, and a concern classifier to retrieve counterarguments that addressed the concerns raised by the users' arguments.	A concern classifier was used to identify the concern that was raised by the users' arguments in order to retrieve counterarguments from the chatbot's knowledge base in form of a concern-argument graph to retrieve counterarguments that addressed the concerns raised by the users' arguments.

^aAdjusted to only those who had a negative or neutral stance

were incentivised to engage with the chatbot for longer, to ensure they get paid. One could expect that if they had not been paid, the chats might have been shorter and possibly less persuasive.

- No longitudinal studies: Although participants were asked about their change of stance after the chats there is no way to verify this without follow-up questionnaires to assess whether people really, for example, got a vaccine or reduced meat consumption, after indicating a shift in their stance in the questionnaires after the chats.

8.3 Future Work & Research Areas of Interest

In this section, I want to overview some potentially interesting research areas that can be explored during future work. The most significant limitation of my research was the lack of big corpora like previous chats on a particular topic which could have been used to design persuasion strategies or fine-tune language models. I collected all the data, and knowledge bases for the chatbots were built from scratch using the little collected data. With more resources, including money and human resources, more data could have been collected and generated (e.g. real people chatting with actual participants to generate chats) that could then be used to develop the chatbots. All chatbots were tested only on a small number of people, which, on the one hand, generated too much data for one person to go through and, for example, annotate, but at the same time did not generate enough data to train a generative chatbot on it. However, with enough resources, the research presented in this thesis offers many attractive future research opportunities, which I am outlining below. This is, by no means, an extensive literature review on the areas mentioned, but rather some pointers to other interesting research I came across while conducting the research presented in this thesis or believe could be valuable extensions to the current work.

8.3.1 Question Answering

During my final study, described in the previous chapter, participants often asked the chatbot questions instead of giving counterarguments. The chatbot could, of course, not answer those. However, a significant amount of the public's mistrust about the COVID-19 vaccine (but also applicable to other domains) may likely arise from misinformation, or ignorance [Chaudhary et al., 2021]. Therefore, the ability to be able to ask genuine questions to address subjects or areas where they may not have been previously exposed to the relevant information is likely to go some distance to placate individuals' concerns born of ignorance and hence increase the persuasiveness of the chats, making it a desirable feature for a chatbot of such kind.

Together with an MSc student, Gabriel Freedman, we extended the chatbot with a question-answering module as a proof of concept (the chatbot was never tested on real participants). The final system consisted of two primary modules: argumentative and question-answering. The system's structure regarding the argumentative module remained largely intact, with the method for concern classification remaining largely as before. A few small changes were made to the corpus of arguments to reflect the most recent changes in the scientific literature. The question-answering component was comprised of two core functionalities in addition to the concern classification stage. These were the capability to identify whether the inputs were questions or not and subsequently the ability to return reasonable answers if a question had been identified. The question-answering component utilised BERT (Bidirectional Encoder Representations from Transformers) [Devlin et al., 2019]¹, which is a transformer-based [Vaswani et al., 2017] encoder-decoder model used for a wide range of NLP tasks. To fine-tune BERT so that it has the capacity for answering questions about a body of text, a version was used that was trained

¹https://huggingface.co/transformers/model_doc/bert.html

on the SQuAD 2.0 dataset [Rajpurkar et al., 2016], a collection of over 150 thousand questions and answer pairs². A body of data has been collected from various internet sources to make up the foundation of the system’s knowledge. However, in contrast to the argumentative approach, the full body of text was not returned, as this approach would not offer enough precision. Instead, the model identified the most relevant passage from the larger body of text.

In the future, it would be desirable to evaluate the system with actual participants. One could also use a more extensive dataset for the question-answering component, for example, the COVID-19 Open Research Dataset (CORD-19), as was done in [Oniani and Wang, 2020] or COVID-Q presented in [Wei et al., 2020].

During the writing of the Master’s thesis, a paper was submitted to Arxiv in July 2021 titled “An Argumentative Dialogue System for COVID-19 Vaccine Information” [Fazzinga et al., 2021] which presents a *theoretical* chatbot that can answer user questions and provide explanations. The theoretical chatbot would have no persuasive element and solely serve the purpose of information seeking regarding COVID-19 vaccines. Further, the authors assume the presence of a scenario-specific knowledge base created by experts in the form of an argumentation graph. And lastly, they leave the representation of the user input as BERT-based sentence embeddings for future works.

8.3.2 Argument Mining & Argument Detection

I have explained before why user argument acquisition via crowdsourcing is necessary to build a corpus that can be used to construct the chatbot’s knowledge base. This is (1) because the topic of interest might have never been discussed on social media, and (2) the language and length of chat messages differ from other forms of media. However, for acquiring the chatbot’s argu-

²<https://huggingface.co/csarron/bert-base-uncased-squad-v1>

ments, argument mining would be an interesting topic. Argument mining is the automatic identification and extraction of the structure of inference and reasoning expressed as arguments presented in natural language [Lawrence and Reed, 2020]. For a review on argument mining please refer to [Lawrence and Reed, 2020] and [Cabrio and Villata, 2018].

One could, for example, apply a similar technique as in [Xu et al., 2020] where the authors used argument mining for legal text summarisation to generate what they called *legal argument triples* which consisted of the major issues a court addressed in a case, the court's conclusion concerning each issue, and characterisation of the court's reason for reaching a conclusion and hence provided a summary of the legal case. Given that during a chat, the chatbot's arguments cannot consist of very long messages, applying a similar approach to long texts (e.g. a scientific text that argues for the safety of the COVID-19 vaccine) is needed. Alternatively, one could use argument mining on some large corpora like a web crawl to create a pool of arguments to choose from instead of researching them manually. The argument search engine ArgumenText is presented in [Stab et al., 2018] and was used in [Rach et al., 2021] to acquire and structure arguments for dialogue systems.

An interesting claim is made in [Galitsky et al., 2018] stating that the accuracy of sentiment analysis remains fairly low and that the recognition of a presence of argument, if done reliably, can potentially substitute some opinion mining tasks when one intends to differentiate strong, opinionated content from neutral one. Although it is questionable whether short chat messages with a lot of implicit content provide enough information for such a task, differentiating opinionated from objective arguments could undoubtedly aid the chatbot when choosing an argumentation strategy.

8.3.3 Dialogue Strategies

As already discussed in Chapter 2, there exist many different dialogue strategies which determine the flow of the dialogue, and it would be interesting to test some of those during chats with real participants.

For example, an abstract bipolar argumentation framework is an extension of the basic argumentation framework, which includes a support relation between arguments which represents the help brought by some arguments to other arguments [Amgoud et al., 2008b]. Currently, the resulting chats follow an argument-counterargument structure. However, it would be interesting to see how bipolarity would affect the nature of the chats. Instead of countering the user's argument, the chatbot could reply with an argument that supports the argument of the user to prevent the user from using that argument in a subsequent move and then counter both of them.

At present, the chatbot and the user can only present one argument at a time. Alternatively, one could allow the chatbot (not the user to ensure the strategy is followed) to present several arguments simultaneously and extend the attack under the condition that the attack includes an argument. This strategy was introduced in [Rach et al., 2020] and is referred to as *chaining*. This allows introducing additional arguments to undermine the current move even before it was challenged. This extension does not reply to an actual attack but to an anticipated one. The authors argued that this way of arguing is more human-like and addresses the drawback of other formalisms that can be perceived as significantly less natural than human discussion. Since making the chatbot sound more natural is desirable, this is certainly an interesting strategy to explore.

8.3.4 Persuasion

In my work, I have solely focused on argumentation and have only briefly touched on different argument types. However, persuasion does not consist of argumentation only. The research conducted in [Halko and Kientz, 2010] concluded that there are eight main types of persuasive strategies in the context of health-promoting mobile applications, which can be grouped into the following four categories, where each category has two complementary approaches: (1) Instruction style which is either authoritative or non-authoritative; (2) social feedback which is either cooperative or competitive; (3) motivation type which can be extrinsic or intrinsic; and (4) reinforcement type which can be either negative or positive. Depending on the personality type, for things that cannot be achieved immediately (e.g. weight loss) or are a continuous process (e.g. long-term health), argumentation might not be the type of persuasion that yields the best results. People might require social feedback, like cooperating as a team with friends (cooperative) or competing against friends or peers to win a competition (competitive) under the supervision of a health professional. There are an extensive plethora of research topics in this domain, but it also requires the researcher to be knowledgeable about health psychology and human-computer interaction (HCI).

In the context of attitude change, depending on the user's personality and the chatbot's argumentation style, argumentation could result in the user getting defensive and even strengthening his current stance. For example, some people might be more receptive to factual causal arguments that focus on how something occurred and how a problem arose due to that occurrence. For instance, arguing why climate change is occurring allows us to explore potential causes and agree on those causes. Other are more receptive to narrative arguments, which might include anecdotal evidence as part of the narrative. For example, telling how someone's relative died due to COVID-

19 and how this most likely would not have happened had the person been vaccinated.

Given that this research aimed to design a chatbot that can be deployed on the internet so that everyone anonymously could engage with it, we do not assume any knowledge of the users' personality traits. Hence, it would be interesting to research what types of arguments the *majority* of people are receptive to or whether users can be probed with different types of arguments at the beginning of the chat and their responses analysed on the fly to establish the types of arguments the user is most receptive to.

Bibliography

[git, a] <https://github.com/lisanka93/FBchatbot>.

[git, b] https://github.com/lisanka93/meatconsumption_chatbot.

[git, c] https://github.com/lisanka93/university_fees_bot.

[git, d] https://github.com/lisanka93/covid-19_vacc_bot.

[app, a] <https://tinyurl.com/y8fjsab8>.

[app, b] https://github.com/lisanka93/Appendix_Impact-of-Arg-Type.

[cor,] https://github.com/lisanka93/Argument_Graph_Corpus,.

[you,] <https://yougov.co.uk/topics/international/articles-reports/2021/01/12/covid-19-willingness-be-vaccinated>.

[Abbott et al., 2016] Abbott, R., Ecker, B., Anand, P., and Walker, M. (2016). Internet argument corpus 2.0: An sql schema for dialogic social media and the corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4445–4452.

- [Alahmari et al., 2017] Alahmari, S., Yuan, T., and Kudenko, D. (2017). Reinforcement learning for abstract argumentation: Q-learning approach. In *Adaptive and Learning Agents workshop (at AAMAS 2017)*.
- [Amgoud et al., 2008a] Amgoud, L., Cayrol, C., Lagasquie-Schiex, M.-C., and Livet, P. (2008a). On bipolarity in argumentation frameworks. *International Journal of Intelligent Systems*, 23(10):1062–1093.
- [Amgoud et al., 2008b] Amgoud, L., Cayrol, C., Lagasquie-Schiex, M.-C., and Livet, P. (2008b). On bipolarity in argumentation frameworks. *International Journal of Intelligent Systems*, 23(10):1062–1093.
- [Amgoud et al., 2000a] Amgoud, L., Maudet, N., and Parsons, S. (2000a). Modelling dialogues using argumentation. In *Proceedings Fourth International Conference on MultiAgent Systems*, pages 31–38. IEEE.
- [Amgoud et al., 2000b] Amgoud, L., Parsons, S., and Maudet, N. (2000b). Arguments, dialogue, and negotiation. *aa*, 10(11):02.
- [Amgoud and Prade, 2009] Amgoud, L. and Prade, H. (2009). Using arguments for making and explaining decisions. *Artificial Intelligence*, 173(3-4):413–436.
- [Andrews et al., 2008] Andrews, P., Manandhar, S., and De Boni, M. (2008). Argumentative human computer dialogue for automated persuasion. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 138–147.
- [Atkinson, 2006a] Atkinson, K. (2006a). Value-based argumentation for democratic decision support. *Frontiers in Artificial Intelligence and Applications*, 144:47.

- [Atkinson, 2006b] Atkinson, K. (2006b). Value-based argumentation for democratic decision support. *Frontiers in Artificial Intelligence and Applications*, 144:47.
- [Atkinson et al., 2017] Atkinson, K., Baroni, P., Giacomin, M., Hunter, A., Prakken, H., Reed, C., Simari, G., Thimm, M., and Villata, S. (2017). Towards artificial argumentation. *AI magazine*, 38(3):25–36.
- [Atkinson et al., 2012] Atkinson, K., Bench-Capon, P., and Bench-Capon, T. J. (2012). Efficiency in persuasion dialogues. In *ICAART (2)*, pages 23–32.
- [Atkinson et al., 2020] Atkinson, K., Bench-Capon, T., Bex, F., Gordon, T. F., Prakken, H., Sartor, G., and Verheij, B. (2020). In memoriam douglas n. walton: the influence of doug walton on ai and law. *Artificial intelligence and law*, pages 1–46.
- [Atkinson et al., 2005] Atkinson, K., Bench-Capon, T., and McBurney, P. (2005). Persuasive political argument. In *Proceedings of the Fifth International Workshop on Computational Models of Natural Argument (CMNA 2005)*, pages 44–51. at IJCAI.
- [Atkinson et al., 2013] Atkinson, K., Bench-Capon, T., and Walton, D. (2013). Distinctive features of persuasion and deliberation dialogues. *Argument & Computation*, 4(2):105–127.
- [Atkinson and Wyner, 2013] Atkinson, K. and Wyner, A. (2013). The value of values: A survey of value-based computational argumentation. *From knowledge representation to argumentation in AI, law and policy making: A Festschrift in honour of Trevor Bench-Capon on the occasion of his 60th birthday*.
- [Atkinson, 2005] Atkinson, K. M. (2005). What should we do?: Computa-

tional representation of persuasive argument in practical reasoning.

- [Baroni et al., 2011] Baroni, P., Caminada, M., and Giacomin, M. (2011). An introduction to argumentation semantics. In *Knowledge Engineering Review* 26(4), pages 365–410.
- [Baroni et al., 2018] Baroni, P., Gabbay, D., Giacomin, M., and Van der Torre, L. (2018). Handbook of formal argumentation.
- [Bench-Capon, 2002a] Bench-Capon, J. (2002a). Value-based argumentation frameworks. In *In Proceedings of Non Monotonic Reasoning*, pages 444–453.
- [Bench-Capon, 2002b] Bench-Capon, T. (2002b). Agreeing to differ: Modelling persuasive dialogue between parties without a consensus about values. *Informal Logic*, 22:231–245.
- [Bench-Capon, 2003] Bench-Capon, T. J. (2003). Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448.
- [Besnard et al., 2008] Besnard, P. et al. (2008). A methodology for action-selection using value-based argumentation. *Computational Models of Argument: Proceedings of COMMA 2008*, 172:264.
- [Besnard et al., 2014] Besnard, P., García, A. J., Hunter, A., Modgil, S., Prakken, H., Simari, G., and Toni, F. (2014). Introduction to structured argumentation. *Argument and Computation*, 5(1):1–4.
- [Bickmore et al., 2005] Bickmore, T., Gruber, A., and Picard, R. (2005). Establishing the computer–patient working alliance in automated health behavior change interventions. *Patient education and counseling*, 59(1):21–30.

- [Black and Atkinson, 2011] Black, E. and Atkinson, K. (2011). Choosing persuasive arguments for action. In *AAMAS'11 The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*, pages 905–912. International Foundation for Autonomous Agents and Multiagent Systems.
- [Black et al., 2014] Black, E., Coles, A., and Bernardini, S. (2014). Automated planning of simple persuasion dialogues. In *International Workshop on Computational Logic and Multi-Agent Systems*, pages 87–104. Springer.
- [Black et al., 2017] Black, E., Coles, A. J., and Hampson, C. (2017). Planning for persuasion. In *16th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2017*, pages 933–942. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS).
- [Boltužić and Šnajder, 2014] Boltužić, F. and Šnajder, J. (2014). Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58.
- [Boltužić and Šnajder, 2015a] Boltužić, F. and Šnajder, J. (2015a). Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115, Denver, CO. Association for Computational Linguistics.
- [Boltužić and Šnajder, 2015b] Boltužić, F. and Šnajder, J. (2015b). Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115.
- [Cabrio and Villata, 2018] Cabrio, E. and Villata, S. (2018). Five years of argument mining: a data-driven analysis. In *IJCAI*, volume 18, pages 5427–5433.

- [Cawsey et al., 1999] Cawsey, A., Grasso, F., and Jones, R. (1999). A conversational model for health promotion on the world wide web. In *Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making*, pages 379–388. Springer.
- [Cayrol and Lagasquie-Schiex, 2005a] Cayrol, C. and Lagasquie-Schiex, M.-C. (2005a). On the acceptability of arguments in bipolar argumentation frameworks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 378–389. Springer.
- [Cayrol and Lagasquie-Schiex, 2005b] Cayrol, C. and Lagasquie-Schiex, M.-C. (2005b). On the acceptability of arguments in bipolar argumentation frameworks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 378–389. Springer.
- [Cerutti et al., 2014] Cerutti, F., Tintarev, N., and Oren, N. (2014). Formal arguments, preferences, and natural language interfaces to humans: an empirical evaluation. In *ECAI*, pages 207–212.
- [Chalaguine and Hunter, 2021] Chalaguine, L. and Hunter, A. (2021). Addressing popular concerns regarding covid-19 vaccination with natural language argumentation dialogues. In *European Conference on Symbolic and Quantitative Approaches with Uncertainty*, pages 59–73. Springer.
- [Chalaguine and Hunter, 2018a] Chalaguine, L. A. and Hunter, A. (2018a). Chatbot design for argument harvesting. In *Proceedings of the Workshop on Systems and Algorithms for Formal Argumentation COMMA 2018*, pages 457–458.
- [Chalaguine and Hunter, 2018b] Chalaguine, L. A. and Hunter, A. (2018b). Chatbot design for argument harvesting. *Computational Models of Argument: Proceedings of COMMA 2018*, pages 457–458.

- [Chalaguine and Hunter, 2019] Chalaguine, L. A. and Hunter, A. (2019). Knowledge acquisition and corpus for argumentation-based chatbots. In *Proceedings of the 3rd Workshop on Advances In Argumentation In Artificial Intelligence*, pages 1–14.
- [Chalaguine and Hunter, 2020] Chalaguine, L. A. and Hunter, A. (2020). A persuasive chatbot using a crowd-sourced argument graph and concerns. In *Computational Models of Argument: Proceedings of COMMA 2020*, pages 9–20. IOS Press.
- [Chalaguine et al., 2018] Chalaguine, L. A., Hunter, A., Hamilton, F. L., and Potts, H. W. W. (2018). Argument harvesting using chatbots. *Computational Models of Argument: Proceedings of COMMA 2018*, pages 149–160.
- [Chalaguine et al., 2019] Chalaguine, L. A., Hunter, A., Hamilton, F. L., and Potts, H. W. W. (2019). Impact of argument type and concerns in argumentation with a chatbot. In *Proceedings of the 31st International Conference on Tools with Artificial Intelligence*, pages 1557–1562.
- [Chalaguine and Schulz, 2017] Chalaguine, L. A. and Schulz, C. (2017). Assessing convincingness of arguments in online debates with limited number of features. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 75–83.
- [Chaudhary et al., 2021] Chaudhary, F. A., Ahmad, B., Khalid, M. D., Fazal, A., Javaid, M. M., and Butt, D. Q. (2021). Factors influencing COVID-19 vaccine hesitancy and acceptance among the Pakistani population. *Human Vaccines and Immunotherapeutics*.
- [Choi and Cardie, 2008] Choi, Y. and Cardie, C. (2008). Learning with compositional semantics as structural inference for subsentential sentiment

- analysis. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 793–801.
- [Cialdini and Cialdini, 2007] Cialdini, R. B. and Cialdini, R. B. (2007). *Influence: The psychology of persuasion*, volume 55. Collins New York.
- [Cocarascu and Toni, 2017] Cocarascu, O. and Toni, F. (2017). Identifying attack and support argumentative relations using deep learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1374–1379.
- [Cramer and Guillaume, 2019] Cramer, M. and Guillaume, M. (2019). Empirical study on human evaluation of complex argumentation frameworks. In *European Conference on Logics in Artificial Intelligence*, pages 102–115. Springer.
- [Devlin et al., 2019] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1.
- [Ding and Pan, 2016] Ding, T. and Pan, S. (2016). Personalized emphasis framing for persuasive message generation. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, pages 1432–1441.
- [Dung, 1995] Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357.
- [Fazzinga et al., 2021] Fazzinga, B., Galassi, A., and Torroni, P. (2021). An

- argumentative dialogue system for covid-19 vaccine information. In *International Conference on Logic and Argumentation*, pages 477–485. Springer.
- [Fine et al., 2011] Fine, P., Eames, K., and Heymann, D. L. (2011). “herd immunity”: a rough guide. *Clinical infectious diseases*, 52(7):911–916.
- [Fitzpatrick et al., 2017] Fitzpatrick, K. K., Darcy, A., and Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785.
- [Fogg, 2002] Fogg, B. J. (2002). Persuasive technology: using computers to change what we think and do. *Ubiquity*, 2002(December):2.
- [Galitsky et al., 2018] Galitsky, B., Ilvovsky, D., and Kuznetsov, S. O. (2018). Detecting logical argumentation in text via communicative discourse tree. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(5):637–663.
- [Georgila and Traum, 2011] Georgila, K. and Traum, D. (2011). Reinforcement learning of argumentation dialogue policies in negotiation. In *Twelfth Annual Conference of the International Speech Communication Association*.
- [Grasso, 1998] Grasso, F. (1998). Exciting avocados and dull pears combining behavioural and argumentative theory for producing effective advice. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, pages 436–441. Routledge.
- [Grasso, 2003] Grasso, F. (2003). Rhetorical coding of health promotion dialogues. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 179–188. Springer.

- [Grasso et al., 2000] Grasso, F., Cawsey, A., and Jones, R. (2000). Dialectical argumentation to solve conflicts in advice giving: a case study in the promotion of healthy nutrition. *International Journal of Human-Computer Studies*, 53(6):1077–1115.
- [Habernal and Gurevych, 2016a] Habernal, I. and Gurevych, I. (2016a). Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599.
- [Habernal and Gurevych, 2016b] Habernal, I. and Gurevych, I. (2016b). Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599.
- [Hadjinikolis et al., 2013] Hadjinikolis, C., Siantos, Y., Modgil, S., Black, E., and McBurney, P. (2013). Opponent modelling in persuasion dialogues. In *Twenty-Third International Joint Conference on Artificial Intelligence*.
- [Hadoux et al., 2015] Hadoux, E., Beynier, A., Maudet, N., Weng, P., and Hunter, A. (2015). Optimization of probabilistic argumentation with markov decision models. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.
- [Hadoux and Hunter, 2017] Hadoux, E. and Hunter, A. (2017). Strategic sequences of arguments for persuasion using decision trees. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- [Hadoux and Hunter, 2019] Hadoux, E. and Hunter, A. (2019). Comfort or safety? gathering and using the concerns of a participant for better persua-

sion. *Argument & Computation*, 10(2):113–147.

[Halko and Kientz, 2010] Halko, S. and Kientz, J. A. (2010). Personality and persuasive technology: an exploratory study on health-promoting mobile applications. In *International conference on persuasive technology*, pages 150–161. Springer.

[Hasan and Ng, 2014] Hasan, K. S. and Ng, V. (2014). Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, Doha, Qatar. Association for Computational Linguistics.

[He et al., 2022] He, L., Basar, E., Wiers, R. W., Antheunis, M. L., and Kraemer, E. (2022). Can chatbots help to motivate smoking cessation? a study on the effectiveness of motivational interviewing on engagement and therapeutic alliance. *BMC Public Health*, 22(1):1–14.

[Hitchcock et al., 2001] Hitchcock, D., McBurney, P., and Parsons, S. (2001). A framework for deliberation dialogues.

[Huang et al., 2007] Huang, J., Zhou, M., and Yang, D. (2007). Extracting chatbot knowledge from online discussion forums. In *IJCAI*, volume 7, pages 423–428.

[Huang and Lin, 2007] Huang, S.-l. and Lin, F.-r. (2007). The design and evaluation of an intelligent sales agent for online persuasion and negotiation. *Electronic Commerce Research and Applications*, 6(3):285–296.

[Hunter, 2014a] Hunter, A. (2014a). Opportunities for argument-centric persuasion in behaviour change. In *European Workshop on Logics in Artificial Intelligence*, pages 48–61. Springer.

- [Hunter, 2014b] Hunter, A. (2014b). Probabilistic strategies in dialogical argumentation. In *International Conference on Scalable Uncertainty Management*, pages 190–202. Springer.
- [Hunter, 2016] Hunter, A. (2016). Computational persuasion with applications in behaviour change. In *Computational Models of Argument: Proceedings of COMMA 2016*, pages 5–18.
- [Hunter, 2018] Hunter, A. (2018). Towards a framework for computational persuasion with applications in behaviour change. *Argument & Computation*, 9(1):15–40.
- [Hunter and Polberg, 2017] Hunter, A. and Polberg, S. (2017). Empirical methods for modelling persuadees in dialogical argumentation. In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 382–389. IEEE.
- [Hunter et al., 2020] Hunter, A., Polberg, S., and Hadoux, E. (2020). Strategic argumentation dialogues for persuasion: Framework and experiments based on modelling the beliefs and concerns of the persuadee. Technical report, University College London.
- [Hunter and Thimm, 2016] Hunter, A. and Thimm, M. (2016). Optimization of dialectical outcomes in dialogical argumentation. *International Journal of Approximate Reasoning*, 78:73–102.
- [Inkster et al., 2018] Inkster, B., Sarda, S., Subramanian, V., et al. (2018). An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11):e12106.
- [Josekutty Thomas et al., 2017] Josekutty Thomas, R., Masthoff, J., and Oren,

- N. (2017). Adapting healthy eating messages to personality. In *International Conference on Persuasive Technology*, pages 119–132. Springer.
- [Jungmann et al., 2019] Jungmann, S. M., Klan, T., Kuhn, S., and Jungmann, F. (2019). Accuracy of a chatbot (ada) in the diagnosis of mental disorders: comparative case study with lay and expert users. *JMIR formative research*, 3(4):e13863.
- [Kaci and van der Torre, 2008] Kaci, S. and van der Torre, L. (2008). Preference-based argumentation: Arguments supporting multiple values. *International Journal of Approximate Reasoning*, 48(3):730–751.
- [Katsumi et al., 2018] Katsumi, H., Hiraoka, T., Yoshino, K., Yamamoto, K., Motoura, S., Sadamasa, K., and Nakamura, S. (2018). Optimization of information-seeking dialogue strategy for argumentation-based dialogue system. *arXiv preprint arXiv:1811.10728*.
- [Krauthoff et al., 2016] Krauthoff, T., Baurmann, M., Betz, G., and Mauve, M. (2016). Dialog-based online argumentation. In *Computational Models of Argument: Proceedings of COMMA 2016*, pages 33–40.
- [Krauthoff et al., 2017] Krauthoff, T., Meter, C., and Mauve, M. (2017). Dialog-based online argumentation: Findings from a field experiment. In *AI³@ AI* IA*, pages 85–99.
- [Lawrence and Reed, 2020] Lawrence, J. and Reed, C. (2020). Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.
- [Le et al., 2018] Le, D.-T., Nguyen, C.-T., and Nguyen, K. A. (2018). Dave the debater: a retrieval-based and generative argumentative dialogue agent. In *Proceedings of the 5th Workshop on Argument Mining*, pages 121–130.

- [Levy et al., 2018] Levy, R., Bogin, B., Gretz, S., Aharonov, R., and Slonim, N. (2018). Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2066–2081.
- [Lukin et al., 2017] Lukin, S. M., Anand, P., Walker, M., and Whittaker, S. (2017). Argument strength is in the eye of the beholder: Audience effect in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, pages 742–753.
- [MacDonald et al., 2015] MacDonald, N. E. et al. (2015). Vaccine hesitancy: Definition, scope and determinants. *Vaccine*, 33(34):4161–4164.
- [Marco-Franco et al., 2020] Marco-Franco, J. E., Guadalajara-Olmeda, N., González-de Julián, S., and Vivas-Consuelo, D. (2020). Covid-19 healthcare planning: Predicting mortality and the role of the herd immunity barrier in the general population. *Sustainability*, 12(13):5228.
- [Mazzotta et al., 2007] Mazzotta, I., De Rosis, F., and Carofiglio, V. (2007). Portia: A user-adapted persuasion system in the healthy-eating domain. *IEEE Intelligent systems*, 22(6):42–51.
- [McBurney et al., 2003] McBurney, P., Van Eijk, R. M., Parsons, S., and Amgoud, L. (2003). A dialogue game protocol for agent purchase negotiations. *Autonomous Agents and Multi-Agent Systems*, 7(3):235–273.
- [Meppelink et al., 2019] Meppelink, C. S., Smit, E. G., Fransen, M. L., and Diviani, N. (2019). “i was right about vaccination”: Confirmation bias and health literacy in online health information seeking. *Journal of health communication*, 24(2):129–140.

- [Miller, 1995] Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.
- [Misra et al., 2016] Misra, A., Ecker, B., and Walker, M. A. (2016). Measuring the similarity of sentential arguments in dialog. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 276–287.
- [Modgil, 2006] Modgil, S. (2006). Value based argumentation in hierarchical argumentation frameworks. *Computational Models of Argument: Proceedings of COMMA 2006*, 144:297–308.
- [Monteserin and Amandi, 2013] Monteserin, A. and Amandi, A. (2013). A reinforcement learning approach to improve the argument selection effectiveness in argumentation-based negotiation. *Expert Systems with Applications*, 40(6):2182–2188.
- [Murphy et al., 2016] Murphy, J., Black, E., and Luck, M. (2016). A heuristic strategy for persuasion dialogues. In *Computational Models of Argument: Proceedings of COMMA 2016*, pages 411–418. IOS Press.
- [Murphy et al., 2018] Murphy, J., Burdusel, A., Luck, M., Zschaler, S., and Black, E. (2018). Deriving persuasion strategies using search-based model engineering. In *Computational Models of Argument: Proceedings of COMMA 2018*, pages 221–232.
- [Neumann-Böhme et al., 2020] Neumann-Böhme, S., Varghese, N. E., Sabat, I., Barros, P. P., Brouwer, W., van Exel, J., Schreyögg, J., and Stargardt, T. (2020). Once we have it, will we use it? a european survey on willingness to be vaccinated against covid-19.
- [Ogden, 2012] Ogden, J. (2012). *Health Psychology: A Textbook (5th ed)*,

page 91. Maidenhead: Open University Press.

- [Oniani and Wang, 2020] Oniani, D. and Wang, Y. (2020). A qualitative evaluation of language models on automatic question-answering for covid-19. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–9.
- [Ontanón and Plaza, 2006] Ontanón, S. and Plaza, E. (2006). Arguments and counterexamples in case-based joint deliberation. In *International Workshop on Argumentation in Multi-Agent Systems*, pages 36–53. Springer.
- [O’keefe and Jensen, 2006] O’keefe, D. J. and Jensen, J. D. (2006). The advantages of compliance or the disadvantages of noncompliance? a meta-analytic review of the relative persuasive effectiveness of gain-framed and loss-framed messages. *Annals of the International Communication Association*, 30(1):1–43.
- [Pazienza et al., 2017] Pazienza, A., Ferilli, S., and Esposito, F. (2017). Constructing and evaluating bipolar weighted argumentation frameworks for online debating systems. In *AI³@ AI* IA*, pages 111–125.
- [Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, page 1532–1543. ACL.
- [Plous, 1993] Plous, S. (1993). *The psychology of judgment and decision making*. Mcgraw-Hill Book Company.
- [Polberg and Hunter, 2018] Polberg, S. and Hunter, A. (2018). Empirical evaluation of abstract argumentation: Supporting the need for bipolar and probabilistic approaches. *International Journal of Approximate Reasoning*, 93:487–543.

- [Prakken, 2006] Prakken, H. (2006). Formal systems for persuasion dialogue. *The knowledge engineering review*, 21(2):163–188.
- [Prasetyo et al., 2020] Prasetyo, P. K., Achananuparp, P., and Lim, E.-P. (2020). Foodbot: A goal-oriented just-in-time healthy eating interventions chatbot. In *Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pages 436–439.
- [Prochaska et al., 2021] Prochaska, J. J., Vogel, E. A., Chieng, A., Kendra, M., Baiocchi, M., Pajarito, S., and Robinson, A. (2021). A therapeutic relational agent for reducing problematic substance use (woebot): development and usability study. *Journal of medical Internet research*, 23(3):e24850.
- [Rach et al., 2018] Rach, N., Minker, W., and Ultes, S. (2018). Markov games for persuasive dialogue. In *Computational Models of Argument: Proceedings of COMMA 2018*, pages 213–220.
- [Rach et al., 2020] Rach, N., Minker, W., and Ultes, S. (2020). Increasing the naturalness of an argumentative dialogue system through argument chains. In *Computational Models of Argument: Proceedings of COMMA 2020*, pages 331–338. IOS Press.
- [Rach et al., 2021] Rach, N., Schindler, C., Feustel, I., Daxenberger, J., Minker, W., and Ultes, S. (2021). From argument search to argumentative dialogue: A topic-independent approach to argument acquisition for dialogue systems. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 368–379.
- [Rahwan et al., 2010] Rahwan, I., Madakkatel, M. I., Bonnefon, J.-F., Awan, R. N., and Abdallah, S. (2010). Behavioral experiments for assessing the abstract argumentation semantics of reinstatement. *Cognitive Science*, 34(8):1483–1502.

- [Rajpurkar et al., 2016] Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuad: 100,000+ questions for machine comprehension of text. In *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*.
- [Rakshit et al., 2019] Rakshit, G., Bowden, K. K., Reed, L., Misra, A., and Walker, M. (2019). Debbie, the debate bot of the future. In *Advanced Social Interaction with Agents*, pages 45–52. Springer.
- [Rémy et al., 2015] Rémy, V., Langeron, N., Quilici, S., and Carroll, S. (2015). The economic value of vaccination: why prevention is wealth. *Journal of market access & health policy*, 3(1):29284.
- [Rienstra et al., 2013] Rienstra, T., Thimm, M., and Oren, N. (2013). Opponent models with uncertainty for strategic argumentation. In *Twenty-Third International Joint Conference on Artificial Intelligence*.
- [Riley et al., 2011] Riley, L., Atkinson, K., Payne, T., and Black, E. (2011). An implemented dialogue system for inquiry and persuasion. In *International Workshop on Theorie and Applications of Formal Argumentation*, pages 67–84. Springer.
- [Rosenfeld and Kraus, 2016a] Rosenfeld, A. and Kraus, S. (2016a). Providing arguments in discussions on the basis of the prediction of human argumentative behavior. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 6(4):1–33.
- [Rosenfeld and Kraus, 2016b] Rosenfeld, A. and Kraus, S. (2016b). Strategic argumentative agent for human persuasion. In *ECAI 2016*, pages 320–328. IOS Press.
- [Sadri et al., 2001] Sadri, F., Toni, F., and Torroni, P. (2001). Logic agents,

- dialogues and negotiation: an abductive approach. In *Proceedings of the AISB 2001 Symposium on Information Agents for E-Commerce*, volume 39. UK.
- [Santos et al., 2009] Santos, R., Marreiros, G., Ramos, C., Neves, J., and Bulas-Cruz, J. (2009). Using personality types to support argumentation. In *International Workshop on Argumentation in Multi-Agent Systems*, pages 292–304. Springer.
- [Sherman et al., 2021] Sherman, S. M., Smith, L. E., Sim, J., Amlôt, R., Cutts, M., Dasch, H., Rubin, G. J., and Sevdalis, N. (2021). Covid-19 vaccination intention in the uk: results from the covid-19 vaccination acceptability study (covaccs), a nationally representative cross-sectional survey. *Human vaccines & immunotherapeutics*, 17(6):1612–1621.
- [Singhal, 2001] Singhal, A. (2001). Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24:35–43.
- [Stab et al., 2018] Stab, C., Daxenberger, J., Stahlhut, C., Miller, T., Schiller, B., Tauchmann, C., Eger, S., and Gurevych, I. (2018). Argumenttext: Searching for arguments in heterogeneous sources. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: demonstrations*, pages 21–25.
- [Swanson et al., 2015] Swanson, R., Ecker, B., and Walker, M. (2015). Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue*, pages 217–226.
- [Tan et al., 2016] Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., and Lee, L. (2016). Winning arguments: Interaction dynamics and persuasion

strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624.

[Toniuc and Groza, 2017] Toniuc, D. and Groza, A. (2017). Climebot: An argumentative agent for climate change. In *13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 63–70. IEEE.

[Tversky and Kahneman, 1985] Tversky, A. and Kahneman, D. (1985). The framing of decisions and the psychology of choice. In *Behavioral decision making*, pages 25–41. Springer.

[Vargheese et al., 2016] Vargheese, J. P., Sripada, S., Masthoff, J., and Oren, N. (2016). Persuasive strategies for encouraging social interaction for older adults. *International Journal of Human-Computer Interaction*, 32(3):190–214.

[Vargheese et al., 2013] Vargheese, J. P., Sripada, S., Masthoff, J., Oren, N., Schofield, P., and Hanson, V. L. (2013). Persuasive dialogue for older adults: Promoting and encouraging social interaction. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 877–882.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, , and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 2017-December.

[Villata et al., 2017] Villata, S., Cabrio, E., Jraidi, I., Benlamine, S., Chaouachi, M., Frasson, C., and Gandon, F. (2017). Emotions and personality traits in argumentation: An empirical evaluation. *Argument & Computation*, 8(1):61–87.

- [Wachsmuth et al., 2017a] Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T. A., Hirst, G., and Stein, B. (2017a). Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187.
- [Wachsmuth et al., 2017b] Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T. A., Hirst, G., and Stein, B. (2017b). Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187.
- [Wachsmuth et al., 2018] Wachsmuth, H., Syed, S., and Stein, B. (2018). Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251.
- [Walker et al., 2012] Walker, M. A., Anand, P., Tree, J. F., Abbot, R., and King, J. (2012). A corpus for research on deliberation and debate. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 812–817.
- [Walton and Krabbe, 1995] Walton, D. and Krabbe, E. C. (1995). *Commitment in dialogue: Basic concepts of interpersonal reasoning*. SUNY press.
- [Walton et al., 2008] Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation schemes*. Cambridge University Press.
- [Ward et al., 2020] Ward, J. K., Alleaume, C., Peretti-Watel, P., Seror, V., Cortaredona, S., Launay, O., Raude, J., Verger, P., Beck, F., Legleye, S., et al. (2020). The french public’s attitudes to a future covid-19 vaccine:

The politicization of a public health issue. *Social science & medicine*, 265:113414.

[Wei et al., 2020] Wei, J., Huang, C., Vosoughi, S., and Wei, J. (2020). What are people asking about COVID-19? a question classification dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

[Weizenbaum, 1966] Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

[Wells, 2014] Wells, S. (2014). Argument mining: Was ist das? In *Proceedings of the 14th International Workshop on Computational Models of Natural Argument*.

[Xu et al., 2020] Xu, H., Avelka, J., and Ashley, K. D. (2020). Using argument mining for legal text summarization. In *Proceedings of the 33rd International Conference on Legal Knowledge and Information Systems (JURIX 2020)*.

[Zaal et al., 2017] Zaal, E., Mills, G., Hagen, A., Huisman, C., and Hoeks, J. (2017). Convincing conversations: Using a computer-based dialogue system to promote a plant-based diet. In *CogSci*.