Fostering Critical Reflection

Unraveling Reflection Bias in Argumentation through Awareness and Mitigation

Dissertation

Zur Erlangung des Doktorgrades an der
Fakultät für Angewandte Informatik
der Universität Augsburg

vorgelegt von

Klaus Weber

Juli 2024

Universität
Augsburg
University

Erstgutachterin:	Prof. Dr. Elisabeth André
Zweitgutachter:	Prof. Dr. Jörg Hähner
Tag der mündlichen Prüfung:	24. September 2024

In celebration of intellectual curiosity and the pursuit of truth - may this thesis inspire continued critical reflection.

Acknowledgements

Undertaking a dissertation is a profound journey that does not transpire in isolation. It necessitates the support, guidance, and collaboration of numerous individuals, without whom this endeavor would remain insurmountable.

Foremost, I extend my deepest appreciation to *Prof. Dr. Elisabeth André* for her pivotal role as my supervisor. Her unwavering support and invaluable insights have been instrumental in shaping not only the existence of this thesis but also its quality. Her constructive feedback has been a guiding light throughout the dissertation project of the last seven years.

I am also immensely grateful to my esteemed colleagues at Ulm University, *Dr. Niklas Rach* and *Annalena Aicher*. Our collaboration within the BEA and EVA projects has significantly enriched the scope and depth of my research.

Working alongside dedicated colleagues is both challenging and rewarding. I extend my heartfelt thanks to my student assistants, *Eva Pohlen*, *Lukas Tinnes*, *Marc-Leon Reinecker*, and *Natalie Hogh*. Their dedicated support and involvement were instrumental in advancing the progression and depth of my research work and added a layer of enjoyment to the process.

In addition, I wish to express gratitude to *Dr. Katharina Weitz*. Her dual expertise in Psychology and unwavering support, especially during the inception of my thesis, provided invaluable insights that were instrumental in shaping my early direction.

My deepest appreciation goes to my closest friends, *Björn Petrak*, *Lisa Petrak*, and *Laura Grabowski*. Their unwavering presence and support have been an anchor in the tumultuous seas of this dissertation journey.

Furthermore, I would like to express special thanks to *Grammarly* and *ChatGPT* for providing invaluable lectorate support and refining the writing style to its best form. In addition, I would like to thank *ChatGPT* for its beautifully generated lion images for the chapter headers.

To all those mentioned and countless others who may not be explicitly named here, your support, encouragement, and belief in my endeavors have been the cornerstone of this academic pursuit. Thank you.

"We do not learn from experience... we learn from reflecting on experience." (John Dewey)

Abstract

With the steadily growing abundance of online information, whether through news portals on the internet or social networks, two increasingly pressing problems have arisen: On the one hand, online users in social networks are frequently confronted with distorted and one-sided information due to filter algorithms, while on the other hand, there is a diminishing willingness for open discourse.

These issues have been particularly evident during the recent COVID-19 pandemic, where divergent viewpoints were quickly rejected. One reason is that it is easier to engage with arguments from one's own side, while arguments from the opposing side (referred to challenger arguments) are often blocked or perceived as provocative.

This phenomenon can be explained psychologically through affective reactions and peripheral information processing, where information contradicting one's own opinion triggers strong emotional reactions, thus complicating rational understanding and leading to a reflection bias. This means that certain arguments are either misinterpreted or ignored (content-based reflection bias), or people are influenced by subliminal cues, such as emotions (behavior-based reflection bias), of which many people are not aware, though.

This dissertation focuses on the reflection bias from two perspectives: 1) Raising awareness of the behavior-based reflection bias through the use of explainable Artificial Intelligence, and 2) mitigation of the content-based reflection bias using an argumentative dialog system.

In the first part of the thesis, we examine how, with the help of explainable Artificial Intelligence and Neural Networks, we can make the behavior-based reflection bias, specifically regarding gestures, visible. The goal is to draw attention and awareness to the influence of gestures on the perceived persuasive effect through visual explanations.

We investigate whether we can generate satisfactory explanations when training Neural Networks with subjective data that significantly differ in quality and accuracy from gold standard data due to noise. Furthermore, we explore whether these explanations are suitable for highlighting the behavior-based reflection bias and differences between individuals by examining whether the focus of the networks aligns with insights from the literature.

Our analysis shows that Neural Networks primarily focus on hand gestures, which is identified in the literature as an important indicator of persuasion.

In the second part of the thesis, we explore how to mitigate the content-based reflection bias. To achieve this, we develop an argumentative dialog system that encourages users through targeted interventions to move away from a one-sided argument exploration and to consider arguments from the opposing side. The system utilizes a metric, gauging the extent to which users predominantly focus on arguments that align with their viewpoint. In three studies, we examine the effects of interventions on reflection and exploration behavior.

The results of the studies demonstrate that users significantly engage more with challenger arguments and spend considerably more time considering these arguments when the system applies intervention strategies. Additionally, we present some interaction effects with personality traits, supporting the idea that systems aiming to improve reflection should also take into account the user's personality.

Keywords: Reflective Engagement, Reflective Argumentation, Awareness and Mitigation of Reflection Bias, (Explainable) Artificial Intelligence, Conversational Intelligent Agent, Conversational Dialogue System

Zusammenfassung

Durch die stetig wachsende Fülle an Online-Informationen, sei es über Nachrichtenportale im Internet oder soziale Netzwerke, treten zwei immer drängendere Probleme auf: Zum einen werden Online-Nutzer in sozialen Netzwerken vermehrt durch Filteralgorithmen mit verzerrten und einseitigen Informationen konfrontiert, während zum anderen die Bereitschaft zum offenen Diskurs abnimmt.

Diese Probleme wurden insbesondere während der jüngsten Corona-Pandemie deutlich, als abweichende Standpunkte schnell abgelehnt wurden. Ein Grund dafür ist, dass es einfacher ist, sich mit Argumenten der eigenen Seite zu befassen, während Gegenargumente (sogenannte Herausforderungsargumente) oft abgeblockt oder als provokativ empfunden werden.

Dieses Phänomen kann psychologisch durch affektive Reaktionen und periphere Informationsverarbeitung erklärt werden, bei denen Informationen, die der eigenen Meinung widersprechen, starke emotionale Reaktionen hervorrufen und somit das rationale Verständnis erschweren und zu einem Reflexions-Bias führen können. Das bedeutet, dass bestimmte Argumente entweder falsch interpretiert oder ignoriert werden (inhaltsbasierter Reflexions-Bias), oder man subtilen, nichtverbalen Signalen wie Emotionen erliegt (verhaltensbasierter Reflexions-Bias). Viele Menschen sind sich dessen jedoch nicht bewusst.

Diese Dissertation fokussiert sich daher auf den Reflexions-Bias aus zwei Perspektiven: 1) Bewusstseinssteigerung des verhaltensbasierten Reflexions-Bias durch Einsatz erklärbarer künstlicher Intelligenz und 2) Mitigation des inhaltsbasierten Reflexions-Bias mithilfe eines argumentativen Dialogsystems.

Im ersten Teil der Dissertation untersuchen wir konkret, wie wir durch den Einsatz von erklärender künstlicher Intelligenz und neuronalen Netzen den verhaltensbasierten Reflexions-Bias, insbesondere im Zusammenhang mit Gesten, sichtbar machen können. Das Ziel ist es, durch anschauliche Bild-Erklärungen das Bewusstsein für den Einfluss von Gesten auf die wahrgenommene Überzeugungswirkung zu schärfen.

Dabei analysieren wir einerseits, ob wir zufriedenstellende Erklärungen generieren können, wenn wir neuronale Netze mit subjektiven Daten trainieren, die sich durch Rauschen in Qualität und Genauigkeit deutlich von Goldstandard-Daten unterscheiden. Andererseits untersuchen wir, ob diese Erklärungen geeignet sind, um auf den verhaltensbasierten Reflexions-Bias und die Unterschiede zwischen Personen aufmerksam zu machen, indem wir überprüfen, ob der Fokus der Netze mit Erkenntnissen aus der Literatur übereinstimmt.

Unsere Analyse zeigt, dass sich die neuronalen Netze vor allem auf Handgesten fokussieren, was in der Literatur als wichtiger Indikator für Überzeugung benannt wird.

Im zweiten Teil der Dissertation untersuchen wir, wie wir den inhaltsbasierten Reflexions-Bias mitigieren können. Hierzu entwickeln wir ein argumentatives Dialogsystem, das durch gezielte Interventionen den Nutzer dazu bringt, von einer einseitigen Argumentexploration abzurücken und vor allem Argumente der Gegenseite zu betrachten. Dabei verwendet das System eine Metrik, die das Ausmaß misst, in dem sich Benutzer überwiegend auf Argumente konzentrieren, die mit ihrer Sichtweise übereinstimmen. In drei Studien untersuchen wir die Effekte der Interventionen auf Reflexion und Explorationsverhalten.

Die Ergebnisse der Studien zeigen, dass sich Nutzer signifikant mehr mit Argumenten der Gegenseite befassen und sich mit Argumenten deutlich länger auseinandersetzen, wenn das System Interventionsstrategien anwendet. Darüber hinaus zeigen wir einige Interaktionseffekte mit Persönlichkeitsmerkmalen auf, die belegen, dass Systeme, die Reflexion verbessern wollen, auch die Persönlichkeit des Nutzers berücksichtigen sollten.

Stichwörter: Reflektive Auseinandersetzung, Reflektive Argumentation, Bewusstsein und Minderung des Reflexions-Bias', (Erklärbare) Künstliche Intelligenz, Intelligenter Gesprächsagent, Konversationelles Dialogsystem

Contents

Appendices

Annexes

List of Algorithms

Acronyms

ACC	Acceptability (ITU-T questionnaire) *(pp. 152, 153, 155, 249, 251)*
ADAM	Adaptive Moment Estimation (Optimization algorithm for machine learning models) *(pp. 24, 25, 82, 83)*
ADAMAX	Adaptive Moment Estimation with Infinity Norm (Optimization algorithm for machine learning models) *(pp. 24, 25, 72)*
ADF	Average Duration of Fixations (Eye gaze dependent measure) *(pp. xvii, 170, 172–180, 188, 245, 246)*
AE	Aesthetic and Appeal Scale (User engagement questionnaire, O'Brien et al. (2018)) *(pp. 154, 155, 254)*
AI	Artificial Intelligence *(pp. 29, 30, 184)*
ANCOVA	Analysis of Covariance (A statistical test) *(p. 170)*
ANF	Average Number of Fixations (Eye gaze dependent measure) *(pp. xvii, 170, 172–180, 188, 189, 245, 246)*
ANOVA	Analysis of Variances (A statistical test) *(pp. 149, 150)*
AOI	Area of Interest *(pp. 169, 170, 172–177, 245, 246)*
ARG	Argumentation (ITU-T questionnaire) *(pp. 152, 153, 155, 180, 189, 249, 251)*
ATI	Affinity for technology (Independent measure) *(pp. 168, 169, 247, 248)*
AVQ	Argument Visitation Quotient (Subjective dependent measure). A metric to measure challenging argument exploration *(pp. xv–xvii, xxiv, 10, 12, 13, 66, 107, 108, 110–112, 126, 130, 131, 133, 136, 137, 139–141, 143, 145–150, 157–160, 164–167, 171–173, 179, 180, 186, 187, 190, 237–244)*
BEA	Building Engaging Argumentation (Name of our developed system) *(pp. 95, 97, 156, 166, 169)*
BERT	Bidirectional Encoder Representations from Transformers (Type of NN architecture, specifically used for natural language processing) *(p. 99)*
BWAG	Bipolar Weighted Argument Graph *(p. 112)*

CAM	Class Activation Mapping (XAI visualization technique) *(p. 32)*
CE	Conversational Engagement (Subjective dependent measure) *(pp. xv, xvii, 10–12, 145–148, 151, 152, 154, 155, 179, 180, 187, 189, 237, 249, 251, 253, 255)*
CML	Cooperative Machine Learning (Machine learning technique) *(p. 70)*
CNN	Convolutional Neural Network (Type of NN architecture, specifically used for image processing and grid-like data) *(pp. xiii, 22, 23, 31, 32, 71, 92)*
COM	Communication with the system (ITU-T questionnaire) *(pp. 152, 153, 155, 249)*
CS	Conscientiousness (Independent measure) *(pp. 13, 167, 169, 247, 248)*
DI	Dialogue (ITU-T questionnaire) *(pp. 152, 153, 155, 156, 180, 189, 249, 250)*
DS	Deprivation Sensitivity (Independent measure) *(pp. 168, 169, 247, 248)*
EASI	Emotions as Social Influence (Psychological theory model of emotional influences) *(pp. xiv, 53, 54, 58)*
ELM	Elaboration Likelihood Model (Psychological model of information processing) *(pp. xiv, 52, 54, 56)*
F	Familiarity Scale (User trust questionnaire, Körber (2019)) *(pp. 154, 157, 161–164, 252)*
FA	Focused Attention Scale (User engagement questionnaire, O'Brien et al. (2018)) *(pp. 154, 155, 254)*
GLM	General Linear Model (A statistical test) *(pp. 170, 173, 175, 176)*
Grad-CAM	Gradient-weighted Class Activation Mapping (XAI visualization technique) *(pp. xiv, xxvi, 16, 31, 32, 70, 74, 77, 78, 93)*
HSM	Heuristic Systematic Model (Psychological model of information processing) *(pp. xiv, 53)*
ID	Intention of Developers Scale (User trust questionnaire, Körber (2019)) *(p. 162)*
IPS	Information Provided by the System (ITU-T questionnaire) *(pp. 152, 153, 155, 249)*

LLM	Large Language Model *(p. 64)*
LOC	Locus of Control (Independent measure) *(pp. 165, 168, 169, 173, 175, 176, 180, 189, 247, 248)*
LRP	Layer-wise Relevance Propagation (XAI visualization technique) *(pp. xiii, xiv, xxvi, 16, 31, 33, 35, 70, 74, 77, 78, 86, 87, 93, 94)*
LSTM	Long Short-Term Memory (Type of NN architecture, specifically a form of recurrent neural network) *(p. 99)*
MDP	Markov Decision Process (Class of machine learning algorithm) *(pp. 41, 43, 49, 50)*
MSE	Mean Squared Error (A loss function) *(pp. 18, 23–25)*
NFC	Need For Cognition (Independent measure) *(pp. 6, 13, 52, 55, 165, 167, 169, 173–175, 180, 189, 190, 247, 248)*
NLG	Natural Language Generation (Dialog system component) *(pp. xiv, xv, 96, 97, 100, 116, 118, 119)*
NLU	Natural Language Understanding (Dialog system component) *(pp. xiv, 96, 97, 99, 105)*
NN	Neural Network (Supervised machine learning algorithm) *(pp. xiii, xxi–xxiii, 8, 16, 22, 23, 30, 31, 33, 40, 44, 68–72, 81, 185)*
OP	Openness (Independent measure) *(pp. 167, 169, 247, 248)*
PS	Perceptual Speed (Independent measure) *(pp. 13, 165, 168, 169, 173, 176–178, 180, 189, 247, 248)*
PT	Propensity to Trust Scale (User trust questionnaire, Körber (2019)) *(pp. 153, 154, 157, 162, 180, 252)*
PU	Perceived Usability Scale (User engagement questionnaire, O'Brien et al. (2018)) *(pp. 154, 155, 254)*
QLT	Overall Quality (ITU-T questionnaire) *(pp. 152, 153, 180, 189, 249, 251)*
RC	Reliability and Competence Scale (User trust questionnaire, Körber (2019)) *(pp. 161, 162)*

RE Reflective Engagement (Psychological concept, Subjective dependent measure); measured by using a computational metric (AVQ) *(pp. xiv, xv, 7, 8, 10–12, 55–57, 59, 61, 63, 65, 66, 107, 108, 110–112, 114, 116, 118, 120, 122, 124, 126, 186, 190)*

ReLU Rectifier Linear Unit (Activation Function) *(pp. 18, 33, 72)*

RJM Reflective Judgement Model (Psychological model) *(p. 55)*

RL Reinforcement Learning (Unsupervised machine learning algorithm) *(pp. xiii, xv, xxv–xxvii, 16, 40, 41, 47, 64, 65, 129, 134–141, 143)*

RMSPROP Root Mean Square Propagation (Optimization algorithm for machine learning models) *(pp. 24, 25)*

RP Reading Proficiency (Independent measure) *(pp. 168, 169, 247, 248)*

RW Reward Factor Scale (User engagement questionnaire, O'Brien et al. (2018)) *(pp. 154, 155, 189, 254, 255)*

SB System Behavior (ITU-T questionnaire) *(pp. 152, 153, 155, 156, 180, 249, 250)*

SC Social Curiosity (Independent measure) *(pp. 168, 169, 247, 248)*

SGD Stochastic Gradient Descent (Optimization algorithm for machine learning models) *(pp. 24, 82, 83)*

SMOTE Synthetic Minority Oversampling Technique (A technique to reduce imbalanced class distributions) *(p. 84)*

TA Trust in Automation Scale (User trust questionnaire, Körber (2019)) *(pp. 153, 154, 157, 162, 180, 189, 252, 253)*

TTS Text-To-Speech System (Dialog system component) *(p. 155)*

UC User Characteristic *(pp. xvi, xvii, 12, 13, 145, 146, 165–173, 175, 177–180, 189, 190, 247)*

UIS User's Impression of the System (ITU-T questionnaire) *(pp. 152, 153, 155, 249, 251)*

UP Understanding and Predictability Scale (User trust questionnaire, Körber (2019)) *(pp. 154, 161, 162, 252)*

VWM Visual Working Memory (Independent measure) *(pp. 168, 169, 173, 176, 177, 180, 189, 247, 248)*

XAI Explainable Artificial Intelligence *(pp. xxii, xxiii, 7, 9, 10, 16, 29–32, 60, 67–70, 74, 78–80, 86, 91–93, 184, 185)*

Symbols

$\rightarrow$	a relation between components φ_i and φ_j *(pp. xxvi, xxvii, 37–39, 97–99, 114, 118)*
$\_$	a placeholder for the argument relation $attack \veebar support$ *(pp. xxv, 38, 98, 101, 108, 109, 114, 115)*
$*$	a placeholder for any index variable *(pp. xxv, 108, 109)*
$\boldsymbol{\omega}$	a vector encoding learned parameters of an RL action $a_i \in \mathcal{A}$ as vector *(pp. xxvii, 44, 45, 48–51, 138, 141)*
α_c	Chronbach's alpha *(pp. 28, 29, 86)*
act	an activation function of a neural network *(pp. 17, 18, 20, 21)*
a	an RL action *(pp. xxv, xxvii, 41–45, 48–51, 136–141)*
$\mathcal{A}$	a set of RL actions $\{a_1,...,a_n\}$ *(pp. xxv, xxvii, 41–45, 48, 135–137, 139–141)*
α	the learning rate of an optimizer algorithm *(pp. 24, 25, 43, 45, 49–51, 141)*
$\mathcal{A}$	a black box analyzer function *(pp. 69, 70)*
$Args$	a set of arguments $\{\Phi_1, ..., \Phi_n\}$ under graph G with Φ_0 as root. A set of arguments under a sub-graph of G with root $\Phi_i \in Args$ is denoted as $Args(\Phi_i) \subseteq Args$. A set of arguments with the same parent argument $\Phi_i \in Args$ is denoted as $Args_{\Phi_{*\Rightarrow\_i}} \subseteq Args(\Phi_i)$. *(pp. xxv, xxvii, 38, 39, 99, 108–110, 114, 115, 121, 130, 131, 138, 148, 159, 160, 240–242)*
b	an entry of a bias vector $\boldsymbol{b}_i$ of a neural network *(pp. xxv, 17)*
$\boldsymbol{b}$	a bias $(b_i, ..., b_n)^T$ of a neural network *(pp. xxv, 17, 18)*
β	a hyper-parameter of an optimizer algorithm *(pp. 24, 25)*
$\mathcal{B}$	a black box function of a classificator *(pp. 16, 18, 23, 69, 70, 85)*
C	a set of coefficient vectors $\{\boldsymbol{c}_1, ..., \boldsymbol{c}_n\}$ of Fourier basis transformation *(pp. 47, 48)*
$\boldsymbol{c}$	a coefficient vector of a Fourier basis transformation *(pp. xxv, 46–48)*

d the Cohen's d effect size of a statistical t-test *(pp. 150, 160, 161, 163, 171, 173, 188)*

dim the dimension of a vector or matrix *(pp. 16, 17, 47, 50, 51, 141)*

e the Euler's number 2.71828... *(p. 123)*

$\boldsymbol{e}$ a one hot vector with exactly one entry set to 1 *(p. 19)*

ϵ exploration probability for an RL algorithm *(pp. 42, 43, 45)*

E the error function of an optimizer, e.g. mean-squared-error *(pp. 18–20, 23–25, 44)*

η_p^2 the partial eta squared effect size of a statistical test *(pp. 149, 150, 160, 171, 173, 175, 176)*

exp a specific explainer, like Grad-CAM or LRP *(pp. 69, 70)*

F the absolute user focus based on visited arguments *(pp. 110, 111, 136, 141)*

f the user feedback function for a given argument Φ_i *(pp. 115, 121, 122, 235, 236)*

$focus$ the user focus for a given argument Φ_i *(pp. 109, 110)*

γ a discount factor used in in RL algorithms *(pp. 43–45, 49–51)*

G an argument graph consisting of a set of components L_t as nodes and a relation $\rightarrow$ between them defining edges *(pp. xxv, 37–39, 118)*

$\mathcal{I}$ the set of annotators *(pp. 69, 70, 85, 87)*

L_c a communication language consisting of speech acts *(pp. 36, 39, 98)*

L_t a set of argument components $\{\varphi_1, ..., \varphi_n\}$ (premise or evidence, supporting or attacking another premise) *(pp. xxvi, 37–39, 97, 100, 114, 118)*

$\mathcal{M}$ a set of sequential, ordered moves $(m_1, ..., m_n)$ within a dialogue *(p. xxvi)*

m a single move $m_i \in \mathcal{M}$ within a dialogue *(pp. xxvi, 131, 138)*

μ the mean score of a set of values *(pp. 147, 151–155, 162, 233, 234, 249–255)*

$\mathbb{N}$ the set of natural numbers *(p. 110)*

$\mathcal{P}$ the probability function *(pp. 41, 141)*

p the alpha error (type I error) of a statistical test *(pp. 123, 149–155, 160–163, 171–178, 187–189, 249–255)*

Φ a single argument $\Phi_i \in Args := \varphi_i \rightarrow \varphi_j$ defined by two components $\varphi_i$and φ_j *(pp. xxv–xxviii, 38, 39, 99, 101, 108–115, 119, 121, 131, 136, 138, 141, 148, 159, 160, 164, 167, 235, 236, 240–242)*

φ a single component (an argument Φ_i consists of two components φ_i and φ_j and a relation $\rightarrow$ between them) *(pp. xxv–xxviii, 37, 38, 97–101, 114, 115, 118, 121–123)*

$\boldsymbol{\phi}$ a feature vector $= (\phi_0, ..., \phi_n)^T$ encoding an RL state $s_i \in \mathcal{S}$ as vector *(pp. xxvii, 44–51, 141)*

ϕ a single feature value of an RL feature vector $\boldsymbol{\phi}$ *(pp. xxvii, 44–51, 141)*

π a strategy of an RL optimization problem *(pp. 40–45, 138, 139, 141)*

Ψ the numerical strength of an argument *(pp. 112, 113, 119)*

Q the q-value function defining the accumulated rewards given an RL state $s_i \in \mathcal{S}$ encoded as feature vector $\boldsymbol{\phi}(s_i)$, and the parameter vector $\boldsymbol{\omega}_i$ of an action $a_i \in \mathcal{A}$ *(pp. 42–45, 48–51, 138, 139, 141)*

r the Wilcox effect size of a statistical Wilcoxon test *(pp. 123, 150–155, 160, 171–174, 176, 177, 249–255)*

$\mathbb{R}$ the set of real numbers *(pp. 17, 41, 42, 44, 50, 70)*

R a relevance value of a neuron of a neural network to describe its influence on the following neurons *(pp. 33–35, 87)*

$\mathcal{R}$ the reward function defining the reward when executing an action $a_i \in \mathcal{A}$ within state $s_i \in \mathcal{S}$ *(pp. 41–45, 49–51, 136–138, 141)*

ρ the sample Pearson correlation coefficient *(pp. 28, 29)*

$\mathcal{S}$ a set of RL states $\{s_1,...,s_n\}$ *(pp. xxvii, 41–48, 135–137)*

s an RL state *(pp. xxvii, 41–51, 136–138, 141)*

σ the standard deviation score of a set of values *(pp. 28, 147, 151, 153, 233, 234)*

$\mathcal{T}$ the transition probability to get from RL state $s_i \in \mathcal{S}$ to state $s_j \in \mathcal{S}$ using action $a_k \in \mathcal{A}$ *(p. 41)*

ω a weight of a neural network, or the weight vector of RL, or the (normalized) weight of an argument *(pp. 16, 17, 19–21, 33–35, 44, 48–51, 112–115)*

$\boldsymbol{W}$ a matrix of weights *(pp. 16, 18, 21, 22, 24, 25)*

x an entry of an input vector $\boldsymbol{x}_i$ of a neural network *(pp. xxviii, 16, 17, 21–23, 81)*

$\boldsymbol{x}$ an input vector $(x_i, \ldots, x_n)^T$ of a neural network *(pp. xxvii, xxviii, 16, 18, 21, 22, 69, 70, 81, 87, 90)*

$\mathcal{X}$ the set of input samples $\{\boldsymbol{x}_i, \ldots, \boldsymbol{x}_n\}$ for a classificator *(pp. 16, 18, 23, 24, 69, 70, 87–90)*

y an entry of an output vector $\boldsymbol{y}_i$ of a neural network *(pp. xxviii, 16–21, 32–35, 74, 75, 87)*

$\boldsymbol{y}$ an output vector $(y_i, \ldots, y_n)^T$ of a neural network *(pp. xxviii, 16–19, 21, 70)*

$\mathcal{Y}$ the set of desired output samples $\{\boldsymbol{y}_i, \ldots, \boldsymbol{y}_n\}$ for a classificator *(pp. 16, 69, 70)*

z the numerical effectiveness of an argument Φ_i (component φ_i) *(pp. 110, 111, 114, 121–123, 136, 141, 148, 159, 160, 164, 167, 235, 236, 240–242)*

CHAPTER 1

MOTIVATION

"Learning without reflection is a waste. Reflection without learning is dangerous." (Confucius)

1.1 Introduction

While arguments are most commonly associated with debates, they are an essential part of our everyday conversations. We rely on arguments to make decisions, even as simple as choosing what to have for lunch. Our choices are not always driven solely by rational thinking. While many might deny it when asked, various other factors can influence our opinions, such as the emotional tone used by the speaker, the speed of their speech, our pre-existing beliefs, or other personal influences.

Imagine you are in England for an internship, have made lovely local friends, and are planning your Friday night out trip. Everyone has different ideas for how to spend the evening, with some suggesting the cinema and others wanting to go clubbing; you have yet to have a real preference. Eventually, you decide for the cinema, leading to another discussion: Which movie? An intense argument is about to start. As a logical person, you consider all the critical aspects - the type and length of the movie, the actors, etc. What would your list look like to make that decision? Would it matter? What if someone you like suggested a romantic movie? Would that influence your opinion? Even if you consider yourself a hundred percent logical and rational, psychological evidence suggests that peripheral factors facilitate peripheral persuasion of weak arguments (Griskevicius et al., 2010).

Of course, we are susceptible to persuasion not just in daily conversations - the consumer industry constantly manipulates us, from using celebrities in ads to carefully staging products. Did you know that food advertisers often use unwholesome ingredients to make food look more appealing? From deodorant to make fruits look shiny to engine oil instead of syrup (Brightside, 2023) because the food does not absorb it.

When you think of an advert as an argument for why to buy a product, it seems that adverts often rely on pure claims and often manipulation, deception, and emotive persuasion to convince you to buy the products (Danciu, 2014). And people fall for these tactics all the time.

Take wine, for example. There are hundreds of different types of wine, each with a complex set of flavors and aromas. Wine can be tense, woody, fruity, or tannic (Mataillet, 2019), but who knows anything about that? When I buy a bottle of wine, I usually look for something sweet, not too bitter, because I am not fond of bitter wines. There are so many flavors that wine connoisseurs can distinguish. However, the truth is, my final decision depends only on two aspects: 1) Does it say *sweet* on the label, and 2) does the label look good? Many can relate to this. It is not a rational decision since the label hardly says anything about the quality of the wine. However, the truth is that there are too many factors to consider, and thus, in the end, my decision is based on non-rational ones.

Advertising and manipulation are not limited to commercial products but also extend to politics, where arguments should matter more than how they are conveyed. However, in recent years, we have seen a shift towards focusing on the politicians' personalities and demeanor rather than the actual issues at hand. Election campaigns, in general, are meant to inform people about politicians' goals. However, these campaigns are essentially advertisements, and like commercial ads, they use various manipulative techniques to get the people's votes.

Let us take the Brexit campaign as an example since I was living in England and witnessing the entire campaign during that time. The infamous red bus claiming that the UK pays 350 million pounds per week to the EU was a widely circulated claim during the campaign. The question is: Is that true? 1985, the United Kingdom was granted a rebate, reducing its net contribution to approximately

66% (D'Alfonso, 2016; Great Britain & Treasury, 2018).

The problem is that people are susceptible to manipulation by psychological aspects, such as the emotional tone throughout the campaign, which greatly influenced the campaign's outcome. The emotional tone of a message plays a significant role in determining its effectiveness (van Kleef, 2014; van Kleef et al., 2015). This applies to the sender's and receiver's emotions of a message. Drawing from personal experiences, we know the ease with which individuals can be manipulated based on pre-existing biases and prejudices. This phenomenon is particularly prevalent in political campaigns, where emotions are frequently leveraged to get votes. As such, it is important to be aware of the manipulative techniques commonly used to make more informed decisions. However, people are usually unaware that they are being manipulated.

1.2 The Continuum of Manipulation

Manipulation is a general term that can describe any non-rational influence; however, it distinguishes several sub-forms and *degree of freedoms*. Buss (1987) defines manipulation as "*the ways in which individuals intentionally or purposefully [...] alter, change, influence, or exploit others*", not necessarily with evil intent, though (Buss, 1987). This is in line with literature distinguishing between several degrees of freedom, that are *persuasion*, *manipulation*, and *coercion* (Sorlin, 2016). According to Sorlin (2016), manipulation lies between persuasion and coercion on the degree of freedom scale see Fig. 1.1).

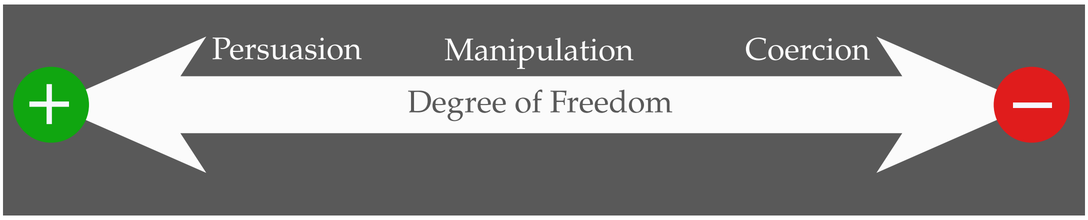


Figure 1.1: Persuasion-Manipulation-Coercion: Degree of Freedom. While persuasion provides the highest degree of freedom to argue, coercion means an obligation to comply (Figure adapted from Sorlin (2016)).

That, however, does not imply that these are distinct. Gass and Seiter (2018) proposed that "*persuasion involves one or more persons who are engaged in the activity of creating, reinforcing, modifying, or extinguishing beliefs, attitudes, intentions, motivations, and/or behaviors within the constraints of a given communication context*", declaring a higher degree of freedom between the individuals involved compared to manipulation. Susser et al. (2019) describes manipulation as "*hidden influence*", i.e., people are not aware of the process of manipulation affecting their decision-making process. This is contrary to persuasion, which "*means attempting to influence someone by offering reasons they can think about and evaluate*" (Susser et al., 2019). Coercion has the lowest level of freedom and means "*influencing someone by constraining their options, such that their only rational course of action is the one the coercer intends*" (Wood, 2014; cited by Susser et al., 2019).

Another form of influencing someone is called *deception*, which is a "*way to covertly influence someone [by planting] [...] false beliefs*" (Susser et al., 2019). For instance, your partner might lie to get you to clean the house by claiming relatives are visiting, thus inevitably arousing false beliefs to facilitate a rational decision that accommodates the manipulator's desires (Susser et al., 2019).

Thus, *persuasion* and *coercion* differ in terms of the degree of freedom, and *coercion* and *deception* differ in terms of conscious awareness. However, all of them can be manipulative following the definition by Buss (1987). Handelman (2009, p.25) grouped them into the continuum of manipulation consisting of three dimensions, that are *level of misleading* (deception), *level of control* (coercion) and *level of influence (persuasion)* (see Fig. 1.2).

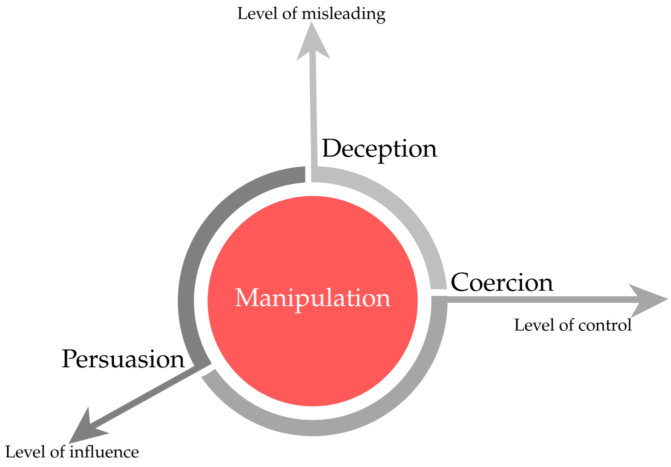


Figure 1.2: Continuum of Manipulation (Figure adapted from Handelman (2009, p.25)).

Think of the *house-cleaning* example. It is *manipulation* because low-level false beliefs deceived you. Contrary to coercion, which needs a conscious level of control, your partner subconsciously controls you. Also, an argument in the form of a lie was used, and thus, it is part of the *persuasion* dimension.

A last form of influence is called *nudging*. Whether it is manipulative depends on the form of nudging (Susser et al., 2019). *Nudges* can be either *overt* or *covert*, but not every covert nudge can be considered *manipulative*. This depends on whether it is trying to exploit or rectify something worse. An advert, for instance, can be regarded as manipulative. However, fair trade labels try to draw attention to how people may have been exploited for other products.

1.3 Motivation and Scope of the Thesis

When it comes to daily decision-making, considering arguments play an essential role. Depending on the context, arguments can be persuasive, manipulative, deceptive, or coercive. In this thesis, we focus primarily on persuasive arguments and, to a lesser extent, manipulative or deceptive ones. There are two main issues that we aim to tackle within this thesis:

- People's tendency to search for arguments in line with pre-existing opinions.
- Subliminal influence of social cues and emotions on argument perception.

People tend to focus on sources aligning with their opinions (Ekström et al., 2022). Why is that? Before the advent of the Internet and the easy accessibility of information, people had to rely on traditional news media such as TV and newspapers. The amount of information we had to process was limited compared to what is available now. However, the more information becomes available, the harder it is to sort through them and establish a well-founded opinion. This was particularly evident from 2020 to about 2022, when the world grappled with the COVID-19 pandemic, accompanied by an overwhelming, constant stream of (often contradicting) ever-shifting information shared over the Internet and social media.

Social media reinforces the problem by filtering out information based on users' past requests (Pariser, 2012). Thus, due to a shift from face-to-face to online discussions, it is even more likely that people focus on sources that further repeat and reinforce a pre-existing opinion, preventing them from re-evaluating established opinions, a phenomenon also known as *confirmation bias* (U. Peters, 2022).

Adding to this issue, emotional discussions in social media comment sections create a *subliminal bias* on argument perception. This is because social cues, such as emotions, play a significant role in how arguments are received and processed (van Kleef, 2014; van Kleef et al., 2015). The problem is that the tone and context are often misinterpreted in online formats, leading to heightened emotional responses.

A driving factor is the people's system of thinking (Kahneman, 2012b) used to process information. *Fast thinking* (System 1) is fast, automatic, intuitive, and often subconscious. It operates quickly to make judgments and decisions with minimal effort. *Slow thinking* (System 2), on the other hand, is conscious, slower and more analytical. It involves critical thinking and requires more cognitive resources (Kahneman, 2012b). When bombarded with vast information from social media and online sources, individuals often rely on *fast thinking* to form judgments and opinions without profoundly analyzing the information. However, navigating through the overwhelming volume of information requires individuals to engage *slow thinking* to critically evaluate arguments, assess evidence, and weigh different perspectives.

Reasons to apply *fast thinking* can be, among others, low personal relevance or low Need for Cognition (NFC) [1] to think about arguments, which is one of many variables that drives the willingness to take that effort (Petty & Cacioppo, 1986). This ultimately gives way to *peripheral processing* (Petty & Cacioppo, 1986), which means instead of being persuaded by the content of arguments, one is persuaded by everything else outside of the argument. *Peripheral processing* is an unconscious process moderated by low elaboration inducing a *subliminal bias* as indicated by J. Peters and Hoetjes (2017) who found that, when gestures are used, people with low elaboration are significantly more likely to rate a given speech as factual accurate (even though it is not) compared to people with high elaboration.

High elaboration is often driven by the individual's NFC (Dole & Sinatra, 1998), which however primarily indicates a *motivation* to engage in cognitive processes but not necessarily the *ability* (APA Dictionary of Psychology, 2023). In addition, *fast thinking* cannot be deactivated (Kahneman, 2012a), making individuals consistently susceptible to biases, such as *confirmation bias* and *subliminal bias*. We refer to any of these biases as *reflection bias*, and within this thesis, we distinguish between two types (Walton, 2005, p.218):

- *Content-based*: Tendency to focus on sources aligning with one's opinion and ignore specific arguments because they do not fit one's point of view.
- *Behavior-based*: The subliminal influence of social cues and (own/other's) emotions on perception of arguments due to peripheral processing.

The tendency to focus on biased sources indicates a general tendency of *low elaboration* by minimal usage of cognitive resources. Thus, there is an increased risk of *peripheral processing* of information, and therefore, a *behavior-based* reflection bias. This exacerbates the issue on platforms like social media, making it important to address these causes by developing better applications for information engagement. This can foster a more reflective and informed public discourse and decision-making.

Consequently, in the scope of this thesis, we address the problems of *reflection biases*. First, we indirectly foster reflection by raising awareness of the *behavior-based*

[1] *"A personality trait reflecting a person's tendency to enjoy engaging in extensive cognitive activity. This trait primarily reflects a person's motivation to engage in cognitive activity rather than their actual ability to do so. Individuals high in need for cognition tend to develop attitudes or take action based on thoughtful evaluation of information."* (Definition by APA Dictionary of Psychology (2023)).

reflection bias as outlined in 1.3.1. Secondly, we directly foster reflection and *slow* thinking, referred to as Reflective Engagement (RE) [2], as outlined in 1.3.2. In addressing behavior-based reflection bias, we focus on analyzing gestures rather than emotions due to the multifaceted role in communication. Drawing upon the findings from J. Peters and Hoetjes (2017), which highlight the interaction between elaboration and gestures regarding rated factual accuracy, we conclude that the concept of *behavior-based* reflection bias encompasses non-verbal behaviors as a whole. Fig. 1.3 gives a total overview of the scope of this thesis, the relevant terms, and their relationship.

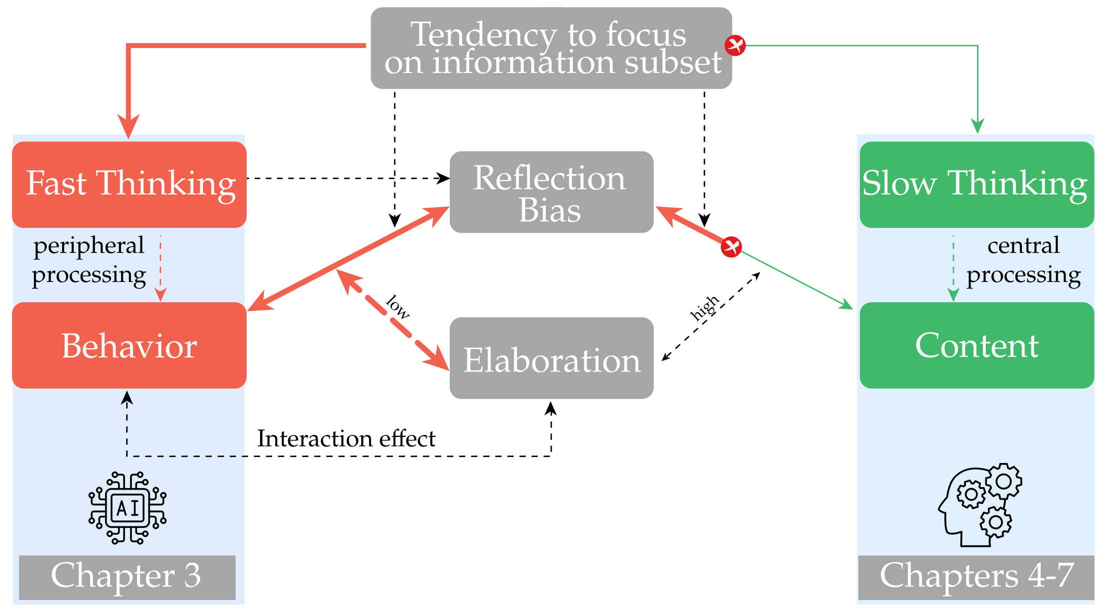


Figure 1.3: Scope of this thesis: The tendency to focus on a subset of information indicates *low elaboration*, and a *content-based* reflection bias (✖), increasing the risk of a *behavior-based* reflection bias. We address both issues by 1) Raising awareness of *behavior-based* reflection bias (left, Sec. 1.3.1) and 2) Mitigating *content-based* reflection bias by fostering RE (right, Sec. 1.3.2).

In Ch. 3, we focus on raising awareness of the *behavior-based* reflection bias induced by gestures in the context of political debates employing XAI. The central focus is on analysis and creating awareness rather than suggesting changes in behavior.

In Ch. 4, we present the prototype of an intelligent conversational agent guiding the user's argument exploration focus using intervention strategies to mitigate the

[2] RE is defined as a *"learner's continual and active participation in their problem inquiry with a continuous and critical judgment of inquiry process and inquiry outcomes for possible improvement"* (Farr and Riordan, 2012; Lyons, 2006; Rodman, 2010; cited by Kong and Song, 2015). Within the domain of argumentation, we define it as *the user's exploration of diverging views*.

content-based reflection bias. For this, we define a metric for RE and analyze the interventions' effectiveness through various user studies and experiments. The central focus encompasses monitoring argument focus using the metric (Ch. 5), employing intervention strategies (Ch. 6), and evaluating the approach (Ch. 7).

1.3.1 Raising Awareness of the Subliminal Reflection Bias

Especially *fast thinking* and *low elaboration* can favor *peripheral processing* and thus lead to a behavior-based reflection bias induced by subliminal persuasion.

A study by Zanot et al. (1983) found that people mainly associate subliminal persuasion with unethical advertising rather than recognizing its role in their daily decision-making process. Thus, many people misunderstand the concept of subliminal persuasion and must be made aware of this.

This is because most literature primarily investigates *what* stimuli make a speaker persuasive but not *why* a speaker is persuasive (see Fig. 1.4). Simply modifying stimuli contributes to an understanding of visible, allegedly supraliminal cues (e.g., gesture vs no gesture); however, it does not help raise awareness of the subliminal influence of these cues and, thus, makes people more vulnerable to fake news. A study by Allcott and Gentzkow (2017) found *"that the average US adult read and remembered [...] one or perhaps several fake news articles during the [2016 presidential] election period"*. Another study by Rogers and Smith (1993) supports this finding, highlighting the need for more awareness and understanding of subliminal persuasion.

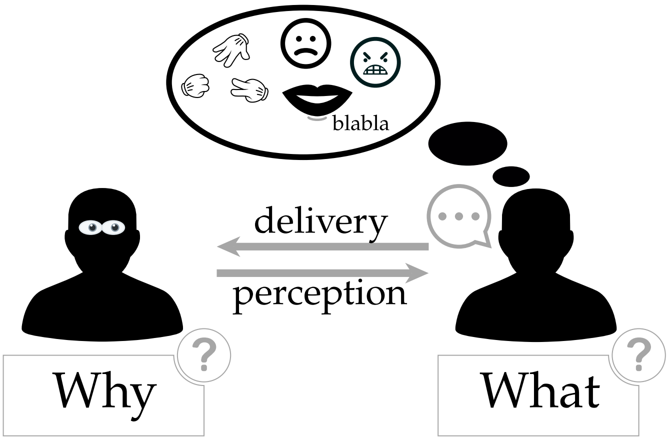


Figure 1.4: Investigation of *why* (receiver's perception) vs. *what* (sender's stimuli).

The first part of this thesis investigates *why* a speaker is found persuasive and compares the different perceptions of several annotators to raise awareness of the subliminal bias and, thus, the induced *behavior-based reflection bias*. We train a Neural Network (NN) model to predict perceived persuasiveness based solely on input video frames (see Sec. 3.2). Three students annotate a political debate based on their perception of persuasiveness using the video's visual and audio output channels. Subsequently, we train a model using only the visual channel and the aggregated annotated data. We then leverage various XAI techniques to highlight what the network focuses on and *why* speakers are perceived as persuasive.

There are two main challenges:

⚙ Challenges

C1.1 Persuasiveness is subjective, making training accurate prediction models quite challenging.

C1.2 The lack of satisfactory explanations as to why a speaker is perceived as persuasive poses another challenge.

The first challenge (**C1.1**) concerns the subjectivity of data, making it challenging to train accurate prediction models. The subjectivity of the data increases the risk of overfitting, where the model may excessively learn from specific nuances of the subjective data, and underfitting, where the model cannot capture essential patterns. Thus, we apply a bias-variance trade-off (Geman et al., 1992), which involves balancing the model's bias to reduce underfitting while also keeping the variance of the model low to minimize overfitting.

Challenge **C1.2** concerns the lack of satisfactory explanations for why a speaker is perceived as persuasive. This gap in research makes it challenging to verify the generated visualizations and explanations. Most existing work focuses on identifying *what* stimuli make speakers persuasive rather than providing comprehensive explanations for perceived persuasiveness. Consequently, analyzing visualizations regarding validity and fidelity becomes particularly challenging and necessary. We compare the findings of the XAI visualizations with existing literature and analyze whether or not the visualizations align with the literature to verify validity and fidelity. After that, we train separate models for each student instead of aggregating the data within one model to allow for comparison of perceived persuasiveness based on an extended dataset of another annotated thirty speeches.

By comparing the highlighted explanations of the models' output and examining the variations in the visual cues identified by each network, we raise awareness of subjective persuasive markers and their individual interpretations if we can show that they also align with similar findings in the literature. If an annotator did not focus on them, the network would likely not learn to focus on them. Consequently, there are two main research questions following the challenges:

Research questions:

Q1.1 Can we effectively uncover behavior-based reflection bias in political speeches and provide satisfactory explanations(**C1.1**, **C1.2**, Sec. 3.2)?

Q1.2 Can XAI contribute to highlighting and understanding subjective differences in persuasive cues in political speeches(**C1.1**, **C1.2**, Sec. 3.3)?

1.3.2 Engaging Users in the Critical Reflection of Arguments

Existing applications of argumentation often revolve around enhancing speaking skills in effective argumentation and debates, which translates to learning how to defend one's points of view most effectively (see Sec. 2.3.2). However, such approaches do not necessarily lead to an unbiased look at diverging views but foster a defense of one's own.

To address this gap, we introduce BEA, a chatbot-like argumentative intelligent agent allowing users to explore the pro and contra sides of a controversial topic. If users stick to arguments supporting their own opinion, the intelligent agent can intervene by proposing so-called *challenger arguments* that challenge the users' position to promote a less biased argument exploration. To do that, we equip the intelligent agent with a computational metric, namely Argument Visitation Quotient (AVQ), gauging the extent to which users predominantly focus on arguments that align with their viewpoint. There are two main challenges:

Challenges

C2.1 Developing an engaging conversational agent (CE) that proactively encourages and motivates users to engage with arguments from diverse viewpoints (RE).

C2.2 Defining a computational metric for RE to assess the user's argument exploration focus during interaction and assess the agent's effectiveness in fostering critical analysis of divergent viewpoints.

The first challenge (**C2.1**) concerns the Conversational Engagement (CE) and Reflective Engagement (RE) of the user. While RE refers to the user's level of exploring diverging views as defined above, the Conversational Engagement (CE) defines how individuals create and sustain a connection while engaging in shared activities (Sidner et al., 2004). We thereby look into three aspects, which are 1) the linguistic style, 2) gamification strategies, and 3) the effect of agent embodiment (see Fig. 1.5).

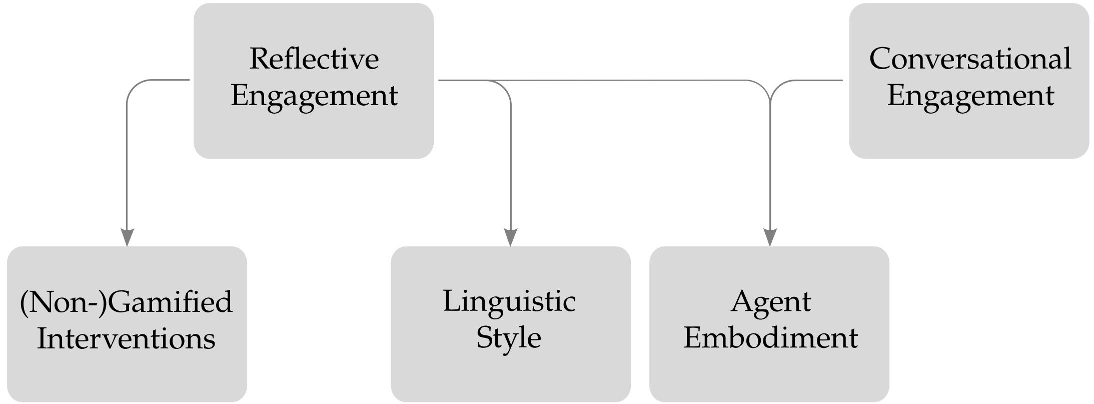


Figure 1.5: Outline: We investigate the effect of linguistic style, gamification strategies, and agent embodiment on RE and CE.

Similar to human-human interaction, how arguments are presented may influence the user's willingness to engage in a critical reflection. Especially when interacting with conversational agents, the user's engagement and motivation are important factors that highly influence the success or failure of such a mixed team. Thus, to increase enjoyment and the motivation to interact with the conversational agent, we equip the agent with two different presentation modalities (*chat* vs. *embodied*) and investigate the effect of the *agent's embodiment* on the CE, and RE.

Further, the way (*linguistic style*) the intelligent agent intervenes (e.g., polite vs. impolite) when guiding the user to be less biased may also influence the perception of the agent and influence the agent's persuasiveness (Hammer et al., 2016). We implement *(non-)adaptive* (linguistic style) and *(non-)gamified* intervention strategies that allow the agent to encourage users to consider alternative perspectives proactively. The agent can employ these strategies to encourage a different argument exploration behavior.

To do so, we define a computational metric for RE (challenge **C2.2**) assessing the argument exploration focus. Subjective measures of RE involve questionnaires (Kember et al., 2000; Lee & Dey, 2011; Leijen et al., 2009), which are unsuitable metrics during interaction. Objective measures encompass measurable outcomes (Govaerts et al., 2012; Kharrufa et al., 2010; Santos et al., 2013), user free-text statements (Farr & Riordan, 2012), time metrics (Dupret & Lalmas, 2013), user interaction behavior (Arapakis et al., 2014; Ponnuswami et al., 2011) and user focus (Yi et al., 2014).

Following the agent's goal to guide the user's argument focus, we use a computational metric based on the user's argument focus, namely Argument Visitation Quotient (AVQ), gauging the extent to which users predominantly focus on arguments that align with their viewpoint. The so-defined metric AVQ thereby yields a low value if the users focus on arguments that align with their stance.

Following the challenges, we overall defined seven research questions to be examined within this thesis.

? Research questions:

Q2.1 How can we formulate a computational metric for RE that is operationally feasible and can be programmatically implemented allowing the agent to guide the user's argument visitation focus (**C2.2**, Ch. 5)?

Q2.2 Does the intervention mechanism impact the user's engagement with challenger arguments, i.e., leads to an increase of AVQ (**C2.1**, Ch. 4, 6 + 7)?

Q2.3 Do the gamification mechanism and agent embodiment affect intervention success positively (**C2.1**, Ch. 4, 6 + 7)?

Q2.4 Does agent embodiment affect system perception, trust, and the user's CE positively (**C2.1**, Ch. 4, 6 + 7)?

Q2.5 Do interventions affect user trust negatively (**C2.1**, Ch. 4, 6 + 7)?

Q2.6 Do interventions impact users' eye gaze behavior (attention to arguments) (**C2.1**, Ch. 4, 6 + 7)?

Q2.7 Is there an interaction effect of User Characteristics (UCs) on the exploration behavior, and agent interactions (**C2.1**, Ch. 4, 6 + 7)?

The first research question (**Q2.1**) is derived from challenge **C2.2** and thus concerns the development of the metric AVQ, which is needed to track and guide the user's argument visitation focus.

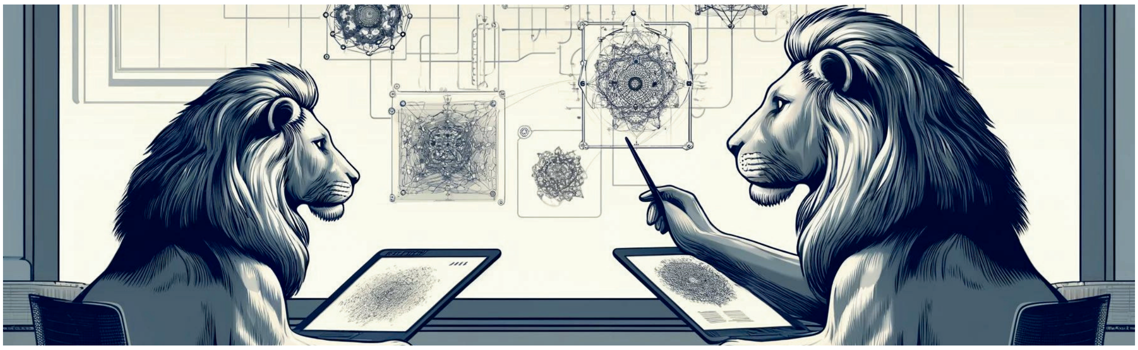
The others are derived from challenge **C2.1**. With the research questions **Q2.2** and **Q2.3**, we investigate the effect of the *gamified* and *non-gamified* intervention strategies on AVQ, meaning that if the metric score increases, the interventions are successful. We thereby also investigate if there is an interaction effect between intervention success and *agent embodiment*. Next, to account for the user's CE, we investigate the impact of *embodiment* on perception, trust, and CE with research question **Q2.4**. Hancock et al. (2011) conducted a meta-analysis of factors affecting trust when interacting with robots, highlighting the significance of performance-related factors such as behavior, dependability, and the level of automation. This aligns with Rezaei Khavas (2021) who also identified performance- and behavior-related factors, such as autonomy level, likability, and personality, that can affect trust. Thus, when actions are perceived as controlling or coercive, they can negatively impact and reduce trust over time due to trust being a dynamic variable (Rhim et al., 2023). As a decline in user trust would hurt the user's willingness to interact with the intelligent agent, we also investigate the effect of interventions on trust (**Q2.5**) to evaluate whether trust decreases.

While the metric AVQ assesses argument selection, we specifically aim for users to actively process the presented arguments, not merely navigate the system and follow interventions. Thus, we also collect eye-tracking data to explore how interventions shape users' processing of arguments, specifically whether users actively engage with opposing viewpoints (**Q2.6**).

Last but not least, personality traits (e.g., Need for Cognition (NFC) and Conscientiousness (CS)) and cognitive abilities (e.g., Perceptual Speed (PS)) have been identified as significant factors affecting user behavior and performance (Conati et al., 2021; Toker & Conati, 2014; Toker et al., 2013; Ziemkiewicz et al., 2011). Thus, with the last research question **Q2.7**, we investigate if the interventions' effectiveness is affected by User Characteristic (UC) to allow for the development methods for personalization and enhancing the subjective experience.

Chapter 2

Related Work and Background



"Critical thinking is thinking about your thinking while you're thinking in order to make your thinking better." (Richard W. Paul)

ℹ Parts of this chapter were previously published by the author in peer-reviewed papers (Weber et al., 2020c, 2023b), *reproduced with permission from Springer Nature*, and other own publication as listed in Annex I.

Related work consists of three parts: 1) technical background (divided into i. neural networks - overview, ii. explainable artificial intelligence, iii. argumentation - theory and notation, and iv. reinforcement learning), 2) background on psychological models on message processing and 3) reflection bias and reflective engagement.

Neural networks (Sec. 2.1.1) are used in Sec. 3 to train models predicting persuasiveness from annotated data. Sec. 2.1.1 gives a technical overview of how neural networks work and introduces the herein-used notation. Explainable artificial intelligence (Sec. 2.1.2) is then used to highlight what the networks focus on to make their prediction and to draw conclusions if the networks focus on persuasive social behavioral cues that are considered persuasive according to literature. This section thereby describes the mathematical background of the applied XAI techniques Grad-CAM (Sec. 2.1.2.1) and LRP (Sec. 2.1.2.2).

In Sec. 2.1.3 we give an overview of the argument notation used in this thesis. As we later present an approach to adapt the intervention strategies (see Ch. 6), we give a thorough overview of RL in Sec. 2.1.4.

Sec. 2.2 covers message processing models; in Sec. 2.3, we cover related work about reflection and establish boundaries that are solved within this thesis.

2.1 Technical Background

2.1.1 Neural Networks - Technical Overview

2.1.1.1 NN: Standard Neural Networks

NNs are among today's most powerful machine-learning techniques. This section is based on Goodfellow et al. (2017) and Rumelhart et al. (1986).

Mathematically, a neural network is a non-trivial, mostly nonlinear function $\mathcal{B} : \mathcal{X} \rightarrow \mathcal{Y}$ with $\boldsymbol{x} \in \mathcal{X}$ the input vector and $\boldsymbol{y} \in \mathcal{Y}$ the target output vector.

A neural network consists of at least two layers, the input layer of dimension $dim(\boldsymbol{x}) = n$ and the output layer of dimension $dim(\boldsymbol{y}) = m$. Each input of the vector defines a neuron of the respective layer. It connects to each neuron of the next layer. To each of these links from the x_i neuron to the y_j neuron, we assign a weight ω_{ij} (see Fig. 2.1).

The weights between the input and output layers are represented as a matrix $\boldsymbol{W}$. Each row i consists of all the weights the input x_i connects to, while each column j consists of all the incoming connections of the neuron y_j.

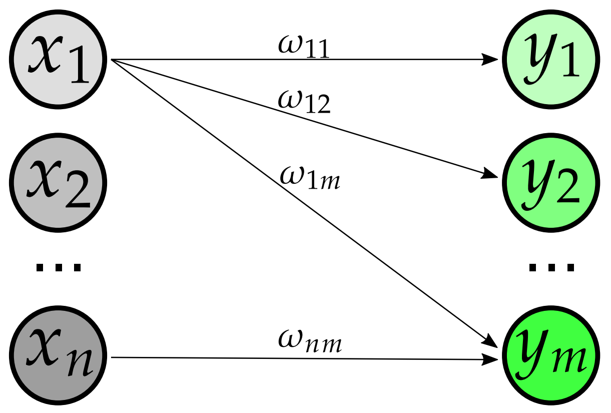


Figure 2.1: Concept of a neural network consisting of an input vector $\boldsymbol{x} = (x_1, \ldots, x_n)$, an output vector $\boldsymbol{y} = (y_1, \ldots, y_m)$ connected by weights $\omega_{ij}, i \leq n, j \leq m$.

$$\text{Input layer}\left\{\overbrace{\begin{pmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1m} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{n1} & \omega_{n2} & \cdots & \omega_{nm} \end{pmatrix}}^{\text{Output layer}}\right. \tag{2.1}$$

The weighted sum vector $\tilde{\boldsymbol{y}}^T$ is calculated by multiplying the input vector by the weight matrix, i.e.,

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}^T \times \begin{bmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1m} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{n1} & \omega_{n2} & \cdots & \omega_{nm} \end{bmatrix} = \begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \vdots \\ \tilde{y}_m \end{bmatrix}^T = \tilde{\boldsymbol{y}}^T \tag{2.2}$$

Each layer can have a bias vector $\boldsymbol{b}$ with $dim(\boldsymbol{b}) = dim(\boldsymbol{y})$ that is added to the multiplication changing Eq. 2.2 to

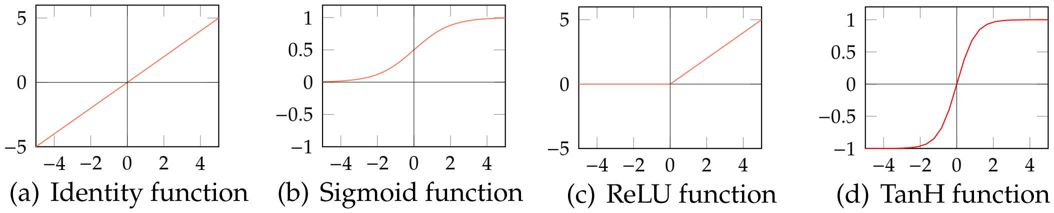
$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}^T \times \begin{bmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1m} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{n1} & \omega_{n2} & \cdots & \omega_{nm} \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}^T = \begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \vdots \\ \tilde{y}_m \end{bmatrix}^T = \tilde{\boldsymbol{y}}^T \tag{2.3}$$

The bias $b_j \in \boldsymbol{b}$ is a scalar value that increases or decreases the influence of all incoming links, depending on whether the bias is positive or negative.

Finally, an activation function $act : \mathbb{R}^m \longrightarrow \mathbb{R}^m$ is applied inputwise to the weighted sum vector $\tilde{\boldsymbol{y}}^T$, giving us the output vector $\hat{\boldsymbol{y}}^T \approxeq \boldsymbol{y}^T$:

$$act\left(\begin{bmatrix} \tilde{y}_1 & \cdots & \tilde{y}_m \end{bmatrix}\right) = \begin{bmatrix} act\left(\tilde{y}_1\right) & \cdots & act\left(\tilde{y}_m\right) \end{bmatrix} = \hat{\boldsymbol{y}}^T \tag{2.4}$$

A trivial activation is the identity function $\hat{\boldsymbol{y}} = \tilde{\boldsymbol{y}}$ (Fig. 2.2(a)). Fig. 2.2 shows other common activation functions (see Goodfellow et al. (2017, pp. 169-189)).



(a) Identity function (b) Sigmoid function (c) ReLU function (d) TanH function

Figure 2.2: Several activation functions: (**a**) Identity, (**b**) Sigmoid, (**c**) Rectifier Linear Unit (ReLU), and (**d**) Tangents Hyperbolic

Between the input layer and the output layer, there may be hidden layers, each of which is connected to the preceding layer so that the network function $\mathcal{B}$ is defined by

- a set of weight matrices $\{\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_n\}$, and
- a set of biases $\{\boldsymbol{b}_1, \boldsymbol{b}_3, \ldots, \boldsymbol{b}_n\}$

Given an input vector $\boldsymbol{x} \in \mathcal{X}$, the final network's output $\hat{\mathcal{B}}(\boldsymbol{x}) = \hat{\boldsymbol{y}}_n$ is calculated recursively as ($n \geq 1$. Note that $\hat{\boldsymbol{y}}_0 := \boldsymbol{x}$):

$$\hat{\boldsymbol{y}}_n = act(\hat{\boldsymbol{y}}_{n-1} \times \boldsymbol{W}_n + \boldsymbol{b}_n) \tag{2.5}$$

The crucial part is to find the weights and biases of the network function $\hat{\mathcal{B}}$ so that the actual output vector $\hat{\boldsymbol{y}}_n$ matches the target output vector $\boldsymbol{y}_n$.

$$\hat{\boldsymbol{y}}_n = \hat{\mathcal{B}}(\boldsymbol{x}) \approxeq \mathcal{B}(\boldsymbol{x}) = \boldsymbol{y}_n, \forall \boldsymbol{x} \in \mathcal{X} \tag{2.6}$$

As a consequence, training a neural network is an optimization problem of finding the weights and biases that minimize the Mean Squared Error (MSE) between the actual output $\hat{y}_j \in \hat{\boldsymbol{y}}_n$ and the target output $y_j \in \boldsymbol{y}_n$.

$$E = -\frac{1}{2}\left(y_j - \hat{y}_j\right)^2 \tag{2.7}$$

A commonly used method for finding the minimum of an error function E is *gradient decent*. The gradient of a multivariate n-dimensional function defines the steepest ascent/descent (Karpfinger, 2017, pp.495-500). Mathematically, it is a partial variable-wise derivative, assuming that the multivariate variables are independent. Thus, it is a special case of the total derivative. The gradient defines the steepest directional slope in all n dimensions, and is defined as (Def. 2.1, see Ex.2.1)

Definition 2.1: Gradient

Let f be a differentiable n-dimensional function with independent multivariate parameters $u_1, \dots, u_n$ and $\boldsymbol{e}_i$, $1 \le i \le n$, be the unit vector, then the partial derivative of f with respect to vector $\boldsymbol{u} = (u_1, \dots, u_n)$ is defined as:

$$\frac{\partial f}{\partial \boldsymbol{u}} = \boldsymbol{e}_1 \cdot \frac{\partial f}{\partial u_1} + \cdots + \boldsymbol{e}_n \cdot \frac{\partial f}{\partial u_n} \tag{2.8}$$

Example 2.1: Partial Derivative

Let $f := -u^2 + v$ a multivariate function with parameters u and v. We assume that u and v are independent. The gradient of $f(u, v)$ is calculated as:

$$\nabla f(u, v) = \begin{pmatrix} \frac{\partial(-u^2+v)}{\partial u} \\ \frac{\partial(-u^2+v)}{\partial v} \end{pmatrix} = \begin{pmatrix} -2u \\ 1 \end{pmatrix} \tag{2.9}$$

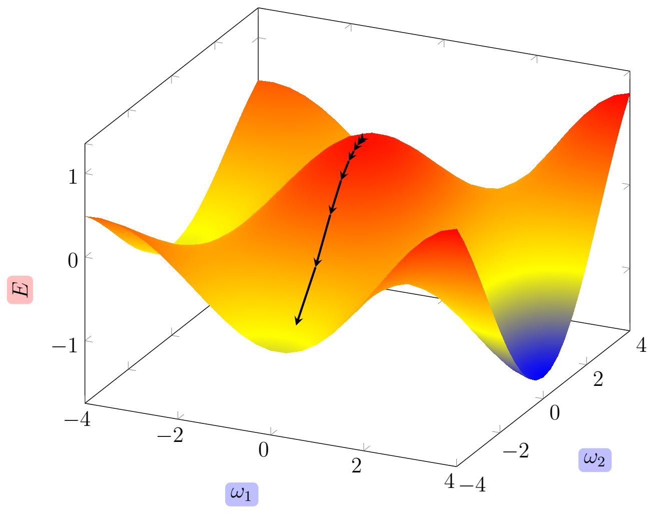


Figure 2.3: Gradient descent: The error function E in a 3-dimensional space depends on the two weights ω_1 and ω_2. The error is minimized by following the path of deepest decent (black) to find a local minimum.

Fig. 2.3 shows a sketch of the concept of the gradient with the two function variables ω_1 and ω_2 and the function result E. Since E is dependent on weights ω_{ij} (incorporated in $\hat{y}_j$), we can calculate the gradient for each of the weights, defined as the partial derivative of E with respect to variable ω_{ij}. The following derivation of the update rule for $\Delta\omega_{ij}$ is called *back-propagation* (Rumelhart et al., 1986):

$$\Delta\omega_{ij} = \frac{\partial E}{\partial \omega_{ij}} \tag{2.10}$$

Because E is a chain function (dependent on $\hat{y}_j$, and $\tilde{y}_j$), we apply the chain rule (Karpfinger, 2017, p.504) twice and get the following:

$$\Delta\omega_{ij} = \underbrace{\underbrace{\frac{\partial E}{\partial \hat{y}_j}}_{y_j - \hat{y}_j} \cdot \underbrace{\frac{\partial \hat{y}_j}{\partial \tilde{y}_j}}_{act\left(\tilde{y}_j\right)\cdot\left(1-act\left(\tilde{y}_j\right)\right)}}_{\delta_j} \cdot \underbrace{\frac{\partial \tilde{y}_j}{\partial \omega_{ij}}}_{\hat{y}_i} \tag{2.11}$$

Note: For the middle derivative $\frac{\partial \hat{y}_j}{\partial \tilde{y}_j}$ we assumed a sigmoid activation function. If another activation function is used, the derivative needs to be altered respectively.

The problem with Eq. 2.11 is that the desired target output vector y_j is only known for the output layer, yet not for the inner layers. Thus, $\frac{\partial E}{\partial \hat{y}_j}$ cannot be calculated. This means that for inner layers, we need to define the error function E as a function of the inputs $\left\{\tilde{y}_{x1}, \ldots, \tilde{y}_{xn}\right\}$ of all subsequent neurons that receive the output $\hat{y}_j$, i.e., the error function for inner neurons takes the form $E\left(\tilde{y}_{x1}, \ldots, \tilde{y}_{xn}\right)$.

Now that E is a multivariate function, dependent on $\left\{\tilde{y}_{x1}, \ldots, \tilde{y}_{xn}\right\}$, we can take the total derivative of E with respect to $\hat{y}_j$ as follows:

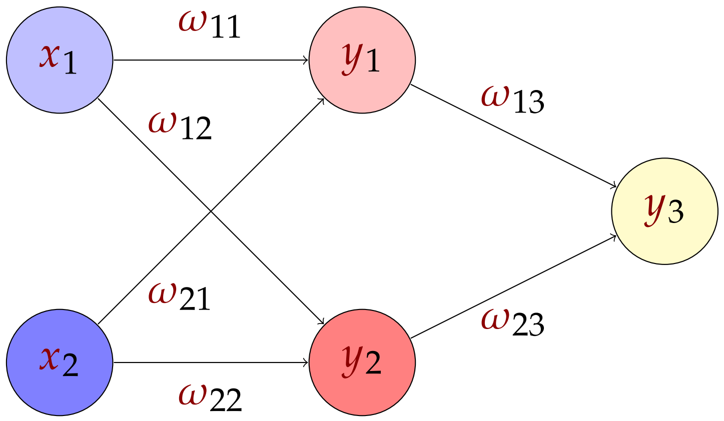
$$\frac{dE}{d\hat{y}_j} = \sum_i^n \frac{\partial E}{\partial \tilde{y}_{xi}} \cdot \frac{\partial \tilde{y}_{xi}}{\partial \hat{y}_j} = \sum_i^n \underbrace{\frac{\partial E}{\partial \hat{y}_{xi}} \cdot \frac{\partial \hat{y}_{xi}}{\partial \tilde{y}_{xi}}}_{\delta_{xi}} \cdot \underbrace{\frac{\partial \tilde{y}_{xi}}{\partial \hat{y}_j}}_{\omega_{j,xi}} \tag{2.12}$$

To summarize, the final weight update rule is as follows:

$$\Delta\omega_{ij} = \hat{y}_i \cdot \delta_j, \text{with } \delta_j = \begin{cases} \left(y_j - \hat{y}_j\right) \cdot \left(\frac{d}{d\,\tilde{y}_j} act\left(\tilde{y}_j\right)\right) & \hat{y}_j \text{ output} \\ \left(\sum_{i=1}^{n} \delta_{xi} \cdot \omega_{j,xi}\right) \cdot \left(\frac{d}{d\,\tilde{y}_j} act\left(\tilde{y}_j\right)\right) & \text{else} \end{cases} \tag{2.13}$$

Example 2.2: Backpropagation - Example

Let us consider the following example of a neural network:



We assume the identity function (cf. Fig. 2.2(a)) as activation function act for all neurons; therefore, it is omitted for readability and comprehensibility. The intended output y_3 is 1. Let $\boldsymbol{x} = \begin{bmatrix} 0.7, 0.8 \end{bmatrix}^T$ be the input vector and $\boldsymbol{W}_{12}$, $\boldsymbol{W}_{23}$ be the weight matrices between layers 1, 2 and 3 as follows:

$$\boldsymbol{W}_{12} = \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{bmatrix} = \begin{bmatrix} 0.1 & 0.3 \\ 0.2 & 0.4 \end{bmatrix}, \quad \boldsymbol{W}_{23} = \begin{bmatrix} \omega_{13} \\ \omega_{23} \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.6 \end{bmatrix}$$

We first compute the hidden output vector: $\hat{\boldsymbol{y}}_h$ (cf. Eq. 2.3 [a]):

$$\hat{\boldsymbol{y}}_h = \boldsymbol{W}_{12}^T \boldsymbol{x} = \begin{bmatrix} 0.1 & 0.2 \\ 0.3 & 0.4 \end{bmatrix} \begin{bmatrix} 0.7 \\ 0.8 \end{bmatrix} = \begin{bmatrix} 0.7 \times 0.1 + 0.8 \times 0.2 \\ 0.7 \times 0.3 + 0.8 \times 0.4 \end{bmatrix} = \begin{bmatrix} 0.23 \\ 0.53 \end{bmatrix}$$

and with it the output vector: $\hat{\boldsymbol{y}}$ of the neural network

$$\hat{\boldsymbol{y}} = \boldsymbol{W}_{23}^T \hat{\boldsymbol{y}}_h = \begin{bmatrix} 0.5 & 0.6 \end{bmatrix} \begin{bmatrix} 0.23 \\ 0.53 \end{bmatrix} = \begin{bmatrix} 0.23 \times 0.5 + 0.53 \times 0.6 \end{bmatrix} = \begin{bmatrix} 0.43 \end{bmatrix}$$

Next, we compute the gradient error vector δ_o of the output layer (cf. Eq. 2.13):

$$\delta_o = \boldsymbol{y} - \hat{\boldsymbol{y}} = \begin{bmatrix} 1.0 \end{bmatrix} - \begin{bmatrix} 0.43 \end{bmatrix} = \begin{bmatrix} 0.57 \end{bmatrix}$$

and update the weight matrix $\boldsymbol{W}_{23}$ (cf. Eq. 2.13):

$$\boldsymbol{W}_{23}' = \begin{bmatrix} 0.5 + 0.23 \cdot 0.57 \\ 0.6 + 0.52 \cdot 0.57 \end{bmatrix} = \begin{bmatrix} 0.63 \\ 0.90 \end{bmatrix}$$

Then, we compute the gradient error vector δ_h of the hidden layer (cf. Eq. 2.13)

$$\delta_h = \boldsymbol{W}_{23}\delta_o = \begin{bmatrix} 0.5 \\ 0.6 \end{bmatrix} \begin{bmatrix} 0.57 \end{bmatrix} = \begin{bmatrix} 0.5 \times 0.57 \\ 0.6 \times 0.57 \end{bmatrix} = \begin{bmatrix} 0.285 \\ 0.342 \end{bmatrix}$$

and update the weight matrix $\boldsymbol{W}_{12}$ (cf. Eq. 2.13):

$$\boldsymbol{W}'_{12} = \begin{bmatrix} 0.1 + 0.7 \cdot 0.285 & 0.3 + 0.7 \cdot 0.342 \\ 0.2 + 0.8 \cdot 0.285 & 0.4 + 0.8 \cdot 0.342 \end{bmatrix} = \begin{bmatrix} 0.30 & 0.54 \\ 0.43 & 0.67 \end{bmatrix}$$

[a] Note that $(\boldsymbol{x}^T\boldsymbol{W})^T = \boldsymbol{W}^T\boldsymbol{x}$

2.1.1.2 CNN: Convolutional Neural Networks

We employ a Convolutional Neural Network (CNN) in Sec. 3.2 and thus, give a short outline of how Convolutional Neural Networks (CNNs) differ from NNs. CNNs are a particular form of NNs that can learn feature masks rather than just sticking to individual weights (Albawi et al., 2017; LeCun et al., 1989, 1998). The advantage of CNNs over standard NNs is that the input is not flattened but remains in the original dimension. Therefore, the network does not lose shape information that occurs when flattening (Ex. 2.3).

Example 2.3: Flattening

Let us take a look at an example. Let $\boldsymbol{x}$ be a 4×4 frame:

$$\boldsymbol{x} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \end{pmatrix} \tag{2.14}$$

If we flatten this matrix, we get the following:

$$\boldsymbol{x}_{flat} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} & x_{21} & x_{22} & x_{23} & x_{24} & \dots & x_{44} \end{pmatrix} \tag{2.15}$$

In Ex. 2.3, we see that the input values x_{13} and x_{14} have split up the shape information of the yellow square pixels x_{11}, x_{12}, x_{21}, and x_{22}. Thus, the network must recognize existing shapes through large distances, which get even larger with higher frame dimensions.

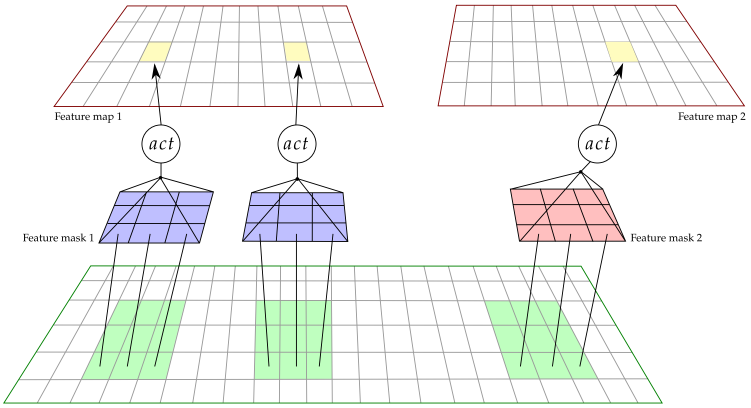


Figure 2.4: Sketch of a CNN. Two 3 × 3 feature masks (*blue mask* and *red mask*) are scanned over the overall frame. At every frame position, the weighted sum (Eq. 2.2) between the *blue mask* / *red mask* and the *green frame section* is calculated and written into the *feature map* as *output*, which is passed to the next neural network layer.

To avoid this problem, CNNs use *feature masks*, each of which learns to recognize specific shapes and objects. The feature masks can be considered scanners that scan the entire frame and find all positions where a particular object or shape is found. The result is then written to a *feature map* and passed to the following neural network layer. In this way, a network can learn feature maps that recognize round shapes while other feature maps recognize square shapes. Fig. 2.4 outlines the concept.

2.1.1.3 Optimization Methods

In order to update NNs, various optimization algorithms can be employed. One such method, commonly known as *standard gradient descent*, was presented in a simplified form in Eq. 2.7 to demonstrate the derivation of the Backpropagation algorithm for a single data sample.

When using *standard gradient descent*, the network's weights are updated by considering the average of the gradients of the entire training data set with respect to the parameters. Mathematically, the *error* function E_{MSE} over the entire dataset $\mathcal{X}$ with $\left|\mathcal{X}\right| = m$ can be expressed using the MSE as (Goodfellow et al., 2017, p.105) (cf. Eq. 2.7, see Eq. 2.16):

$$E_{MSE} = \frac{1}{2m}\sum_{x_i\in\mathcal{X}}\left(\mathcal{B}(x_i) - \hat{\mathcal{B}}(x_i)\right)^2 \tag{2.16}$$

and the weights $\boldsymbol{W}_t$ are updated as follows (cf. Eq. 2.10, see Eq. 2.17):

$$\boldsymbol{W}_t \longleftarrow \boldsymbol{W}_{t-1} - \nabla_{\boldsymbol{W}} E_{MSE} \tag{2.17}$$

When training a neural network, it is uncommon to solely use a single sample or the entire dataset at once. Instead, a minibatch $\mathcal{X}' \subseteq \mathcal{X}$ of the data is often used to compute E_{MSE}. Optimization methods dealing with batches of data are generally referred to as *stochastic* optimization methods (Goodfellow et al., 2017, pp.271-272). A simple variant of *standard gradient descent*, which uses the same update rule as in Eq. 2.17, is called Stochastic Gradient Descent (SGD). Other optimization methods that we apply within this thesis encompass Adaptive Moment Estimation (ADAM) and Adaptive Moment Estimation with Infinity Norm (ADAMAX).

The ADAM optimizer combines two gradient descent techniques, namely *Momentum* and Root Mean Square Propagation (RMSPROP). Momentum is a technique to accelerate learning by computing a moving average of previous gradients, facilitating weight updates based on past gradients in the respective direction (Polyak, 1964; cited in Goodfellow et al., 2017, pp.288-289).

Using Momentum, a velocity matrix $\boldsymbol{V}_t$ is calculated and used to update the weights $\boldsymbol{W}_t$ (see. Eq. 2.18, Eq. 2.19):

$$\boldsymbol{V}_t \longleftarrow \beta_0 \boldsymbol{V}_{t-1} - \alpha \nabla_{\boldsymbol{W}} E_{MSE} \tag{2.18}$$

$$\boldsymbol{W}_t \longleftarrow \boldsymbol{W}_{t-1} + \boldsymbol{V}_t \tag{2.19}$$

where $\beta_0 \in [0, 1]$ is a hyper-parameter defining the decay speed of previous gradients, and $\alpha \in [0, 1]$ is the learning rate.

RMSPROP (Goodfellow et al., 2017, pp.299-300) uses the root mean square error of past gradients to normalize the learning rates. The RMSPROP update rule is defined as follows (see. Eq. 2.20, Eq. 2.21):

$$\boldsymbol{R}_t \longleftarrow \beta_2 \boldsymbol{R}_{t-1} + (1 - \beta_2) \nabla_{\boldsymbol{W}} E_{MSE} \odot \nabla_{\boldsymbol{W}} E_{MSE} \tag{2.20}$$

$$\boldsymbol{W}_t \longleftarrow \boldsymbol{W}_{t-1} - \frac{\alpha}{\sqrt{\delta + \boldsymbol{R}_t}} \odot \nabla_{\boldsymbol{W}} E_{MSE} \tag{2.21}$$

Here, $\beta_2 \in [0, 1]$ is another hyper-parameter controlling the sliding window of the moving gradient average, and $\delta = 10^{-6}$ is a constant to stabilize small number division.

ADAM combines momentum (cf. Eq. 2.18) and RMSPROP (cf. Eq. 2.20) using the following update rule (Goodfellow et al., 2017; Kingma & Ba, 2014) (see Eq. 2.22):

$$\boldsymbol{W}_t \longleftarrow \boldsymbol{W}_{t-1} - \alpha \cdot \frac{\hat{\boldsymbol{V}}_t}{\sqrt{\hat{\boldsymbol{R}}_t} + \delta} \odot \nabla_{\boldsymbol{W}} E_{MSE} \tag{2.22}$$

with slight modifications of momentum (cf. Eq. 2.18, see Eq. 2.23)

$$\tilde{\boldsymbol{V}}_t \longleftarrow \beta_1 \tilde{\boldsymbol{V}}_{t-1} - (1 - \beta_1) \nabla_{\boldsymbol{W}} E_{MSE} \tag{2.23}$$

and the application of a bias correction using decaying rates β_1^t and β_2^t depending on the time step t (see Eq. 2.24, Eq. 2.25):

$$\hat{\boldsymbol{V}}_t \longleftarrow \frac{\tilde{\boldsymbol{V}}_t}{1 - \beta_1^t} \tag{2.24}$$

$$\hat{\boldsymbol{R}}_t \longleftarrow \frac{\boldsymbol{R}_t}{1 - \beta_2^t} \tag{2.25}$$

ADAMAX is yet another variant of ADAM replacing the RMSPROP term in the denominator of Eq. 2.22 with the L-infinity norm of the gradients (maximum absolute value) of past gradients (Kingma & Ba, 2014). This simplification avoids the use of the square root, making it computationally more efficient (see Eq. 2.26, Eq. 2.27):

$$\boldsymbol{U}_t = \max\left(\beta_2 \boldsymbol{U}_{t-1}, \left|\nabla_{\boldsymbol{W}} E_{MSE}\right|\right) \tag{2.26}$$

$$\boldsymbol{W}_t \longleftarrow \boldsymbol{W}_{t-1} - \alpha \cdot \frac{\hat{\boldsymbol{V}}_t}{\boldsymbol{U}_t} \tag{2.27}$$

Note that E_{MSE} is one of many existing error functions, and any other loss function can be used with the optimization algorithms. Within this thesis, we only use E_{MSE} loss.

2.1.1.4 Performance Measures

There are several performance measures to evaluate the network's training.

A *confusion matrix* summarizes correct (*true*) and incorrect (*false*) predictions for each class (see Tab. 2.1).

Table 2.1: Confusion matrix of two classes, and the amount of correctly classified samples ($true_{class1}$, $true_{class2}$), and miss-classified samples ($false_{class1}$, $false_{class2}$).

		Classified as	
		class 1	class 2
Class	class 1	$true_{class1}$	$false_{class2}$
	class 2	$false_{class1}$	$true_{class2}$

Accuracy measures the percentage of correct predictions of all samples:

$$accuracy = \frac{true_{class1} + true_{class2}}{\#class1 + \#class2} \tag{2.28}$$

This, however, is problematic if the number of samples across classes is not balanced. In this case, the average accuracy can be used instead:

$$average\ accuracy = \frac{1}{2}\left(\frac{true_{class1}}{\#class1} + \frac{true_{class2}}{\#class2}\right) \tag{2.29}$$

Precision defines the percentage of how many as class A classified samples are correct (commonly known as *positive prediction*, Eq. 2.30), computed by:

$$precision = \frac{true_{class1}}{true_{class1} + false_{class1}} \tag{2.30}$$

In contrast, *recall* (aka. *sensitivity*) defines the percentage of how many samples of a certain class A have been correctly classified as class A (*true-positives*, Eq. 2.31):

$$recall = \frac{true_{class1}}{true_{class1} + false_{class2}} \tag{2.31}$$

Note the difference in the denominator. When calculating the precision, we are interested in all samples that **were classified** as *class 1*; when calculating the recall, we are interested in all samples that **are of type** *class 1*.

Finally, the F1-score is the weighted value of $precsion$ and $recall$, calculated with weight factor α as follows (Eq. 2.32).

$$\text{F1-score} = \frac{2}{\alpha \cdot precision^{-1} + (1-\alpha) \cdot recall^{-1}} \tag{2.32}$$

2.1.1.5 Overfitting and Underfitting

As outlined in Sec. 1.3.1, due to the subjectivity of data in this work, there is a risk of *overfitting* and *underfitting*, requiring a trade-off between *bias* and *variance* (Fortmann-Roe, 2012; Geman et al., 1992). See Tab. 2.2 below.

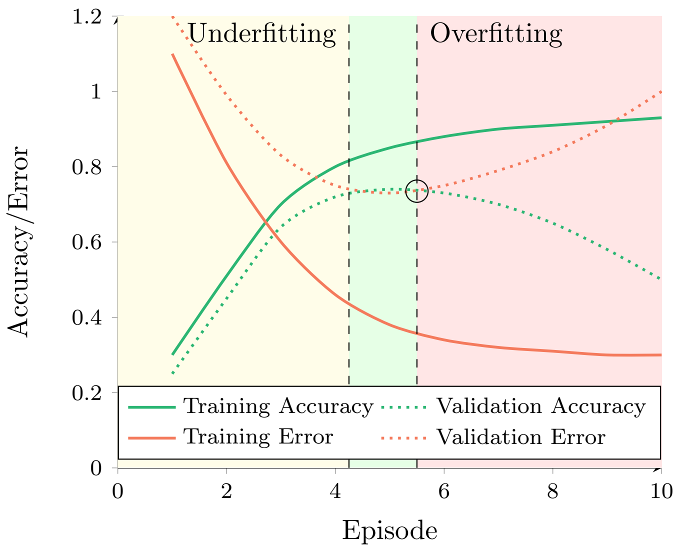


Figure 2.5: *Underfitting* with high error and low accuracies on **training** and **test** data, and *overfitting* with increasing accuracy (decreasing error) on the **training** data but decreasing accuracy (increasing error) on the **test** data.

Overfitting is caused by high variance (Bilmes, 2020). It is indicated by 1) the model having a low error (decreasing) on the **training data** but a high error (increasing) on the **test data**, and 2) a high accuracy (increasing) on the **training data** but a low accuracy (decreasing) on the **test data** (see Fig. 2.5).

Underfitting is caused by high bias (Bilmes, 2020), e.g., due to the subjectivity of data, the network cannot learn any patterns. This is the case if both the network's **training** and **validation** exhibit low accuracy with high error at the same time (see Fig. 2.5).

Table 2.2: Relationship between *overfitting/underfitting*, *accuracy/error*, and *bias/variance*. High bias means the network cannot capture any patterns, and high variance means the network does not generalize. Thus low bias (the network captures existing patterns) and low variance (the network generalizes) are favorable.

Model	Training		Test		*Bias*	*Variance*
	Error	*Accuracy*	*Error*	*Accuracy*		
Overfitting	*Low*	*High*	*High*	*Low*	*Low*	*High*
Underfitting	*High*	*Low*	*High*	*Low*	*High*	*Low*
Balanced	*Low*	*High*	*Low*	*High*	*Low*	*Low*

2.1.1.6 Cronbach's Alpha

In Ch. 3, we annotate videos concerning perceived persuasiveness to train a single neural network employing data of three annotators (Sec. 3.2). However, using such data is inherently challenging for training a neural network, given the subjectivity of the annotations and the absence of ground truth. To ensure the reliability of the data, we, therefore, calculate *Cronbach's alpha* α_c (Cronbach, 1951), which is a statistical measure for determining the average agreement on n different scales (here: annotations). It is calculated by (Eq. 2.33):

$$\alpha_c = \frac{n \cdot \bar{\rho}}{1 + (n-1) \cdot \bar{\rho}} \tag{2.33}$$

where $\bar{\rho}$ defines the average *Pearson correlation coefficient* of all scales. The *Pearson correlation coefficient* (Howell, 2012, p.252 et seq.) between two variables U and V is thereby calculated by (Eq. 2.34):

$$\rho_{uv} = \frac{\frac{1}{n}\sum_i \left(u_i - \bar{u}\right)\left(v_i - \bar{v}\right)}{\sigma_U \, \sigma_V} \tag{2.34}$$

where σ_U is the standard deviation of U.

Example 2.4: Cronbach's Alpha

Let us consider a brief example using the annotations of Sec. 3.2. The following table shows three scales (here: annotators) and the respective correlations between them.

Table 2.3: Pearson correlation coefficients between annotators. (*) denotes a p-value ≤ 0.001.

	Annotator 1	Annotator 2	Annotator 3
Annotator 1	*1*	0.5523*	0.5559*
Annotator 2		*1*	0.4647*
Annotator 3			*1*

First, we compute the average correlation, that is

$$\bar{\rho} = \frac{0.5523 + 0.5559 + 0.4647}{3} = 0.5243 \tag{2.35}$$

and with it ($n = 3$)

$$\alpha_c = \frac{3 \cdot 0.5243}{1 + 2 \cdot 0.5243} = 0.7678 \tag{2.36}$$

A *Cronbach's alpha* α_c is acceptable if > 0.7 (George & Mallery, 2002), which is the case here.

2.1.2 Explainable Artificial Intelligence

ℹ This section was previously published by the author in a similar form in peer-reviewed papers (Weber et al., 2020c, 2023b), *reproduced with permission from Springer Nature*.

XAI is a promising tool for inferring behavioral characteristics of humans, such as persuasiveness, which is a highly subjective task that might include biases.

Earlier works already used XAI on several subjective tasks. For example, Escalante et al. (2017) developed a challenge to test different explainable systems used for first impression analysis in job applications. Weitz et al. (2019) investigated different XAI methods on facial pain and emotion recognition models.

In the context of persuasion and XAI, recent work mainly investigated explainable recommendation systems persuading humans (Donadello et al., 2019; Zhang & Chen, 2020). To the best of our knowledge, this is the first work on explainable systems that investigates *why* a speaker is perceived as persuasive.

Since Artificial Intelligence (AI) systems are becoming increasingly complex, there is an increasing need to increase their explainability. XAI aims at explaining decisions of AI systems to make them more comprehensible for humans. More specifically, such methods try to explain why a system has made certain decisions to, for instance, increase trust towards the system (Weitz et al., 2019).

This is especially important for sensitive tasks that require a high understanding of the system's decisions, like self-driving cars (Schraagen et al., 2020). Especially newer AI methods, such as neural networks, have become better and better in their task accuracy, which, however, came with the caveat that they are challenging to understand by humans. This is mainly due to the vast amount of parameters that the network learns. For instance, the Visual-Geometry-Group-19 network consists of over one hundred million parameters (Simonyan & Zisserman, 2014).

Therefore, people developed methods for explanations. Explanation methods can be evaluated in two different ways: 1) *interpretability* and 2) *completeness*. As of Gilpin et al. (2018), the objective of completeness is "*to describe the operation of a system in an accurate way*", while the objective of interpretability is "*to describe the internals of a system in a way that is understandable to humans*"

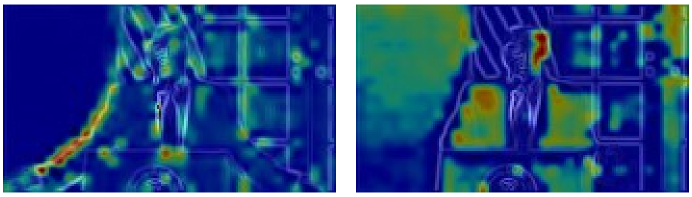


Figure 2.6: Two images that were classified as *very convincing* (left) and *neutral* (right). Using XAI, we can verify what the network focused on: The speaker's contours, especially right arm (left image), as well as the speaker's head (right image).

In the context of AI, the primary focus of explanations lies on *why*-questions (Gilpin et al., 2018). For instance, when classifying an image, why-explanations can give insights about the areas that the network looked at. Such methods can also be used to verify whether a network has learned what was to be learned (Lapuschkin et al., 2019). In our work, the NN should focus on the person. In Fig. 2.6, we can see two images of a speaker. On the left, the focus is almost completely on the speaker; on the right, the focus is also on the background, however, the main focus (indicated by red), is on the speaker's head.

Gilpin et al. (2018) classified common XAI methods into three categories, that are *processing*, *representation*, and *explanation producing*. As the term suggests, *processing* means to explain how a network deals with data, *representation* what the data represents within the network. E.g., an image classifier may recognize specific shapes in certain neurons. *Explanation producing* is a technique to create models that explain themselves. This can be used to create an input image reversely that activates a specific output neuron the most (Nguyen et al., 2016).

A different kind of explanation is *counterfactual* explanation. It is the opposite of reverse imaging, which creates an image different from what the network has seen. For instance, a network that can classify cats and dogs might output a dog as the opposite of a cat. Counterfactuals as explanations have been applied in various research tasks. Heimerl et al. (2022) applied counterfactuals in a job interview training task to suggest what a user needs to change to increase the appeared engagement. Based on a multi-modal analysis, the user engagement was tracked, and features were extracted and converted to counterfactual explanations showing how the user can increase their engagement appearance.

Molnar (2019) categorized interpretability methods into: 1) *intrinsic* vs. *post-hoc*,

2) *model-specific* vs. *model-agnostic*, and 3) *local* vs. *global* explanations. *Intrinsic* methods aim to 1) use machine learning models that are inherently explainable or 2) replace existing models with ones that have intrinsic explainability, whereas *post-hoc* methods analyze a model after its training. *Model-specific* techniques are tailored to specific machine learning model classes, while *model-agnostic* methods are versatile and applicable to any machine learning model. The *local* vs. *global* distinction lies in explaining a single prediction versus the overall model. This work focuses on *local* and *post-hoc* explanations. Since we use CNNs, we focus on both *model-specific* (Grad-CAM, Ras et al., 2022) and *model-agnostic* (LRP, Ras et al., 2022) methods (see Sec. 2.1.2.1 and Sec. 2.1.2.2).

Sixt et al. (2020) tested different LRP variants and concluded that most LRP variants lose much information about the network's last fully connected layers. Instead, they mainly analyze the convolutional layers at the beginning. Grad-CAM, on the other hand, mainly analyzes the last convolutional layer (Zhou et al., 2016).

Alam et al. (2022), who explored the application of XAI in the analysis of chest radiography images, highlighted that LRP provides more fine-grained explanations, offering insights at a *micro level*. In contrast, Grad-CAM sheds light on the final network's layers, enabling analyses at a *macro level*. The rationale behind this lies in the inherent characteristics of a neural network's initial and last layers. The early layers function as edge detectors primarily focused on recognizing basic features and simple shapes, such as lines. On the other hand, the later layers recognize high-level features and more detailed shapes (LeCun et al., 2015). Given the differences in *micro* vs. *macro* level explanations between LRP and Grad-CAM (Alam et al., 2022) we chose a combination of class discriminatory Grad-CAM saliency maps and fine granular LRP saliency maps to understand better the end and the beginning parts of our model, respectively.

Generating saliency maps is the most common, local, post-hoc explanation method for NNs (Adadi & Berrada, 2018). Saliency maps are heat maps that highlight areas of the input that were relevant to a system's decision. Using such methods, we can draw conclusions if the network has focused on persuasive cues that are prevalent for persuasiveness. One of the first kinds of saliency maps was based on the gradient. Simonyan et al. (2014) used Backpropagation to calculate the gradient with respect to each input unit to measure how much a slight change in this input affects the prediction. Selvaraju et al. (2017) made this approach more class discriminatory (Grad-CAM) by stopping the Backpropagation after the fully connected layers and using the gradient with respect to the output of the last convolutional layer.

2.1.2.1 Grad-CAM

Grad-CAM is a *model-specific* (Ras et al., 2022) XAI technique to explain the last layers of a CNNs at *macro level*. Here, Grad-CAM is a generalization of Class Activation Mapping (CAM) (Zhou et al., 2016). The problem with CAM is that it requires changing the system architecture and retraining the model. More specifically, the feature maps must be replaced with a softmax layer. Using the gradients, this issue was solved by Grad-CAM (Selvaraju et al., 2017). In this thesis, we apply Grad-CAM in Sec. 3.2.3.1 to analyze the last layer of our trained network. Fig. 2.7 outlines the concept of Grad-CAM.

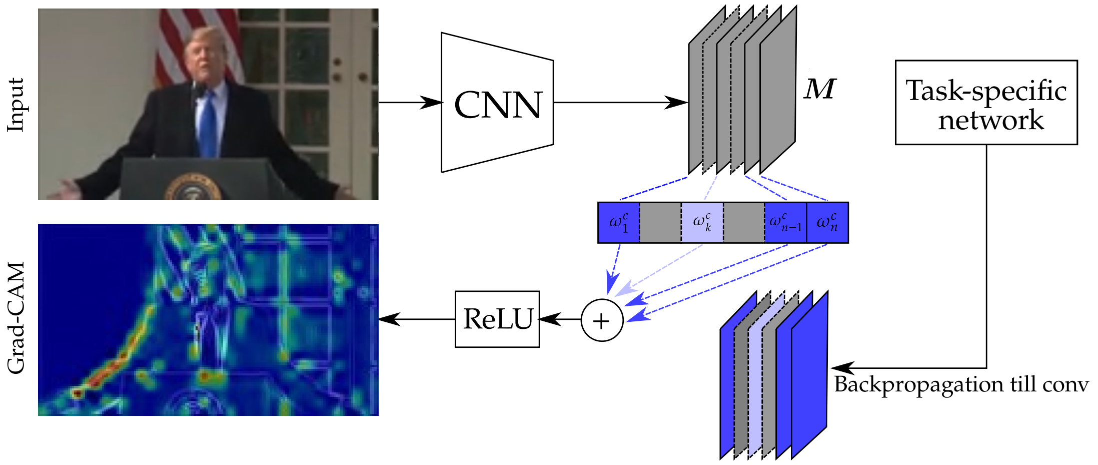


Figure 2.7: Sketch of Grad-CAM using the left image of the example in Fig 2.6 (Figure adapted from Selvaraju et al. (2017)).

To generate a localization map L_c for a specific class c, we first feed-forward the $n \times m$ input through the network. Then, we compute the partial derivative (gradient) of the class-c output $\hat{y}_n^c$ with respect to the last convolutional layer's feature maps $M^k \in \boldsymbol{M}$, i.e.,

$$\frac{\partial \hat{y}_n^c}{\partial M^k} \tag{2.37}$$

Note that M_k is a matrix of activations. Therefore, we iterate over the gradients and calculate the global average, such that we obtain the importance weight α_k^c for the feature matrix M_k:

$$\alpha_k^c = \frac{1}{i \cdot j} \underbrace{\sum_n \sum_m \frac{\partial \hat{y}_n^c}{\partial M_{ij}^k}}_{\omega_k^c} \tag{2.38}$$

Finally, we obtain the localization map L_c for class c by computing the linear combination between the importance weight α_k^c and the feature matrix M_k and applying a Rectifier Linear Unit (ReLU) function:

$$L_c = ReLU \underbrace{\left(\sum_k \alpha_k^c M^k \right)}_{\text{linear combination}} \tag{2.39}$$

2.1.2.2 Layer-wise-Relevance Propagation (LRP)

To analyze further the first convolutional layers of the network and what patterns they learned, we further make use of the *model-agnostic* method LRP (see Sec. 3.2.3.2, some variants can be *model-specific*, Ras et al., 2022).

LRP, introduced by Bach et al. (2015), is a method that assigns a relevance score to each neuron of an NN, indicating the neuron's relevance to a specific prediction. For this, the output of an NN is back-propagated to assign a relevance value R_j to each neuron j. This relevance value defines how much a particular neuron contributed to the input of the following neurons. Fig. 2.8 sketches the concept. Sec. 2.1.1 gives a more detailed description of NNs. Let $\hat{y}_k$ be the activation of the k-th neuron during the forward pass, and let ω_{jk} be the weight that connects neuron j and neuron k. After the forward pass, the relevance propagation starts in the output layer. Here, the activation responsible for the prediction gets assigned its activation as relevance, and every other neuron gets set to zero. That is

$$R_k = \begin{cases} \hat{y}_k & \text{if } k = \arg\max \left\{ \hat{y}_k \right\} \\ 0 & \text{if not.} \end{cases} \tag{2.40}$$

The relevance gets propagated to each preceding layer according to different rules (see Fig. 2.8 and Tab. 2.4). In our experiments we use the z^+- or $\alpha 1 \beta 0$-rule, which is a special case of the LRP-$\alpha\beta$ rule (Bach et al., 2015), defined as:

$$R_j = \sum_k \frac{\left(\hat{y}_j \, \omega_{jk} \right)^+}{\sum_i \left(\hat{y}_i \, \omega_{ik} \right)^+} R_k \tag{2.41}$$

where $(\hat{y}_j\omega_{jk})^+$ is defined as $\max(\hat{y}_j\omega_{jk}, 0)$.

Fig. 2.8 shows an example propagation of the relevance of neuron j_1. Neuron j_1 has two successor neurons, which are k_1 and k_2.

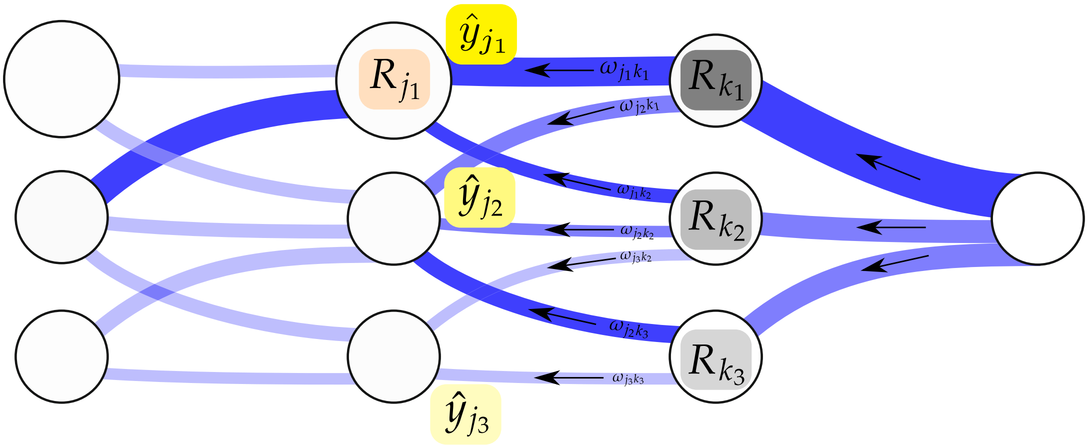


Figure 2.8: Sketch of relevance propagation. The relevance R_{j_1} of neuron j_1 is computed by back-propagating its influence on the successor neurons, which are the neurons k_1 and k_2. (Figure adapted from Montavon et al. (2019, p.4))

The relevance R_{j_1} of neuron j_1 is, assuming all weights are > 0, computed by:

$$\underbrace{\underbrace{\frac{\hat{y}_{j_1}\omega_{j_1k_1}}{\hat{y}_{j_1}\omega_{j_1k_1}+\hat{y}_{j_2}\omega_{j_2k_1}}}_{\text{Note the denominator: Neuron } k_1 \text{ gets inputs from } j_1 \text{ and } j_2\text{, but not } j_3.} \cdot R_{k_1} + \frac{\hat{y}_{j_1}\omega_{j_1k_2}}{\hat{y}_{j_1}\omega_{j_1k_2}+\hat{y}_{j_2}\omega_{j_2k_2}+\hat{y}_{j_3}\omega_{j_3k_2}} \cdot R_{k_2}}_{\text{Note the summands: Neuron } j_1 \text{ only influences } k_1 \text{ and } k_2\text{, but not } k_3.} \tag{2.42}$$

There are two essential steps we notice: 1) We only consider neurons k_i (here: k_1 and k_2) that receive the output from neuron j_1 (summands). 2) within each neuron k_i, we check the percentage influence of neuron j_1 with respects to all other input neurons j_1, j_2, and j_3 (see denominator).

Table 2.4: Common LRP methods. (Table adapted from Montavon et al. (2019)). $(\cdot)^+$ and $(\cdot)^-$ are $max(0, \cdot)$ and $min(0, \cdot)$.

Name	Formula	Layers
LRP-0 [3]	$R_j = \sum_k \frac{\hat{y}_j \omega_{jk}}{\sum_{0,j} \hat{y}_j \omega_{jk}} R_k$	Upper
LRP-ϵ [3]	$R_j = \sum_k \frac{\hat{y}_j \omega_{jk}}{\epsilon + \sum_{0,j} \hat{y}_j \omega_{jk}} R_k$	Middle
arg max [4]	$R_j = \sum_{\arg\max\{\hat{y}_j \omega_{jk}\}} R_k$	Mid/Low
LRP-γ [5]	$R_j = \sum_k \frac{\hat{y}_j (\omega_{jk} + \gamma \omega^+_{jk})}{\sum_{0,j} \hat{y}_j (\omega_{jk} + \gamma \omega^+_{jk})} R_k$	Lower
LRP-$\alpha\beta$ [3]	$R_j = \sum_k (\alpha \frac{(\hat{y}_j \omega_{jk})^+}{\sum_{0,j} (\hat{y}_j \omega_{jk})^+} - \beta \frac{(\hat{y}_j \omega_{jk})^-}{\sum_{0,j} (\hat{y}_j \omega_{jk})^-}) R_k$	Lower
flat [6]	$R_j = \sum_k \frac{1}{\sum_j 1} R_k$	Lower
ω^2-rule [5]	$R_j = \sum_k \frac{\omega^2_{jk}}{\sum_j \omega^2_{jk}} R_k$	First
$z^{\mathcal{B}}$-Rule [5]	$R_j = \sum_k \frac{\hat{y}_j \omega_{jk} - l_j \omega^+_{jk} - h_j \omega^-_{jk}}{\sum_j \hat{y}_j \omega_{jk} - l_j \omega^+_{jk} - h_j \omega^-_{jk}} R_k$	

Another take on saliency maps comes with occlusion or perturbation-based visualizations. Zeiler and Fergus (2014) zero out windows inside the input and measure how much the prediction changes. The more the output changes, the more relevant this window is for this particular prediction.

Greydanus et al. (2018) uses a similar approach but perturbs the windows with noise to see how much the introduced uncertainty affects the prediction.

The LIME framework from Ribeiro et al. (2016) first separates the input picture into super-pixels by a segmentation algorithm. Afterward, a more interpretable model is trained to estimate which super-pixels are the most relevant for a given decision. One advantage of those methods is that they are not dependent on the model's structure, but this comes with the limitation that they are less precise than some model-specific methods.

[3] Bach et al. (2015)

[4] Huber et al. (2019)

[5] Montavon et al. (2017)

[6] Lapuschkin et al. (2019)

2.1.3 Argumentation - Theory, and Notation

ℹ This section provides a brief overview of computational argumentation and the formalism used in this thesis. Part of this section was previously published by the author in a similar form in peer-reviewed papers (Rach et al., 2021; Weber et al., 2020b, 2020a).

To allow seamless communication between the human user and the conversational agent, we got inspired by so-called *dialogue games* (Prakken, 2000, 2005). Originally designed for agent-agent interactions within a dispute task, dialogue games incorporate elements such as *speech acts*, a sequence of *moves*, *protocol*, *commitment*, *locution*, and *termination* rules. Unlike traditional dialogue systems that rely on *templates* and *slot filling*, dialogue games offer a more flexible structure, not constrained by predetermined *templates* and *slot filling* requirements.

Dialogue systems can be classified as *task-oriented*, *conversational*, or *question-answering* (Deriu et al., 2021). While traditional task-oriented dialogue systems have a restricted domain, are highly structured, and are mostly short, conversational dialogues allow for more flexibility, are domain-independent, and are mostly longer than task-oriented dialogue systems (Deriu et al., 2021).

Further, dialogues can be classified by type of goals, that is *persuasion*, *negotiation*, *deliberation*, *information-seeking*, *inquiry*, *eristic* (Walton, 2005, p.183) and *discovery* (Walton, 2010), which often refer to *conversational* dialogues as they require a more complex communication structure.

Templates are often used for *task-oriented* dialogues (Deriu et al., 2021) and define conditions under which a user utterance is appropriate. This is the case if the utterance fits into a slot of the template. Slots for a *restaurant-finder* may be *area*, *food* or *price range* (Novikova et al., 2017).

To fill slots, *dialogue acts* or *speech acts* are used. *Dialogue acts* are functions with which the dialogue context (slots) can be filled. *Speech acts* are simpler forms of *dialogue acts*, i.e., *dialogue acts* can require other actions before being satisfied. Generally, *dialogue* and *speech acts* allow communication between user and system to express their goals and intents.

The intent is the expressed user's intention or plan via speech or text in a natural form linked and mapped to one of the existing *speech acts* (P. R. Cohen & Perrault, 1979). The set of speech acts that is available to the conversational agent and the user is called the communication language L_c (see Sec 4.1.1 for employed speech acts). Within the context of this work, there are speech acts defined for the user to ask for a supporting argument (why_{pro}) or for an attacking argument

(why_{con}), and for the conversational agent to present an argument ($argue$).

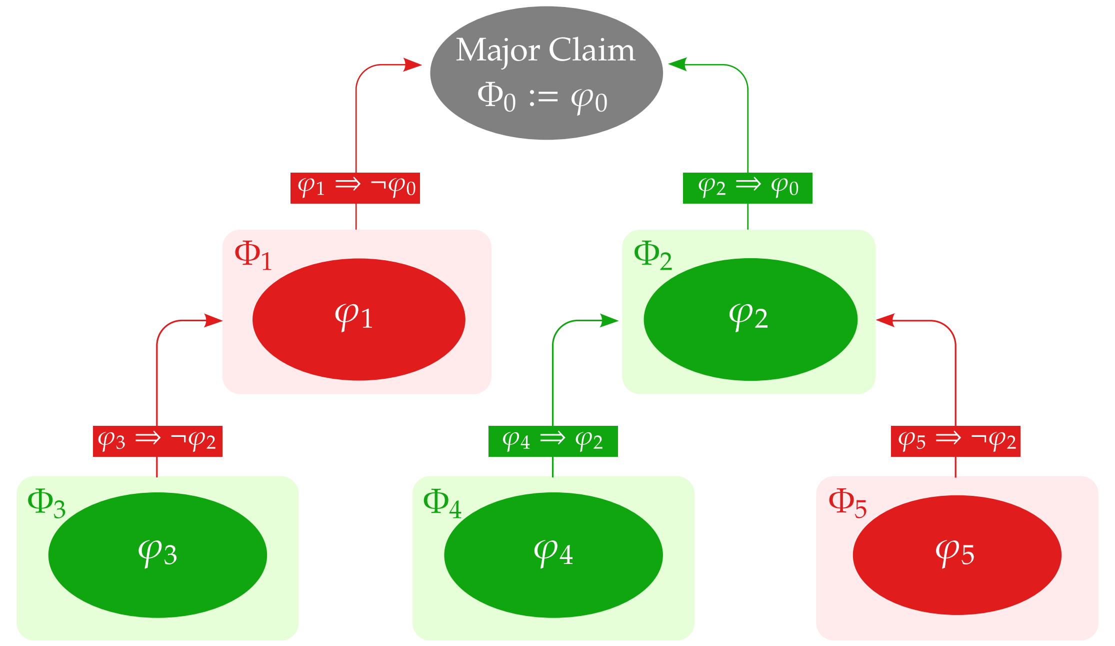

Major Claim
$\Phi_0 := \varphi_0$
$\varphi_1 \Rightarrow \neg\varphi_0$
$\varphi_2 \Rightarrow \varphi_0$
Φ_1
φ_1
Φ_2
φ_2
$\varphi_3 \Rightarrow \neg\varphi_2$
$\varphi_4 \Rightarrow \varphi_2$
$\varphi_5 \Rightarrow \neg\varphi_2$
Φ_3
φ_3
Φ_4
φ_4
Φ_5
φ_5

Figure 2.9: Sketch of the argument graph consisting of pro and con arguments (nodes, also called components) towards the *Major Claim*, defined as relations support and attack (edges) between components. **Note**: An argument can have multiple supporting/attacking arguments.

To be able to present arguments, the conversational agent requires a *knowledge base* containing all available arguments about the topic and needs to know, for each argument, which arguments are in favor (pro) or against (con) it.

Therefore, our system utilizes a *knowledge base* based on an argument structure organized as an acyclic-directed graph $G = (L_t, \longrightarrow)$. The graph **nodes** are defined by a set of **argument components** $L_t = \left\{\varphi_0, \ldots \varphi_n\right\}$. Each argument component has a unique natural language realization that it is referred to, e.g., φ_1 = (*the air conditioner was not working*). The graph **edges** are defined as directed logical relation $\longrightarrow := \left\{\left(\varphi_i, \varphi_j, \Rightarrow\right) \middle| \varphi_i, \varphi_j \in L_t\right\}$ between **nodes** with $\Rightarrow \in$ $\left\{\text{supporting}(+), \text{attacking}(-)\right\}$ the directed relation (see Fig. 2.9).

If a component $\varphi_i \in L_t$ has a logical relation towards a component $\varphi_j \in L_t$, we say that φ_j is the target (of φ_i) and each component (apart from the root node φ_0) has exactly one target. The left-hand side of the logical operator $\Rightarrow$ is thereby called *evidence*, while the right-hand side is called *conclusion* (Rach et al., 2021).

The outcome of the *conclusion*, i.e., whether it is *true* or *false*, is thereby defined by the *relation*. For instance, φ_1 (*the air conditioner was not working*) $\Rightarrow$ φ_2 (*the rooms were bad*). Since φ_1 is a supporting argument component, φ_2 is *true*.

Based on the argument components L_t, and the relation $\longrightarrow$, we can build two types of arguments $\Phi_i \in Args$, that are supporting and attacking (Stab and Gurevych (2014), Fig. 2.9, see Def. 2.2):

Definition 2.2: Notion of Supporting and Attacking Arguments in G

Let $\varphi_i, \varphi_j \in L_t$ be argument components, and $\longrightarrow$ be a relation between them, then an attacking argument is defined as:

$$\Phi_i = \varphi_i \Rightarrow \neg \varphi_j \tag{2.43}$$

while a supporting argument is defined as:

$$\Phi_i = \varphi_i \Rightarrow \varphi_j \tag{2.44}$$

obtaining a set of arguments under G, $Args = \left\{ \Phi_k, \Phi_l, \Phi_m, \Phi_n \Phi_o, \ldots \right\}$ with Φ_0 as root, defined as φ_0 for the sake of simplicity.

A set of arguments under a sub-graph of $G' \subseteq G$ with root $\Phi_i \in Args$ is denoted as $Args(\Phi_i) \subseteq Args$. To refer directly to the target argument Φ_j of argument Φ_i, we can express it as $\Phi_{i \Rightarrow j}$, indicating that Φ_i is a supporting argument of Φ_j. Analogously, we can represent Φ_i as an attacking argument of Φ_j by $\Phi_{i \Rightarrow \neg j}$. If the relation is either of them, we write $\Phi_{i \Rightarrow \_ j}$ with $\_ : attack \oplus support$ [7].

In certain contexts, it is necessary to talk about the underlying components φ_i and φ_j of an argument Φ_i and the relation between them rather than the overall argument. To express that, we use an analogous notion:

- Component with attacking relation: $\varphi_{i \Rightarrow \neg j}$
- Component with supporting relation: $\varphi_{i \Rightarrow j}$
- Component with either relation: $\varphi_{i \Rightarrow \_ j}$

The root argument Φ_0 is called *Major Claim*, while arguments directly attacking or supporting the *Major Claim* are called *Claims*. Any other argument $\Phi_{i \Rightarrow \_ j}$ with $j \neq 0$ is called *Premise* (Stab & Gurevych, 2014).

In argumentation theory, relations are allowed from *Claims* to the *Major Claim*, *Premises* to *Claims*, and *Premises* to *Premises*. Within the scope of this work, there

[7] Note: $\oplus$ defines the *XOR* operator.

are no relations from *Premises* to the *Major Claim*, i.e., when we speak about *Claims*, we always mean arguments targeting the *Major Claim*.

Based on the relation $\longrightarrow$, each argument $\Phi_i \in Args$ refers to one of the two existing *stances* $\in \{+\,(\text{pro}), -\,(\text{con})\}$ of the topic. The stance of Φ_0 is considered as $+$ throughout this work. The other arguments' stances are computed considering the arguments' relations, i.e., the stance of supporting arguments is always the same as its target's. In contrast, the stance of attacking arguments is the opposite of the target's.

Summarized, the employed dialog framework is defined as triple $((L_t, \longrightarrow), L_c,$ *protocol rules*) consisting of the argument graph $G = (L_t, \longrightarrow)$, and the communication language L_c available to the conversational agent and the user. The *protocol rules* define what *speech acts* are allowed, what arguments are valid, and other communication rules, such as turn-taking. For instance, an argument is only valid if its conclusion has been presented to the user.

2.1.4 Reinforcement Learning (RL)

ℹ This section was previously published by the author in a similar form in Weber (2017) and Weber et al. (2018a).

In Ch. 6, we present an approach to adapt the verbal style of the agent when performing interventions. This adaptation is based on RL and *linear function approximation* using a *Fourier basis* transformation. Therefore, we give a brief overview of RL, linear *function approximation*, and the *Fourier basis*.

2.1.4.1 Concept

The following section is based on Sutton and Barto (2018). Contrary to NNs, RL is an unsupervised machine learning method based on the trial-and-error method. A system or agent aims to learn the correct behavior to optimally solve a task by experimenting with available actions to solve a specific task. No expert guides the agent on the correct order of actions using unsupervised learning. Instead, the agent learns what is right or wrong by receiving positive feedback for goal-directed actions and negative feedback for non-goal-directed ones. The agent does not know in advance which actions are best; it only learns from the feedback.

To do so, the agent first observes its environment. The environment contains everything surrounding the agent, abstracted into a simplified form called a *state*. Then, the agent chooses an action based on a strategy π and executes it. An *action* manipulates or changes the environment and thus the state. After execution, the agent receives a *reward* that it uses for learning the correct behavior. More specifically, the agent learns for every state-action pair how good the executed action is to solve the task (see Fig. 2.10).

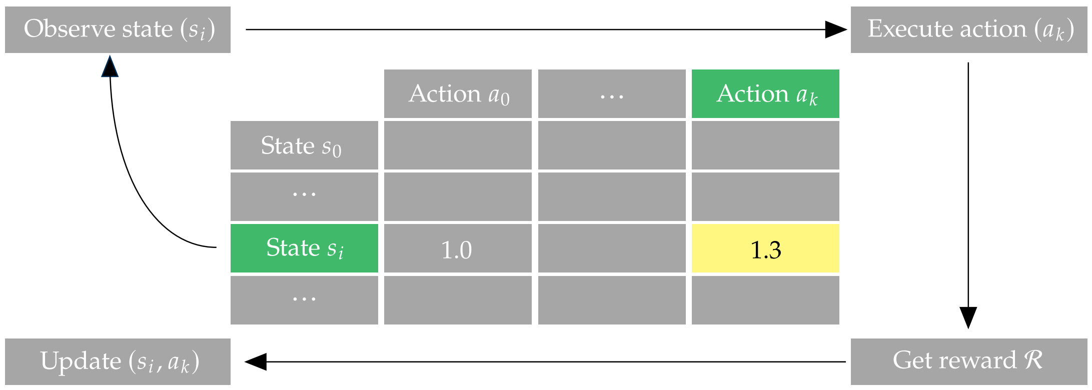


Figure 2.10: Learning process of a self-learning agent using RL

RL problems are defined as Markov Decision Processes (MDPs). MDPs only require the *current environment's state* for decision-making, and are defined as seen in Def. 2.3 (Sutton & Barto, 2018, p. 47-71), consisting of states $s_i \in \mathcal{S}$, actions $a_k \in \mathcal{A}$, a transition function $\mathcal{T}$ between states, and the reward function $\mathcal{R}$.

Definition 2.3: Markov Decision Process (MDP)

An (in-)finite MDP is defined as a 4-tuple:

$$M = < \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R} > \tag{2.45}$$

$\mathcal{S} = \{s_0, \ldots, s_n\}$ represents the infinite state space, and $\mathcal{A} = \{a_0, \ldots, a_m\}$ denotes the agent's action space. $\mathcal{T}$ indicates the probability of transitioning to state s_j when action a_k is chosen in state s_i, accompanied by the reward $\mathcal{R}$, for all $s_i, s_j \in \mathcal{S}$ and $a_k \in \mathcal{A}$:

- $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0,1] := \mathcal{P}\left(s_j | s_i, a_k\right)$
- $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$

A state s at time t is denoted as s_t, and an end state is represented by s_T. To indicate a reward at time step t, we can also write $\mathcal{R}_t$.

The reward $\mathcal{R}$ given for executing a single action a only indicates short-term success. To achieve long-term success (and thus solve the problem most efficiently), the agent accumulates rewards over time, prioritizing actions with higher accumulated rewards; a strategy π known as exploitation or *greedy*.

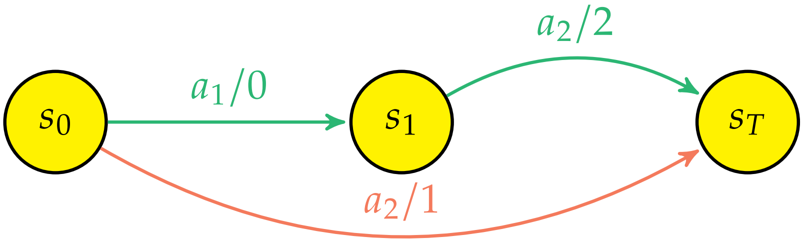
Example 2.5: Short vs. Long-Term Rewards: Stanford-Marshmallow test

Consider the *Stanford-Marshmallow test* (Mischel & Ebbesen, 1970) as an example of short- (immediate) and long-term (accumulated) rewards:

> In this test, children could decide whether to receive one small reward immediately (such as a marshmallow), or they could wait for a short period (typically 15 minutes) to receive a larger reward (like two marshmallows). This test aimed to measure children's ability to delay gratification and resist the temptation of immediate rewards in favor of greater future rewards.

In RL, we can think of this scenario as an agent making decisions to maximize cumulative rewards over time. The immediate reward is analogous to the short-term reward, while the larger reward obtained by waiting corresponds to the long-term reward.

See the figure below, which illustrates two potential sequences of actions (a_1 = wait and a_2 = get) starting from an initial state s_0. In this representation, action a_1 signifies waiting without receiving a marshmallow, while action a_2 entails obtaining the marshmallow(s). The transitions between states indicate the action a_k taken and the subsequent reward obtained (a_k/reward).



It is easy to verify that opting for action a_1 in state s_0 (green path) yields a higher cumulative, long-term reward (**2** > **1**), despite action a_2 offering a higher immediate, short-term reward (**1** > **0**).

The agent usually prefers state-action pairs (s, a) with the highest accumulated rewards. Since the agent cannot know for certain if the currently executed best action is optimal in the long-term, it occasionally needs to choose a different, seemingly less favorable action with a low probability ϵ. This behavior is referred to as exploration. Striking a balance between exploitation (favoring known high-reward actions) and exploration (trying potentially less rewarding actions) is therefore necessary. This strategy π is known as *ϵ-greedy* with ϵ representing the exploration probability (Sutton & Barto, 2018, pp. 26-27)

Using the given rewards $\mathcal{R}_{t+1}$ at time step t, an accumulated value $Q(s, a)$, referred to as *q-value*, can be computed for every state-action pair (s, a) (Sutton & Barto, 2018, pp. 58–67)). The value describes the overall expected accumulated reward when choosing an action $a \in \mathcal{A}$ in state s. Q^* defines the optimal *q-value* function that solves the task most efficiently and accurately (see Def. 2.4).

Definition 2.4: Optimal Q-Value Function

For all $s_t \in \mathcal{S}$ and for all $a_t \in \mathcal{A}$, the optimal *q-value* function Q^* is defined as

$$Q^* : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R} \tag{2.46}$$

It is calculated as the sum of all discounted rewards:

$$Q^*(s_t, a_t) := \sum_{k=0}^{\infty} \gamma^k \mathcal{R}(s_{t+k}, a_{t+k}, s_{t+k+1}) \tag{2.47}$$

with γ the discount factor with $0 \le \gamma < 1$ to ensure convergence.

To compute $Q^*(s_t, a_t)$ in Eq. 2.47, all future rewards $\mathcal{R}_{t+1}, \dots, \mathcal{R}_T$ are needed. Since we want the agent to learn during interaction, we use *Q-Learning* to $Q_\pi(s_t, a_t)$ incrementally based on the strategy π (Watkins, 1989; Watkins & Dayan, 1992):

$$Q_\pi(s_t, a_t) = \mathcal{R}(s_t, a_t, s_{t+1}) + \gamma \max_{a_{t+1}} Q_\pi(s_{t+1}, a_{t+1}) \tag{2.48}$$

until

$$Q_\pi(s_t, a_t) \approx Q^*(s_t, a_t) \tag{2.49}$$

Algorithm 1 shows the full *Q-Learning* algorithm (Sutton & Barto, 2018). The parameter $\alpha \in [0, 1]$ defines the learning rate, i.e., the step size of the update rule.

```
Algorithm 1: Q-Learning
  Data: $Q_\pi(s, a) = 0 \; \forall s \in \mathcal{S} \; \forall a \in \mathcal{A}$, Initial state $s$
  foreach steps $t = 0, 1, 2, \dots$ do
      1. Select action $a$ according to strategy $\pi$
         $a \leftarrow \begin{cases} \arg\max_{a'} (Q_\pi(s, a')) & \text{with probability } 1 - \epsilon \\ \text{uniform random action } a' \in \mathcal{A} & \text{with probability } \epsilon \end{cases}$
      2. Apply $a$
      3. Measure next state $s'$ and reward $r = \mathcal{R}(s, a, s')$
      4. Update $Q_\pi(s, a)$
         $Q_\pi(s, a) \leftarrow Q_\pi(s, a) + \alpha \left[ r + \gamma \max_{a'} Q_\pi(s', a') - Q_\pi(s, a) \right]$   (2.50)
      $s \leftarrow s'$
```

2.1.4.2 Linear Function Approximation

In finite MDPs, the state-action value $Q_\pi(s, a)$ for all $s \in \mathcal{S}, a \in \mathcal{A}$ is directly calculated and stored in table form. However, this approach suffers from the "*curse of dimensionality*", which means that the memory and time required for learning increase exponentially when the amount of states increases.

One potential way to address this issue is to discretize the state space; that is, for example, to differentiate a finite set of levels of valence (Gordon et al., 2016), engagement (Ritschel et al., 2017) or in our work, the user's stance and focus, which then allows using table-based algorithms. However, it causes an information loss, as these algorithms cannot generalize knowledge across similar states.

A second potential way is to use algorithms based on function approximation, which do not use a discrete set of states but parameters representing a target function to learn and represent the state space.

As described by Busoniu et al. (2010), a state $s \in \mathcal{S}$ with m features is represented by a vector $\boldsymbol{\phi}(s) \in [0,1]^m$. The concept of linear function approximation aims to find a finite parameter vector $\boldsymbol{\omega}_a \in \mathbb{R}^m$ for all $a \in \mathcal{A}$. This vector is used to approximate the state-action value $Q_\omega(s,a) = Q_\pi(s,a)$ so that $Q_\omega(s,a) \approx Q^*(s,a)$ (Busoniu et al., 2010; Sutton and Barto, 2018; see Def. 2.5).

Definition 2.5: Q-Value Approximation

For each state $s \in \mathcal{S}, a \in \mathcal{A}$, the state-action value $Q_\omega(s,a)$ is calculated by

$$Q_\omega(s,a) := \boldsymbol{\phi}(s) \circ \boldsymbol{\omega}_a = \sum_{i=1}^{m} \phi_i(s)\, \omega_{a,i}\, , \forall s \in \mathcal{S}, a \in \mathcal{A} \tag{2.51}$$

During learning, the objective is to minimize the error between the optimal state-action values $Q^*(s,a)$ and the approximated values $Q_\omega(s,a)$. As for NNs, we compute the mean squared error between approximated q-value $Q_\omega(s,a)$ and target q-value $Q^*(s,a)$ (cf. Eq. 2.7, see Eq. 2.52; Busoniu et al. (2010, p.61)).

$$E = \Big(Q^*(s,a) - Q_\omega(s,a) \Big)^2 \tag{2.52}$$

The gradient of E with respect to the parameter vector $\boldsymbol{\omega}_a$ $\forall s \in \mathcal{S}, a \in \mathcal{A}$ is computed by (Busoniu et al., 2010, p.61):

$$\begin{aligned}
\Delta \boldsymbol{\omega}_a &= -\frac{1}{2} \nabla_\omega E \\
&= -\frac{1}{2} \nabla_\omega \Big(Q^*(s,a) - Q_\omega(s,a) \Big)^2 \\
&= \Big(\underbrace{Q^*(s,a)}_{\mathcal{R}(s,a,s') + \gamma \max_{a'} Q_\omega(s',a')} - \underbrace{Q_\omega(s,a)}_{\boldsymbol{\phi}(s) \circ \boldsymbol{\omega}_a} \Big) \underbrace{\nabla_\omega Q_\omega(s,a)}_{\boldsymbol{\phi}(s)}
\end{aligned} \tag{2.53}$$

Algorithm 2 shows the full *Q-Learning with function approximation* algorithm (Busoniu et al., 2010, p.61).

Algorithm 2: Q-Learning with Function Approximation

```
Data: $Q_\pi(s,a) = 0 \ \forall s \in \mathcal{S} \ \forall a \in \mathcal{A}$, Initial state $s$
foreach steps $t = 0, 1, 2, \ldots$ do
    1. Select action $a$ according to strategy $\pi$
        $$a \leftarrow \begin{cases} \arg\max_{a'} (Q_\omega(s,a')) & \text{with probability } 1-\epsilon \\ \text{uniform random action } a' \in \mathcal{A} & \text{with probability } \epsilon \end{cases}$$
    2. Apply $a$
    3. Measure next state $s'$ and reward $r = \mathcal{R}(s,a,s')$
    4. Update $\boldsymbol{\omega}_a$
        $$\boldsymbol{\omega}_a \leftarrow \boldsymbol{\omega}_a + \alpha \left[ r + \gamma \max_{a'} Q_\omega(s',a') - Q_\omega(s,a) \right] \boldsymbol{\phi}(s) \qquad (2.54)$$
    $s \leftarrow s'$
```

2.1.4.3 Fourier Basis

Using linear function approximation, the agent cannot learn non-linear functions. However, often features $\phi_i \in \boldsymbol{\phi}(s)$ depend on each other and require the learning agent to learn non-linear dependencies. Using neural networks is impractical for user adaptation due to the high complexity. The Fourier basis transformation is utilized to learn non-linear functions while simultaneously employing linear function approximation.

In Konidaris et al. (2011), it has been demonstrated that the Fourier basis performs very well compared to commonly used radial and polynomial basis functions. Since it is also straightforward to apply, it is employed in this thesis. Using the Fourier basis, any function can be written as a sum of trigonometric functions. That way, any function can be easily linearized since the function parameters are combined through linear multiplication and addition.

While Konidaris et al. (2011) also describe the Fourier transformation for *uni-variate* functions, we are more interested in the Fourier basis for *multi-variate* functions (see Def. 2.6) as the transformed function depends on each feature $\phi_i \in \boldsymbol{\phi}(s)$ and generally has more than one feature.

Definition 2.6: nth Fourier Transformation of Multi-variate Functions

For any continuous, T-periodic, **uni-variate** function $f(x)$, the nth Fourier transformation is given by (Karpfinger, 2017, p.767):

$$\overline{f}(x) = \frac{a_0}{2} + \sum_{k=1}^{n} \left[a_k \cos\left(k\frac{2\pi}{T}x\right) + b_k \sin\left(k\frac{2\pi}{T}x\right)\right] \tag{2.55}$$

where

$$a_k = \frac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(x)\cos\left(k\frac{2\pi}{T}x\right)dx \tag{2.56}$$

$$b_k = \frac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(x)\sin\left(k\frac{2\pi}{T}x\right)dx \tag{2.57}$$

The Fourier transformation of **multi-variate** functions is similar, except that instead of scalar values (k, see Eq. 2.55), vectors ($\boldsymbol{c}$, Eq. 2.58) are used (Konidaris et al., 2011).

Let $f(\boldsymbol{x})$ be a multi-variate function with $\dim \boldsymbol{x} = m$, then the nth multi-variate Fourier transformation is defined as:

$$\overline{f}(\boldsymbol{x}) = \frac{a_0}{2} + \sum_{\boldsymbol{c}} \left[a_{\boldsymbol{c}} \cos\left(\frac{2\pi}{T}\boldsymbol{c} \circ \boldsymbol{x}\right) + b_{\boldsymbol{c}} \sin\left(\frac{2\pi}{T}\boldsymbol{c} \circ \boldsymbol{x}\right)\right] \tag{2.58}$$

where:

$$\boldsymbol{c} = \begin{pmatrix} c_1 \\ \vdots \\ c_m \end{pmatrix} \quad \text{with } c_j \in \{0, \dots, n\} \text{ and } 1 \leq j \leq m \tag{2.59}$$

For any function $f(x)$ it holds that (Karpfinger, 2017, p.770):

$$\begin{aligned} f(-x) = f(x) \Leftrightarrow (f(x)\text{is even}) \Rightarrow b_k = 0 \\ f(-x) = -f(x) \Leftrightarrow (f(x)\text{is odd}) \Rightarrow a_k = 0 \end{aligned} \tag{2.60}$$

Konidaris et al. (2011) argue that because functions generally are neither *even* or *odd* and when projecting input variables $\phi_i(s) \in \boldsymbol{\phi}(s)$ to [0,1], one of the two terms ($a_{\boldsymbol{c}}, b_{\boldsymbol{c}}$) can be dropped. They suggested dropping the sine term because the approximation of such *"half-even"* functions is easier [8], and proposed a Fourier transformation to convert the feature vector $\boldsymbol{\phi}(s)$ for any state $s \in \mathcal{S}$ into a Fourier basis vector $\overline{\boldsymbol{\phi}}(s)$ (see Def. 2.7):

[8] Karpfinger (2017, p.212) point out that if necessary, the sine term can be added.

Definition 2.7: Coupled Fourier Basis

Let $\phi(s) = [0,1]^m$ be a normalized feature vector representing a state $s \in \mathcal{S}$.

Additionally let $C = \left\{ c_1, \dots c_k \right\}$ be a set of coefficient vectors with $dim\, c_i = m, 1 \leq i \leq k = (n+1)^m$, and $\forall c_j \in c_i : c_j \in \{0, \dots, n\}, 1 \leq j \leq m$.

The nth Fourier basis $\overline{\phi}(s)$ for all entries $\overline{\phi}_i(s) \in \overline{\phi}(s)$ is defined as (Konidaris et al., 2011):

$$\overline{\phi}_i(s) = \cos(\pi c_i \circ \phi(s)) \tag{2.61}$$

Note that for all entries $\overline{\phi}_i(s) \in \overline{\phi}(s)$: $\overline{\phi}_i(s) \in [-1,1]$. The following example 2.6 illustrates the calculation to provide a clearer understanding of the coefficient vectors $C = \{c_1, \dots c_k\}$ and the Fourier basis transformation.

Example 2.6: Coupled Fourier Basis

Suppose we have an RL state $s \in \mathcal{S}$ represented by the feature vector $\phi(s) = \left(\frac{1}{2}, \frac{1}{4}\right)^T$, we want to calculate the $2nd$ order Fourier basis.

There is a total of $k = (n+1)^m = (2+1)^2 = 9$ coefficient vectors. This follows directly from Def. 2.7, as the maximum entry of the coefficients vectors must be equal to the nth order, thus each entry in c_i can take values from 0 to 2. Consequently, we get the following coefficient vectors:

$$C = \left\{(0,0),(0,1),(0,2),(1,0),(1,1),(1,2),(2,0),(2,1),(2,2)\right\} \tag{2.62}$$

Now, for each c_i, we calculate the corresponding Fourier basis:

$$\overline{\phi}_i(s) = \cos(\pi c_i \circ \phi(s)) \tag{2.63}$$

For $c_1 = (0,0)^T$, the corresponding Fourier basis would be:

$$\overline{\phi}_1(s) = \cos\left(\pi c_1 \circ \phi(s)\right) = \cos(\pi \cdot 0 \cdot 0.25 + 0 \cdot 0.75) = \cos(0) = 1. \tag{2.64}$$

Repeating this process for each $c_i \in C_i$, we get the Fourier basis $\overline{\phi}(s)$ that describes the given feature state $\phi(s)$ using the Fourier transformation:

$$\overline{\phi}(s) = \left(1, \frac{\sqrt{2}}{2}, 0, 0, -\frac{\sqrt{2}}{2}, -1, -1, -\frac{\sqrt{2}}{2}, 0\right)^T \tag{2.65}$$

The Fourier basis is computed for each state $s \in \mathcal{S}$. Using this basis, the agent subsequently learns the parameter vectors $\boldsymbol{\omega}_a$ for every action $a \in \mathcal{A}$ and thus the underlying function $Q(s,a)$ relevant for behavior. Note that $\overline{\phi}_5$, $\overline{\phi}_6$, $\overline{\phi}_8$, and $\overline{\phi}_9$ consist of multiple features. Thus, they allow for learning dependencies between features.

Not all feature values are typically dependent, or it is sufficient if they are not considered. Thus, one can compute a variable coupled Fourier basis (Konidaris et al., 2011; see Def. 2.8). The advantage of the variable coupled Fourier basis is the reduction of complexity of the vector dimension from $k = (n+1)^m$ to $k = \sum_{i=0}^{q} \binom{m}{i} n^i$ (Weber, 2017), as fewer coefficient vectors are needed (Konidaris et al., 2011).

Definition 2.8: Variable Coupled Fourier Basis (cf. Def. 2.7)

In addition to the conditions stated in Def. 2.7, the following constraint holds: For every $\boldsymbol{c}_i \in C$, there exist at most q indices $j \in \{0, \dots, m\}$ such that $c_j \in \boldsymbol{c}_i > 0$.

The nth Fourier basis is defined analogously to Def. 2.7, utilizing the modified coefficient vectors $\boldsymbol{c}_i$.

The additional constraint of Def. 2.8 states that for no coefficient vector $\boldsymbol{c}_i \in C$ there are more than q entries that are not equal to zero. In example 2.6, for $q = 1$, the set of coefficient vectors changes to $C = \big\{(0,0),(0,1),(0,2),(1,0),(2,0)\big\}$.

Using the Fourier basis, Eq. 2.51 changes to:

$$Q_{\omega}(s,a) := \overline{\boldsymbol{\phi}}(s) \circ \boldsymbol{\omega}_a = \sum_{i=1}^{n} \overline{\phi}_i(s)\, \omega_{a,i}, \forall s \in \mathcal{S}, a \in \mathcal{A} \tag{2.66}$$

What exactly does the agent learn? Using the cosine terms as feature approximation, the agent learns the $a_{\boldsymbol{c}}$ values of the Fourier transformation term as the following equation shows (cf. Eq. 2.58 in Def. 2.6 and Eq 2.61 in Def. 2.7):

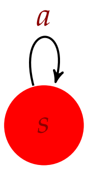
$$Q_{\omega}(s,a) = \underbrace{\omega_{a,1}}_{a_{\boldsymbol{c}_1}} \cdot \underbrace{\overline{\phi}_1(s)}_{\cos(\pi \boldsymbol{c}_1 \circ \boldsymbol{\phi}(s))} + \dots + \underbrace{\omega_{a,n}}_{a_{\boldsymbol{c}_n}} \cdot \underbrace{\overline{\phi}_n(s)}_{\cos(\pi \boldsymbol{c}_n \circ \boldsymbol{\phi}(s))} = \sum_{\boldsymbol{c}} \left[a_{\boldsymbol{c}} \cos\left(\pi\, \boldsymbol{c} \circ \boldsymbol{\phi}(s)\right) \right] \tag{2.67}$$

2.1.4.4 Convergence Criteria

When employing *linear function approximation* along with *Q-Learning*, the algorithm can diverge (Dann, 2012). Ex. 2.7 illustrates the problem (Weber, 2017):

Example 2.7: Divergence Example 1

Consider the following simple MDP:



Now, consider the parameters $\boldsymbol{\phi} = (3)$, $\alpha = 1.0$, $\gamma = \frac{1}{2}$, and $\mathcal{R}(s, a, s') = 1.0$.

Table 2.5: Oscillating divergence example.

t	ω_1	$q = \boldsymbol{\phi} \circ \boldsymbol{\omega}$	$\Delta\omega_1 = \alpha\,[1 + \gamma q - q]\,\phi_1$	$\omega_1 = \omega_1 + \Delta\omega_1$
1	0	0	3	3
2	3	9	-10.5	-7.5
3	-7.5	-22.5	36.75	29.25
⋮	⋮	⋮	⋮	⋮
10	52544	157633	-236447	-183902
⋮	⋮	⋮	⋮	⋮
n	$+-\infty$	$+-\infty$	$+-\infty$	$+-\infty$

As seen in Ex. 2.7, the weight ω_1 diverges to infinity. Why is that? The problem is the multiplicator $\boldsymbol{\phi}$ in the equation of the *Q-Learning* update rule. It can only be guaranteed that (cf. Eq. 2.51, and Eq. 2.54)

$$\underbrace{\phi_1(s)\underbrace{\left(\omega_1 + \left[\mathcal{R}_{t+1} + \gamma \max_a Q_\omega(s,a) - Q_\omega(s,a)\right] \cdot \phi_1(s)\right)}_{Eq.\ 2.54}}_{Eq.\ 2.51} \leq Q^*(s_t, a_t) \tag{2.68}$$

if $\left|\phi_1(s)\right| \leq 1$, otherwise:

$$
\begin{aligned}
&Q_\omega(s,a) = \\
&= \phi_1(s)\left(\omega_{a,1} + \left[\mathcal{R}_{t+1} + \gamma \max_a Q_\omega(s,a) - Q_\omega(s,a)\right] \cdot \phi_1(s)\right) \\
&= \underbrace{\phi_1(s)\,\omega_{a,1}}_{Q_\omega(s,a)} + \left[\mathcal{R}_{t+1} + \gamma \max_a Q_\omega(s,a) - Q_\omega(s,a)\right] \cdot \phi_1(s)^2 \\
&= \left(\mathcal{R}_{t+1} + \gamma\mathcal{R}_{t+2} + \gamma^2\mathcal{R}_{t+3} + \cdots + \gamma^{T-t-1}\mathcal{R}_T\right) \cdot \phi_1(s)^2 \\
&= \phi_1(s)^2 \underbrace{\sum_{k=0}^{T-t-1} \gamma^k \mathcal{R}_{t+k+1}}_{Q^*(s,a)} \\
&> Q^*(s,a)
\end{aligned}
\tag{2.69}
$$

causing the algorithm to oscillate as seen in Tab. 2.5. Therefore, we require that

$$\forall \phi_i(s) \in \boldsymbol{\phi}(s) : \phi_i(s) \in [-1, 1] \tag{2.70}$$

to avoid divergence. This however, only solves the problem if $dim\boldsymbol{\phi} = 1$ as the following Ex. 2.8 illustrates:

Example 2.8: Divergence Example 2

For the MDP in Ex. 2.7, consider the parameters $\boldsymbol{\phi} \in \mathbb{R}^{10}$ with $\phi_i = 1.0, \forall \phi_i \in \boldsymbol{\phi}, \alpha = 1.0, \gamma = \frac{1}{2}$, and $\mathcal{R}(s,a,s') = 1.0$.

Table 2.6: Oscillating divergence example.

t	ω_i	$q = \boldsymbol{\phi} \circ \boldsymbol{\omega}$	$\Delta\omega_{a,i} = \alpha\,[1 + \gamma q - q]\,\phi_i$	$\omega_i = \omega_i + \Delta\omega_i$
1	*0*	*0*	*1*	*1*
2	*1*	*10*	*-4*	*-3*
3	*-3*	*-30*	*16*	*13*
⋮	⋮	⋮	⋮	⋮
10	*52429*	*524290*	*-262144*	*-209715*
⋮	⋮	⋮	⋮	⋮
n	$+-\infty$	$+-\infty$	$+-\infty$	$+-\infty$

Again, the weight ω_i diverges to infinity. The problem here is how the delta difference ΔQ_ω of the *q-value* is added to the existing parameter vector:

$$\begin{aligned}
\boldsymbol{\omega}_a &= \boldsymbol{\omega}_a + \underbrace{\left(\mathcal{R}_{t+1} + \gamma \max_a Q_\omega(s,a) - Q_\omega(s,a)\right)}_{\Delta Q_\omega} \boldsymbol{\phi}(s) \\
&= \begin{pmatrix} \omega_{a,1} + \Delta Q_\omega \, \phi_1(s) \\ \vdots \\ \omega_{a,m} + \Delta Q_\omega \, \phi_m(s) \end{pmatrix} \overset{\phi_i(s) = 1.0}{=} \begin{pmatrix} \omega_{a,1} + \Delta Q_\omega \\ \vdots \\ \omega_{a,m} + \Delta Q_\omega \end{pmatrix}
\end{aligned} \tag{2.71}$$

which leads to a total increase of the *q-value* $Q_\omega(s,a)$ by m times greater than it should be:

$$\begin{aligned}
Q_\omega(s,a) &= \boldsymbol{\phi} \circ \begin{pmatrix} \omega_{a,1} + \Delta Q_\omega \\ \vdots \\ \omega_{a,m} + \Delta Q_\omega \end{pmatrix} \\
&= \left(\omega_{a,1} + \Delta Q_\omega\right) \phi_1(s) + \cdots + \left(\omega_{a,m} + \Delta Q_\omega\right) \phi_m(s) \\
&\overset{\phi_i(s) = 1.0}{=} \left(\sum_{i=1}^{m} \omega_{a,i}\right) + m \cdot \Delta Q_\omega
\end{aligned} \tag{2.72}$$

To solve that issue, we need to make sure that only $\frac{1}{m}$ of ΔQ_ω is added to each $\omega_{a,i} \in \boldsymbol{\omega}_a$, which can be achieved by setting the algorithm's learning rate to

$$\alpha = \frac{1}{dim \, \boldsymbol{\phi}(s)} \tag{2.73}$$

2.2 Psychological Models of Message Processing

The following section gives an overview of three established *psychological models* of message processing, describing how messages can be influenced by *subliminal biases*. Most people associate *subliminal bias* with unethical advertising (Allcott & Gentzkow, 2017). However, as seen within the models, it plays an essential role in everyday message processing, whether persuasion or decision-making.

2.2.1 Elaboration Likelihood Model (ELM)

In Sec. 1.3 and Sec. 1.2, we already described how a *subliminal bias* can be caused by *peripheral processing*, which is one route of the psychological model Elaboration Likelihood Model (ELM) by Petty and Cacioppo (1986). The ELM is a psychological model that describes how a message is processed based on the listener's NFC, which means the inclination of an individual to engage in challenging cognitive tasks (Bauer & Stiner, 2020; Cacioppo et al., 1984).

Two routes for *central* and *peripheral* processing are defined within the ELM (see Fig. 2.11). Central processing means that the speaker focuses on the content of the message if they have, among other factors (e.g., personal relevance; Petty and Cacioppo, 1986, p.150), sufficiently high NFC. They are more motivated to think about the message and critically deal with the content.

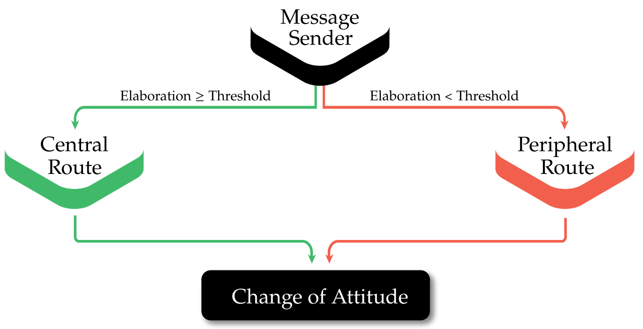


Figure 2.11: Elaboration Likelihood Model (ELM) describing how a message is processed based on two routes: 1) *central* and 2) *peripheral*. The *central* route is activated if a certain threshold is reached.

Individuals with low NFC, low personal involvement, or no motivation are inclined to be influenced by other factors, wherein the substance and quality of arguments within the message become secondary. In that case, when arguments have identical content but are delivered along with non-verbal behaviors, the persuasiveness of an argument can significantly vary. The peripheral route is often active when the listener lacks adequate knowledge of the topic or the motivation to engage with it thoroughly. Cues, emotions, or the listener's mood are then the priority of processing. For instance, slower speech or overuse of body language can be perceived as not convincing (Streeck, 2008; Yokoyama & Daibo, 2012), even though the content of the argument has stayed the same. This is because non-verbal behavior is essential to a speaker's credibility (Burgoon et al., 1990).

2.2.2 Heuristic Systematic Model (HSM)

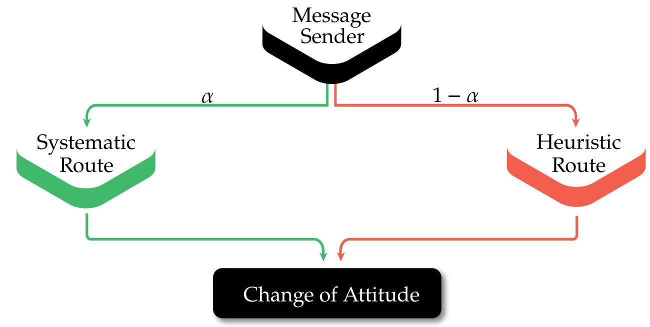


Figure 2.12: Heuristic Systematic Model (HSM) describing how a message is processed based on two routes: 1) *systematic* and 2) *heuristic*. The *systematic* route is activated if a certain threshold is reached, and both routes are then used concurrently based on an elaboration factor α.

The Heuristic Systematic Model (HSM) (Chaiken, 1989) is another similar psychological model that also describes two processing routes: *Systematic* and *heuristic* processing. These align with the central and peripheral routes. The Heuristic Systematic Model (HSM) states that both processes often operate concurrently, to some extent, whereas a certain elaboration threshold is required to activate systematic processing. In that case, the elaboration factor α is greater than zero. If systematic processing is not activated, α is zero.

2.2.3 Emotion As Social Influence Theory Model (EASI)

The Emotions as Social Influence (EASI) theory model by van Kleef (2014) describes how specifically emotional expressions can influence one's attitude, behavior, and understanding of the social environment (see Fig. 2.13). The theory distinguishes between *inter-personal* and *intra-personal* effects of emotions. Inter-personal influence refers to the effects caused by the emotions of others, while intra-personal influence determines the effects caused by one's own emotions.

The EASI theory further states that the effect of emotions is moderated by 1) the personal *epistemic motivation*, which triggers the way people process the underlying information, i.e., *"What is the reason for the sender's emotion?"*, *"Was there something we did wrong?"* and 2) the *perceived appropriateness* of the emotions.

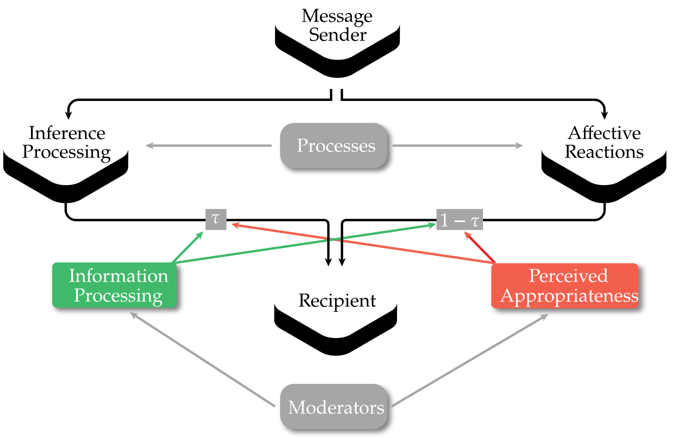


Figure 2.13: EASI Theory: Two processes (*inference processing* and *affective reaction*) moderated by two moderators (*information processing* and *perceived appropriateness*.)

Imagine the following scenario: Friends have arranged to meet at a café at 2 PM. One person arrives an hour late, showing up at 3 PM without informing the others of the delay. The friends, who have been waiting the entire time, express their frustration and anger.

When the latecomer's epistemic motivation is high, they are more likely to recognize the inconsiderate nature of their lateness. This understanding can lead to a positive outcome, as they may empathize with their friends' frustration and even offer a sincere apology for the delay. Conversely, if the latecomer's epistemic motivation is low, they may react defensively and perceive their friends' reaction as unfair.

Epistemic motivation is not a fixed value, though, but varies depending on the current situation (van Kleef, 2014).

In addition to the two moderators, EASI further describes that there are two processes, *inference processing* and *affective reaction*, which the moderators ultimately influence. Inference processing is simply the logical reasoning and reflection of the observed emotions, including the underlying information and situation, similar to the aforementioned ELM, while *affective reaction* corresponds to the emotional response triggered by the stimuli.

The processing of *inference*, *affective reaction*, and the effect of the moderators mostly takes place subconsciously based on a parameter τ.

2.3 Reflection Bias and Reflective Engagement (RE)

In this section, we narrow the literature to the scope of this work, *fostering critical reflection*, and establish the boundaries addressed within this thesis.

In the beginning, we discussed the people's tendency to focus on a biased subset of information, which indicates *low elaboration* and *fast thinking*. Kahneman (2012b) introduced the concept of slow and fast thinking. *Slow thinking* is necessary to be able to think about thinking and necessitates more cognitive resources than *fast thinking*. *Fast thinking*, on the other hand, is automatic, intuitive, and often subconscious and operates quickly to make judgments and decisions with minimal effort (Kahneman, 2012b).

While NFC can be the basis for *high elaboration* (Dole & Sinatra, 1998) and the motivation to use *slow thinking* and think reflectively, it does not necessarily imply the ability to do so. The Reflective Judgement Model (RJM) (see Fig. 2.14) distinguishes between seven stages of reflective judgment categorized into three levels, which are pre-reflective, quasi-reflective, and reflective. Quasi-reflective is also called *critical thinking* and considered a pre-form of *reflective thinking*. *Critical thinking* is a dynamic process that considers context and allows for self-correction; *reflective thinking* goes beyond, requiring ongoing reevaluation of one's beliefs, assumptions, and hypotheses (King & Kitchener, 1994).

The levels within the RJM are based on two fundamental assumptions, the first pertaining to the problem's structure.

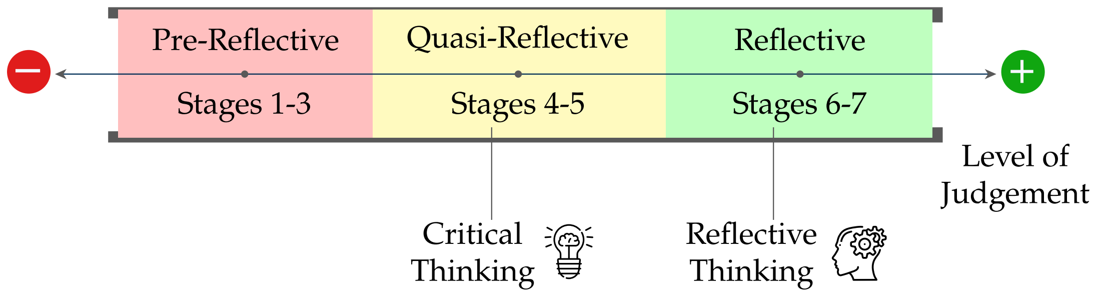


Figure 2.14: Reflective Judgment Model: Seven stages of judgment.

Problems with high certainty and completeness are characterized as *well-structured* and can be effectively addressed through *critical thinking*. In contrast, *ill-structured* problems, which are marked by uncertainty and lack a straightforward solution, make *reflective thinking* necessary. Awareness of uncertainty is an integral element for *reflective thinking* (King & Kitchener, 1994), which is called the *epistemological assumption about knowledge*.

Reflection can thereby manifest either as *weak sense* or *strong sense* reflection (Paul, 1981, 1990). *Strong sense* refers to reflective thinking and thus tends to correlate with *higher stages* of judgment and *slow thinking* through analytical processing and thorough examination of information and arguments. In contrast, *weak sense* thinkers often defend their own opinion without genuine reflection (Mason, 2007). Thus, *weak sense* reflection correlates with *lower stages* of judgment, and *fast thinking* by failing to acknowledge uncertainty in own opinions and making quick judgments about one-sided, biased information.

This phenomenon was often observed during the COVID-19 pandemic, as politicians often treated COVID-19 challenges as *well-structured* problems, even though they inherently showed characteristics of *ill-structured* ones. People then often defended their own opinion (*weak sense*) and declared every other contradicting opinion as factually wrong, which we learned later was not always true. This misjudgment underscores the importance of recognizing the nature of a problem and employing the appropriate mode of thinking, especially in contexts where uncertainty prevails.

Recognizing uncertainty and thinking critically about one's opinion requires *meta-cognitive* skills. Dean and Kuhn (2003) say that *meta-cognition* is the learnable skill of "*awareness and management of one's own thought, or 'thinking about thinking'*", and that *meta-cognition* is essential for the cognitive development, making it therefore necessary for *reflective thinking*.

Therefore, being reflective is "*a systematic, rigorous, disciplined way of thinking*" (Rodgers, 2002). This deliberate act of engagement is called Reflective Engagement (RE), defined as "*learner's continual and active participation in their problem inquiry with a continuous and critical judgment of inquiry process and inquiry outcomes for possible improvement*" (Farr and Riordan, 2012; Lyons, 2006; Rodman, 2010; cited by Kong and Song, 2015).

Weak sense reflection and *fast thinking* can both lead to: 1) a *content-based* reflection bias, where one relies on personal beliefs without deeper understanding and ignores opposing viewpoints, and 2) a *behavior-based* reflection bias, which involves using fewer cognitive resources and results in *low elaboration*, causing *peripheral processing* of information, as explained by the psychological model ELM (see Fig. 2.11).

Thus, to counteract *reflection bias*, one must actively engage in cognitive processes and be aware of the existing uncertainties and biases that can occur. By reducing content-based reflection bias, we can foster RE, and by raising awareness of subliminal biases, we can reduce behavior-based reflection bias.

2.3.1 Raising Awareness of Reflection Bias

Research of persuasiveness has ancient roots, tracing back to the era of the Roman Empire when the focus was on unraveling the qualities that define an ideal orator (Cicero, 2001). In today's age, research seeks to unveil the essential elements contributing to persuasive speech by examining verbal and non-verbal elements to 1) train people to become better orators and 2) build persuasive agents, robots, and chat-bots. Most studies focus on *what* stimuli affect persuasion rather than analyzing *why* someone is perceived as persuasive. Such stimuli can be among others *emotions* (DeSteno et al., 2004; Wang et al., 2015), *gaze* (Fischer et al., 2020; Ham et al., 2011; Kipp & Gebhard, 2008; Poggi & Vincze, 2009), and *gestures* (Maricchiolo et al., 2009; J. Peters & Hoetjes, 2017).

2.3.1.1 Emotions

Humans can experience a range of affects (emotions) such as excitement, boredom, happiness, or disappointment.

In literature, three primary approaches are used to label users' emotional states (André, 2011). Some of them are focused on classifying emotions; others describe the origin of an emotion:

- *Categorical models* characterize emotions as distinct categories. E.g., Ekman and Friesen (1971) categorized emotions into *basic emotion*, such as anger, happiness, and sadness.
- *Dimensional models* characterize emotions with respect to dimensions, e.g., valance and arousal, such as the Russels' circumplex model (Russell, 1980) which consists of two dimension, namely valence (negative, positive) and arousal (low, high). For instance, an emotion with negative valence and positive arousal is characterized as anger.
- *Appraisal models* describe emotions "*as valued reactions to emotion-eliciting stimuli*" (André, 2011). One of the most famous appraisal models is the so-called OCC (Ortony, Clore, and Collins) model (Ortony et al., 2022).

The effectiveness of a persuasive message depends on the appropriate usage of emotions. For instance, DeSteno et al. (2004) showed that the effectiveness of persuasive messages increases when their overtones align with the recipient's emotional state. Wang et al. (2015) showed the influence of emotions depending on the power dynamics between the message sender and recipient. Another study with virtual agents showed that people tend to agree more with speakers who show anger rather than happiness, indicating the significant effect that emotions can have (de Melo et al., 2012).

The aforementioned EASI theory model by van Kleef (2014) also indicates that emotions can influence one's attitude, behavior, and understanding of the social environment, and thus play an important role when it comes to the overall persuasion process.

One possible reason for this is that emotions are conveyed through multiple channels, such as *facial expressions*, and *speech* (Tomar et al., 2024), and *gestures* (Castellano et al., 2008).

2.3.1.2 Gestures

Gestures can be distinguished between *co-speech* and *non-co-speech* gestures. *Non-co-speech gestures* are gestures that occur independently of speech (Ekman, 2004):

- *Emblems* are gestures with a specific meaning, like a thumbs-up.
- *Regulators* control the flow of conversation.
- *Adaptors* (also called *manipulators*) are gestures related to self-touch (self-addressed) or manipulation of objects (object-addressed).

Co-speech gestures, also referred to as *illustrators* (Ekman, 2004), are hand and body movements that accompany spoken language, enhancing the communication and providing additional context or emphasis to the spoken words (McNeill, 1992). There are several types of co-speech gestures as identified by McNeill (1992):

- *Iconic gestures* visually represent the content of speech, helping to illustrate and clarify the spoken words.
- *Metaphoric gestures* are abstract gestures that represent ideas or concepts, aiding in conveying complex or intangible content.
- *Deictic gestures* are pointing gestures that direct the listeners' attention to specific objects, locations, or directions.
- *Beat gestures* are rhythmic hand movements that align with the speech's natural cadence, helping to maintain the listeners' attention and emphasizing the structure of the speech.

Maricchiolo et al. (2009) explored the effects of ideational gestures (iconic, metaphoric, deictic), conversational gestures (e.g., beat), object- and self-addressed adaptors as well as no gestures on their perceived persuasiveness, style effectiveness, composure, and competence. They found ideational gestures to be more persuasive than adaptors. In addition, they found that when gestures were present, the participants paid a lot more attention to them. Further, participants paid more attention to adaptors that were object-addressed than ideational gestures.

C.-M. Huang and Mutlu (2013) investigated the effect of illustrator gestures of a robot narrator on task performance (information recall), perceived performance (gesture effectiveness, competence, naturalness), social and affective evaluation (rapport, engagement), as well as narration behavior (gesture use, narration duration). Among many other effects, they showed that, for instance, deictic gestures significantly enhanced information recall for females, while for males, both deictic and metaphoric gestures significantly enhanced information recall. They also showed that for females, deictic, beat, and metaphoric gestures significantly enhanced the robot's gesture effectiveness, while for males, only beat gestures did so.

J. Peters and Hoetjes (2017) showed how gestures can significantly affect viewer perception based on the abovementioned tendency to process information *peripherally*. First, the study found that participants who viewed a speech with hand gestures rated the speech more persuasive than the control group without gestures in line with the other literature. In addition, they found that this was even more pronounced for participants with low elaboration. They showed an interaction effect between elaboration, hand gestures, and the rated factual accuracy of the content.

2.3.1.3 Gaze

Gaze behavior can also be categorized into five types (Argyle & Cook, 1976):

- *Mutual gaze* refers to maintaining eye contact with another person.
- *Averted gaze (Avoidance)* refers to looking away from another person.
- *Gaze aversion* refers to looking away while thinking.
- *Fixed gaze* refers to fixing one's gaze on an object or something else.
- *Gaze following* refers to looking where another person is looking.

Ham et al. (2011) had participants listen to a Nao robot telling the Greek story "*The boy who cried wolf*" (Aesop, 2020). The robot either used gestures, gazing, both, or neither. Persuasiveness was measured by having participants tell the lying character in the story. Their results showed that gestures only increased persuasiveness when accompanied by gazing. Gazing alone without gestures positively affected persuasion as well.

Conversely, Chen et al. (2013) showed that too much eye contact can even increase resistance to persuasion and decrease persuasiveness.

Fischer et al. (2020) explored the impact of coordinating gaze and speech behaviors to enhance a robot's persuasiveness. The study was conducted at a public event using the SMOOTH (Juel et al., 2020) robot to serve water to attendees. The study examined how the robot's verbal messages were received depending on whether it established mutual gaze with the participants. The results revealed that when the robot gazed at participants while saying "*skål*" (cheers), a significantly higher percentage of them drank immediately compared to when the robot did not gaze at them. Similarly, water-related jokes told by the robot were more effective in eliciting laughter when mutual gaze was present.

These findings indicate that a persuasive message's effect depends not just on a particular stimulus but on the proper usage and combination of stimuli.

Raising Awareness of Reflection Bias

While the presented studies demonstrated *what stimuli* (e.g., gestures, gaze, emotion) contribute to persuasiveness, they did not investigate *why* a speaker is perceived as more or less persuasive. More specifically, most existing studies primarily focused on systematically varying stimuli to investigate the effects of specific non-verbal cues, while our research takes a different approach. Rather than analyzing manipulated stimuli, we examine existing video material, leveraging annotations to understand the persuasiveness of speakers. In contrast to traditional studies that assume certain features may have an effect, we employ XAI to uncover why a speaker is persuasive.

This shift from manipulating stimuli to analyzing real-world, annotated video material allows us to explore the intricacies of persuasion in a novel way, shedding light on the underlying factors that contribute to individual persuasiveness. Our methodology aims to provide a deeper understanding of persuasion, introducing a unique dimension to the existing research in this field.

Previous studies have laid the groundwork for understanding the persuasive impact of non-verbal cues; this thesis explores how reflection bias can be practically identified and compared among individuals. Examining *why* people are perceived as persuasive fosters a deeper understanding of the impact of individual cues and, thus, increases awareness.

2.3.2 Fostering Reflective Engagement

Baumer et al. (2014) identified two main application types in which reflection was explored: 1) *design* and 2) *education*.

2.3.2.1 Design

Regarding *design* and *reflection*, the literature focuses on the design process and the creator rather than cognitive processes and outcomes.

Dijk et al. (2011) presented an interactive tangible system called NOOT supporting reflection during brainstorming sessions of a design process. Their system consists of *tangible clips*. These clips create a spatial context for audio tag files, allowing users to revisit and replay recorded conversations associated with specific clips. This enables users to reflect on the previous brainstorming ideas more easily.

Hailpern et al. (2007) proposed an interaction model with *spatial maps*, allowing a team of designers to work efficiently by enabling them to work simultaneously on multiple design ideas. Some of the key features of this model include, among others, *multi-level sharing of ideas across several devices*, *showing multiple ideas at the same time*, and *allowing rapid access to own and others' designs and ideas*. The results of their initial evaluation show that design teams could utilize the system to enhance their creative design process effectively. In comparison to alternative models like *Tabs* using Microsoft® Office OneNote®, and *layers and canvases* using Adobe® Photoshop® CS3, the *spatial map* model was rated higher for reflection and viewing all ideas, as shown in Smith et al. (2010), which was evaluated based on participants' feedback.

Gennari et al. (2021) explored how workshops could encourage children to engage in reflective thinking and develop an awareness of different aspects of the design process. Their workshop centered around creating smart objects for a park, utilizing a structured design approach based on a card board game. The results indicate that the workshop positively influenced how children contemplated design and their overall understanding of it.

In summary, when it comes to design, the act of *reflection* often revolves around thinking about different ideas during the design process.

2.3.2.2 Education - General Overview

In the education context, the primary focus often lies on the outcomes for learners.

An *activity meter*, proposed by Govaerts et al. (2012), is utilized to visualize students' actions and offer insights into metrics like *time spent*, thereby enhancing self-awareness and resource utilization. The learning progress, understanding, and self-reflection were assessed based on post-task interviews.

Kharrufa et al. (2010) presented a collaborative learning application. It employs a tabletop to facilitate externalizing thinking and improve schoolchildren's higher-level thinking skills. The system demonstrated an increased likelihood of effective learning and fostering higher-level thinking through reflective interaction with the application and other children.

Santos et al. (2013) introduced a learning analytic platform that supports awareness and self-reflection by portraying students' activities as bar charts and comparing them to those of others. The authors identified the most prevalent learning problems through brainstorming sessions and evaluated how their dashboard addresses these issues. The evaluation revealed that while many students did not perceive significant added value from the platform, it impacted groups more positively than individual students.

Blasco et al. (2015) explored the use of cinema in education to stimulate learners' reflection and impact their affective domain. They argue that emotions are key in shaping learning attitudes and driving behavioral change. Therefore, educators must address learners' affective domains to ensure adequate education. They investigated a movie-clip methodology, which involved showing multiple movie clips rapidly, accompanied by facilitator comments. They argue that this method can promote reflective thinking by providing a discussion forum and improving teaching skills.

Similarly, Yip et al. (2019) conducted a study examining the impact of *Augmented Reality Videos* compared to hand-outs, revealing a significant main effect on task comprehension. They measured the learning effect through post-test questionnaires and time-based scores.

Silpasuwanchai et al. (2016) conducted a study demonstrating that various *gamification* methods can effectively enhance skill acquisition and transfer. They specifically examined the impact of employing *badges*, *points*, and *leaderboards* on different aspects of engagement: *behavioral engagement* (measured by *number of attempts* and *effort*), *emotional engagement* (measured by *valence and arousal* and *endurability*), and *cognitive engagement* (measured by *focused attention* and *reflection*). The results showed significant influences on *effort*, *attention*, and *reflection*, with a preference for *leaderboards* over *points* and *points* over *badges*. The study highlights that *gamification* significantly impacts users' problem-solving performance.

Ward and Litman (2011) evaluated the effect of providing a reflective reading after dialog-based tutoring in physics. Following each tutoring session, the system presented a problem statement and asked questions related to it. To assess reflection and learning, they conducted pre and post-tests requiring students to employ Newton's laws in scenarios distinct from the problems covered during the tutoring sessions. The results showed that the reflective text did not significantly improve learning compared to the control condition. However, they found a moderate interaction effect between the student's motivation and reflection.

Grigoriadou et al. (2005) developed a dialogue-based learning platform to enhance learners' understanding of historical texts by means of reflective dialogues. Employing individual cognitive profiles, this system tailors conversations to each learner, thereby fostering reflection and reasoning. The student is asked about several factors, such as *position* and *justification*. Based on the student's answers (selected from a list of alternatives), the system analyzes the overall comprehension based on three factors: 1) *remember*, 2) *understand*, and 3) *analyze*. The alternative answers are thereby classified as either *valid*, *towards-valid* and *non-valid*. An answer is considered *complete* if both *position* and *justification* are *valid*, else *non-complete*, which is the case if *position* and *justification* contradict. Thereby, *completeness* is defined on a discrete scale ranging from *complete* to *incomplete*. The degree of *answer completeness* measures the cognitive profile. They employed dialogue strategies, namely *theory of inquiry teaching* by Collins (1986). These strategies aim to enhance cognitive abilities rather than focusing solely on knowledge specific to certain subjects.

2.3.2.3 Education - Argumentative Applications

In the context of educational applications, argumentation plays an important role as well. It often focuses on enhancing speaking skills (Darmawansah et al., 2022) and learning the process of effective argumentation (Guo et al., 2023; Iordanou & Constantinou, 2015).

Various platforms have been developed to enable students to practice metacognitive skills required for argumentation, often using gamification elements (e.g., L. Huang and Yeh, 2017). Most of the earlier systems are based on a visual representation of arguments students can interact with using graphical interfaces. A prominent example is https://debategraph.org/. It allows users to create, visualize, and explore argument structures, provides tools for collaboration in different fields (e.g., education, conflict resolution, and media), and overall community engagement in debates. Unlike our work, it does not allow users to discuss with an agent, like a chat-bot.

Petukhova et al. (2017) developed a virtual debate coach that trained young politicians' multimodal rhetorical skills to engage in a political debate. The system's task was to give feedback on inappropriate debate behaviors addressing arguments' structure, quality, and presentation. Goda et al. (2014) developed a chat-bot with which foreign-speaking students could interact to enhance their discussion skills before entering a human-human group discussion. They showed that people interacting with the chat-bot engaged much more in the subsequent group discussion.

2.3.2.4 Other Argumentative Applications

Beyond educational contexts, argumentative applications are explored in various domains. Roussou et al. (2019) developed a chat-bot discussing controversial topics like "Would you bury someone under your bed" provocatively to enable users to think more critically about the topic and get out of their comfort zone. However, a mere confrontation with opposing arguments in such a competitive setup (similar to a debate) can lead to cognitive dissonance (Hart et al., 2009), which can have a negative effect (defensive attitude (Harmon-Jones, 2000)). Therefore, a confrontation in a competitive rather than cooperative scenario is more likely to lead to rejection.

Most approaches to human-agent argumentation are embedded in a competitive setting with or without embodied agent (Rach et al., 2021; Rosenfeld & Kraus, 2016; Slonim et al., 2021; Weber et al., 2020b). They utilize different models to structure the interaction (similarity model to retrieve counterarguments (Rakshit et al., 2019), retrieval- and generative-based models (Le et al., 2018), or pretrained dialog strategy (Rach et al., 2018a; Weber et al., 2020b)).

In contrast, (Aicher et al., 2021b) introduced a cooperative chat-based argumentative dialogue agent that provides arguments upon users' request without trying to persuade or win a debate against the user.

Neither the competitive nor cooperative systems have a mechanism to foster an unbiased argument exploration, though. Instead, existing systems often have a different goal, such as opinion building (Aicher et al., 2021b), persuasion (Chalaguine & Hunter, 2020; Mishra et al., 2022), information/argument retrieval (Fazzinga et al., 2021; Rakshit et al., 2019), and inquiry (Black & Hunter, 2009).

Advanced state-of-the-art Large Language Models (LLMs), such as *InstructGPT* (Ouyang et al., 2022) and *ChatGPT*, allow users to discuss nearly any topic (depending on the data set they were trained on). *InstructGPT* and similar *instruction-based* LLMs are specifically refined through RL using human feedback (Stiennon et al., 2020) to improve its ability to follow explicit instructions. *ChatGPT*, on the other hand, is optimized for conversational contexts. While it also leverages RL from human feedback to enhance its performance, it is designed to handle longer, back-and-forth interactions more effectively[9].

The behavior, and whether the discussion is reflective and unbiased, depends on the user's prompts. Responses by language models like *ChatGPT* tend to mirror the biases of the individuals asking the questions. Given the people's tendency to focus on sources aligning with one's opinion (Ekström et al., 2022), this can result in a cycle of confirmation biases rather than providing diverse viewpoints, fostering a so-called echo chamber effect (Sharma et al., 2024).

Thus, unlike earlier platforms for teaching argumentation, debates of controversial topics, or argument retrieval, we aim to promote metacognitive skills, such as being open to new opposing views or topics instead of training specific argumentation strategies. Earlier work made use of interface agents whose role was, for example, to identify new topics for discussion to keep the conversation going. For example, Isbister et al. (2000) introduced an interface agent to enhance cross-cultural human-human interaction in a virtual environment. More recently, Kusajima and Sumi (2018) presented an agent to activate group discussion by suggesting web pages that matched keywords related to the current discussion. However, the main objective of these approaches was to ensure the progress of a conversation as opposed to engaging with diverging views.

Fostering Reflection

Reflective practices have been central to education, teaching, and learning. Dewey (1910, 1933) underscores the significance of reflection within the educational context. His prolific use of the term *education* (approximately 85 times) compared to *reflection* (roughly 55 times) exemplifies its correlation. Another expert in reflective thinking, Schön (1983), similarly emphasized education in reflective practices. In argumentation, the main focus often revolves around enhancing speaking skills and learning the process of effective argumentation rather than exploring diverging points of view.

Thus, we have develop an intelligent argumentative dialogue agent that enables users to engage with diverging points of view on a particular topic. Our intelligent agent integrates intervention strategies to promote a less biased argument exploration. The system actively tracks the user's argument focus during interaction and encourages a more diverse argument focus based on

[9] https://openai.com/blog/chatgpt/

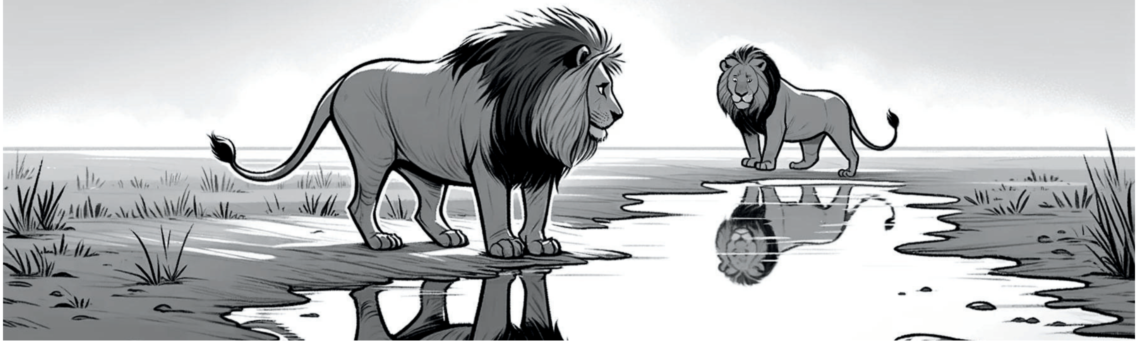
a computational metric AVQ for RE.

While there are subjective measures for RE, such as questionnaires, to assess reflective thinking (Kember et al., 2000; Lee & Dey, 2011; Leijen et al., 2009), such tests are not suitable for real-time assessment of reflective argument exploration and, thus, impractical for our approach. Objective measures used in teaching practices often encompass measurable outcomes (Govaerts et al., 2012; Kharrufa et al., 2010; Santos et al., 2013), reflective markers in free-text statements (Farr & Riordan, 2012), time metrics (Dupret & Lalmas, 2013), interaction behavior, such as cursor movements by leveraging mouse data to predict the perceived significance of specific web content (Arapakis et al., 2014), and area focus on websites (Yi et al., 2014). Using the computational metric AVQ, we assess users' reflective argument exploration by considering their explicitly expressed stance on a given topic in relation to their focus during the interaction.

Based on the metric, we can calculate a normalized score, indicating the user's focus on *challenger arguments*, which are arguments contradicting their point of view. The intelligent agent uses this metric to challenge the user's argument focus into a less biased one by encouraging the user to look into *challenger arguments* rather than sticking to arguments supporting their own opinion.

Chapter 3

Raising Awareness of the Subliminal Reflection Bias



"The eye sees only what the mind is prepared to comprehend." (Henri Bergson)

❗ Read Sec. 2.1.1 (neural networks) - 2.1.2 (background on XAI).

ℹ This chapter answers research questions **Q1.1** and **Q1.2**:

Q1.1: *"Can we effectively uncover behavior-based reflection bias in political speeches and provide satisfactory explanations?"* (Sec. 3.2)

Q1.2: *"Can XAI contribute to highlighting and understanding subjective differences in persuasive cues in political speeches?"* (Sec. 3.3) ℹ Most of the work presented in this chapter was previously published by the author in peer-reviewed papers (Weber et al., 2020c, 2023b). It is *reproduced with permission from Springer Nature* in the scope of this work. The paper Weber et al. (2023b) is based on a Bachelor's thesis by Tinnes (2022) that I supervised.

In this chapter, we investigate the research questions **Q1.1** and **Q1.2**. We investigate *why* a speaker is perceived as persuasive rather than investigate *what* stimuli make a speaker persuasive. To do that, we train (a) neural network(s) on video frames annotated regarding their persuasiveness, and employ XAI to uncover what the network focuses on.

In the first part (see Sec. 3.2), we investigate if we can uncover behavior-based reflection bias and provide satisfactory explanations (**Q1.1**). This is done by training a single neural network based on semi-subjective data comprising aggregated data of three annotators and analyze the feasibility of learning persuasiveness based solely on video frames. The biggest challenge arises from the subjectivity of the data. Using subjective data has a lot of noise compared to objective gold-standard data. To account for subjectivity, we verify the agreement of annotations among all three annotators. We then investigate XAI methods to highlight whether the network has effectively learned what makes a person persuasive.

In the second part (see Sec. 3.3), we investigate if XAI can contribute to highlighting and understanding subjective differences in persuasive cues (**Q1.2**). To do so, we train three distinct individual networks using entirely subjective data sourced from 30 videos, each annotated by three annotators. We then apply XAI to analyze the video frames of each network, compare the findings with the existing literature, and highlight differences among annotators.

3.1 Methodology

The strength of our method lies in its ability to reveal what individual users might have focused on by using XAI to explain why an NN rated a video as *persuasive* or not based on the users' annotations. By highlighting what the network focused on when predicting persuasiveness from the annotated data, we can *raise awareness* of *individual subliminal biases* within the persuasion process. Analyzing whether the network's focus aligns with other findings in the literature creates *awareness* of *subliminal cues*.

However, as mentioned in Sec. 2.1.2, annotators may still focus on different cues than those used by the NN when categorizing videos as persuasive or non-persuasive. NNs may not even focus on the primary learning objective (Lapuschkin et al., 2019), which, in our context, is the person and their gestures, as illustrated in Fig 2.6. Nevertheless, XAI remains a robust methodology for several reasons:

1) XAI is a promising tool for uncovering the underlying reasons behind the NN's predictions of persuasive and non-persuasive categorizations.

2) By providing insights into the cues that the NN considers important, XAI can bridge the gap between the annotators' diverse perspectives and the NN's learned patterns.

3) By understanding the NN's decisions, we can identify the specific cues that contribute to persuasive categorization. This, in turn, facilitates the identification of shared persuasive elements among annotators, even if their focus differs. Especially in cases of varying annotator emphasis, XAI is a suitable choice to highlight and understand potential biases, thereby contributing to a heightened awareness.

The process of our methodology is divided into four steps (see Fig. 3.1):

1. Annotation of videos based on visual and audio output channels.
2. Training and fine-tuning the neural network(s) using video frames as input.
3. Generating explanation images using XAI.
4. Analysis of the explanation images and comparison with existing literature.

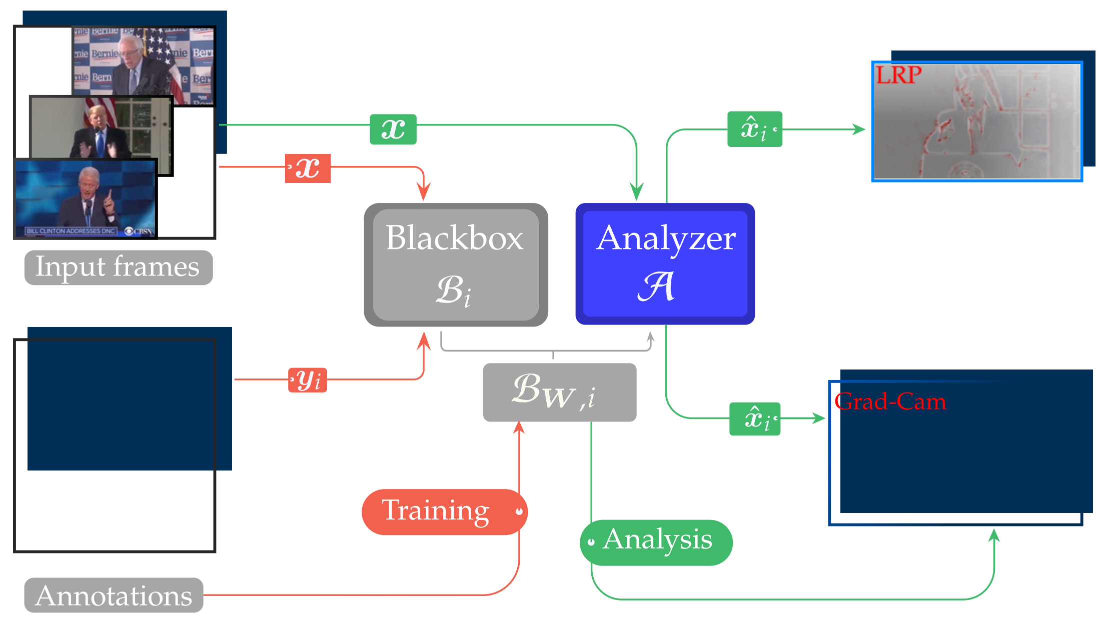


Figure 3.1: Methodology: Annotations $\mathcal{Y}_i$ of an annotator $i \in \mathcal{I}$ are used to train the network $\mathcal{B}_i(\boldsymbol{x})$ for all video frames $\boldsymbol{x} \in \mathcal{X}$. Then the analyzer $\mathcal{A}_{exp,i}$ is applied to obtain explanation images $\hat{\boldsymbol{x}}_i$, $\forall \boldsymbol{x} \in \mathcal{X}$.

We denote the NN for each annotator $i \in \mathcal{I}$ as function $\mathcal{B}_i : \mathcal{X} \rightarrow \mathcal{Y}_i$. The input frame of $\mathcal{B}_i$ is denoted as $\boldsymbol{x} \in \mathcal{X}$ with $\boldsymbol{x} \in \mathbb{R}^{60\times190\times3}$. Further, we denote $\boldsymbol{y}_{\boldsymbol{x}}^{i} \in \mathcal{Y}$ as the 5−dimensional target output vector for an arbitrary given input frame $\boldsymbol{x} \in \mathcal{X}$ for annotator $i \in \mathcal{I}$. Finally, $\hat{\boldsymbol{y}}_i = \mathcal{B}_i(\boldsymbol{x})$ denotes the actual output of the NN.

First, we take the annotations $\mathcal{Y}_i$ of any arbitrary annotator $i \in \mathcal{I}$ and train the respective network $\mathcal{B}_i(\boldsymbol{x})$ with all video frames $\boldsymbol{x} \in \mathcal{X}$. We then feed the same video frame into the Analyzer $\mathcal{A}_{exp,i}$ (see Def. 3.1) with a specified explainer exp, such as Grad-CAM or LRP, along with the trained network $\mathcal{B}_i$ and obtain explanation images $\hat{\boldsymbol{x}}_i$, $\forall \boldsymbol{x} \in \mathcal{X}$.

Definition 3.1: Analyzer

Let $\mathcal{B}_i : \mathcal{X} \rightarrow \mathcal{Y}$ be the network model, $\boldsymbol{x} \in \mathcal{X}$ the input frame and $exp : \mathcal{Y} \times \mathcal{X} \rightarrow \hat{\mathcal{X}}_i$ be the specified explainer of model $\mathcal{B}_i$, then the network analyzer model $\mathcal{A}_{exp,i} : \mathcal{X} \rightarrow \hat{\mathcal{X}}_i$ is defined as

$$\mathcal{A}_{exp,i} := exp\left(\mathcal{B}_i(\boldsymbol{x}), \boldsymbol{x}\right) \tag{3.1}$$

3.2 Uncovering the Behavior-Based Reflection Bias

ℹ This section answers research question **Q1.1**:

Q1.1: *“Can we effectively uncover behavior-based reflection bias in political speeches and provide satisfactory explanations?”*

3.2.1 Corpus and Annotation Process

The training corpus consists of a public speech delivered by Donald J. Trump in 2019 [10], approximately 50 minutes. The data were annotated using NOVA (Baur et al., 2020), an annotation tool for annotating and analyzing behavior in social interactions. The NOVA user interface was designed to annotate continuous recordings with multiple modalities and subjects. It supports techniques from the latest developments in current research, such as Cooperative Machine Learning (CML) and XAI, to speed up the standard annotation process using automation. Within this work, we do not apply CML, as this could potentially falsify the results due to the expected high subjectivity of the annotations.

[10] https://www.youtube.com/watch?v=DU6BnuyjJqI

Three experienced labelers annotated the video at a sampling rate of 25 Hz. They were asked to rate how convincing they found the speaker, distinguishing between five different levels (from *not at all convincing* to *very convincing*). Despite the subjectivity, the annotators achieved an agreement of 0.77 (*Cronbach's alpha*, cf. Sec. 2.1.1.6), which seems sufficient for our purposes given the high subjectivity of perceived persuasiveness (Kaptein et al., 2010). The annotations were merged (see Fig. 3.2) to obtain a *gold standard* stream with over 50,000 sample video frames. The two lowest classes were barely present in the annotated data set, meaning that the annotators found the video generally more persuasive than not.

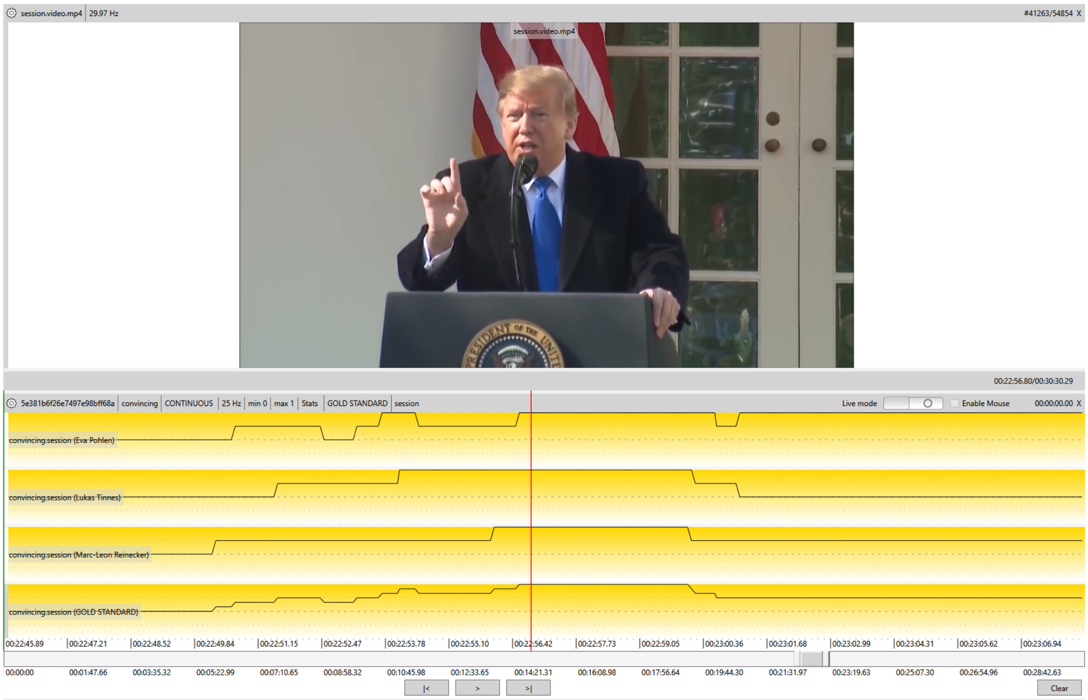


Figure 3.2: The NOVA tool depicting the video at the top and four annotation streams below (3 annotators + merged gold standard data).

3.2.2 Model Architecture and Training

Standard NNs have been shown to be effective in predicting persuasiveness based on visual, audio and text features (Nojavanasghari et al., 2016).

As we only use the video frames as input, we make use of a CNN (see Fig. 3.3), which consists of three successive convolutional layers. Layers for batch normalization and max-pooling follow the last two layers. The output of the last convolutional layer is flattened and then, to obtain probabilities for all five classes, fed into a five-fold softmax activation function (Goodfellow et al., 2017,

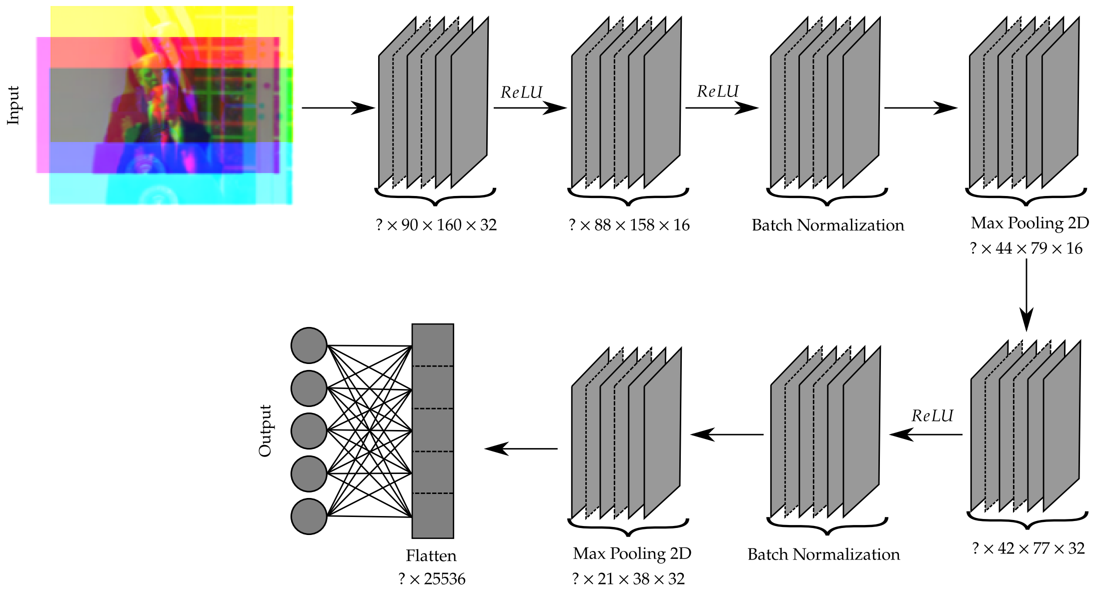


Figure 3.3: An illustration of the network architecture. The network consists of three convolutions. The first layer expands the 3-channel RGB to 32 channels before being fed into the last two convolution layers, after each of which batch normalization and max-pooling are applied. The network outputs a 5-vector, estimating the probability of each class.

pp. 178-182), which has a similar form as the sigmoid function (cf. Fig. 2.2(b)).

We first extracted the video frames with a sampling rate of 25 Hz and downsampled them to 160x90 RGB frames. The first convolution layer extends the RGB channel of the input frame to 32 channels. The idea behind this is that we allow the network to define colors for different pixel combinations, similar to how humans, for example, see a combination of yellow and blue as green.

The network outputs a five-dimensional vector describing the probability of each class. A ReLU activation is used in each layer, except the output layer, where the softmax activation is applied. We further use ADAMAX ($\beta_1 = 0.9$, $\beta_2 = 0.999$, Kingma and Ba (2014), cf. Sec. 2.1.1.3) as the optimizer. To avoid zero initialization, the NN's weights are initialized employing the Xavier initialization (Glorot & Bengio, 2010). The Xavier initialization ensures that the variances of the weights remain the same across all layers, i.e., the weights neither explode nor vanish to zero. To tackle overfitting, we use batch normalization (Bjorck et al., 2018; Ioffe & Szegedy, 2015) and L2 regularization. Batch normalization is a technique that re-centers and scales the input of a layer. It is applied after the second and third convolutional layers, followed by pooling layers.

The model was trained for 100 epochs with a batch size of 32, splitting the data set 4:1 into training and validation data. Fig. 3.4 shows that after only 20 epochs, the neural network could reliably predict classes with an accuracy of > 98% on the training set. Since the validation loss after 20 epochs shows slight overfitting, the network analyzed in this thesis was trained for only 20 epochs.

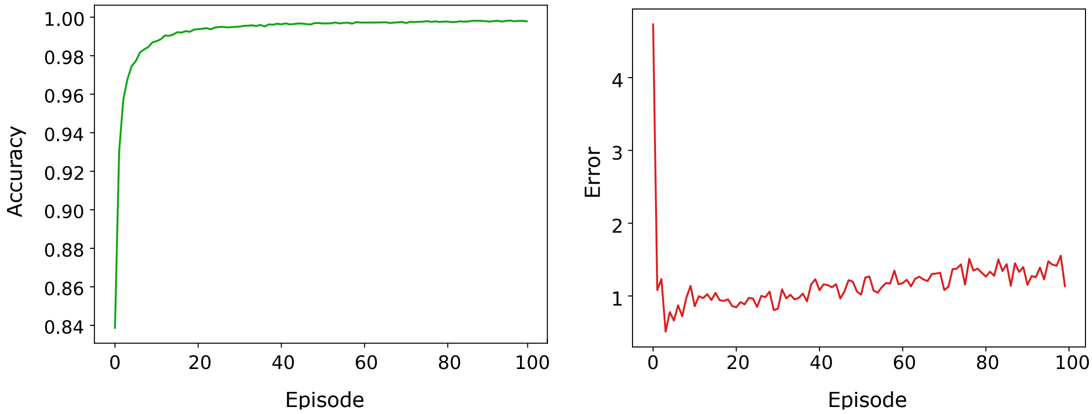
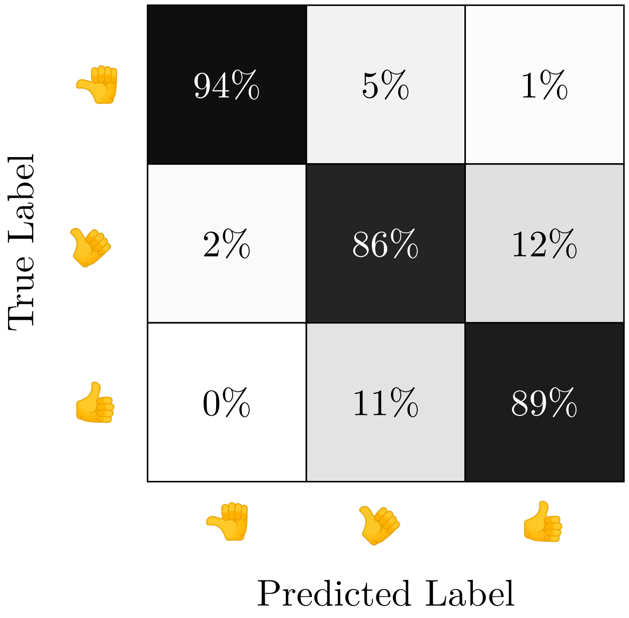


Figure 3.4: Training accuracy (left) and validation loss (right) over 100 epochs.

To validate the overall performance of the network, we computed the confusion matrix on the training data set as visualized in Fig. 3.5. As we did not aim to train a general predictor for persuasiveness, we evaluated our model on the training data set to ensure that our network is sufficiently accurate on the learned samples. Since the lowest two classes were not annotated at the current stage, they are not listed in the matrix. Table 3.1 shows the trained model's precision, recall, and F1 scores (cf. Sec. 2.1.1.4), indicating a high model performance.



	👈	🤙	👍
👈	94%	5%	1%
🤙	2%	86%	12%
👍	0%	11%	89%

Figure 3.5: Confusion matrix computed on the training data set to ensure that our network is sufficiently accurate on the learned samples. (👈) neutral, (🤙) moderately convincing, (👍) very convincing.

Table 3.1: Network Performance across the classes *neutral*, *convincing*, and *very convincing*.

Measure	Class		
	👈	🤙	👍
Precision	*0.93*	*0.93*	*0.77*
Recall	*0.94*	*0.86*	*0.88*
F1-Score	*0.93*	*0.89*	*0.82*

3.2.3 Highlighting the Reflection Bias

Since we trained the network on video frames only, it was forced to base its prediction on body cues. The interesting question is, which sections were the most relevant for making a (correct) prediction and whether there are features consistent with the existing literature, i.e., did the network learn to focus on frame sections that are indicators of perceived persuasiveness? To investigate this, we applied two XAI techniques: (1) Grad-CAM and (2) LRP.

As outlined, using both Grad-CAM and LRP helps garner both *macro* and *micro* level insights (cf. Sec. 2.1.2). Specifically, Grad-CAM highlights the global contributions of various image regions towards the network's predictions by emphasizing significant features at a *macro* level. On the other hand, LRP delves into the intricate details, offering a fine-grained (*micro* level) understanding of how individual features and neurons influence the model's predictions. Using these methods together allows for a more complete and balanced interpretation.

3.2.3.1 Grad-CAM - Visualizations

We first analyzed the last layer of the network with Grad-CAM (Selvaraju et al., 2017) using keras-vis (Kotikalapudi & contributors, 2017) (cf. Sec. 2.1.2.1 for technical details), a high-level toolkit for visualizing trained neural network models. For better visualization, we created edge images of the input frames and overlaid the visualization maps of the network. For our visualizations, we only used frames with $\arg\max_k(\hat{y}_k) = \arg\max_k(y_k)$ and $\hat{y}_k \geq 0.95$, i.e., we only visualized frames that were correctly classified with very high confidence probability resulting in about $33,000$ frames. Table 3.2 outlines the predictive power based on different minimum confidence scores of the network.

Table 3.2: Network Predictions. The number of correctly predicted frames based on different minimal output confidences from 60% to 99%.

Min Confidence (y_k)	99%	95%	90%	75%	60%
Number of Frames Total	27,305	33,043	35,365	38,102	39,446
Number of Additional Frames	-	+5,738	+2,322	+2,737	+1,344

Fig. 3.6 shows the generated saliency maps indicating the intensity of focus, with red regions signifying strong attention and blue areas representing neglect.

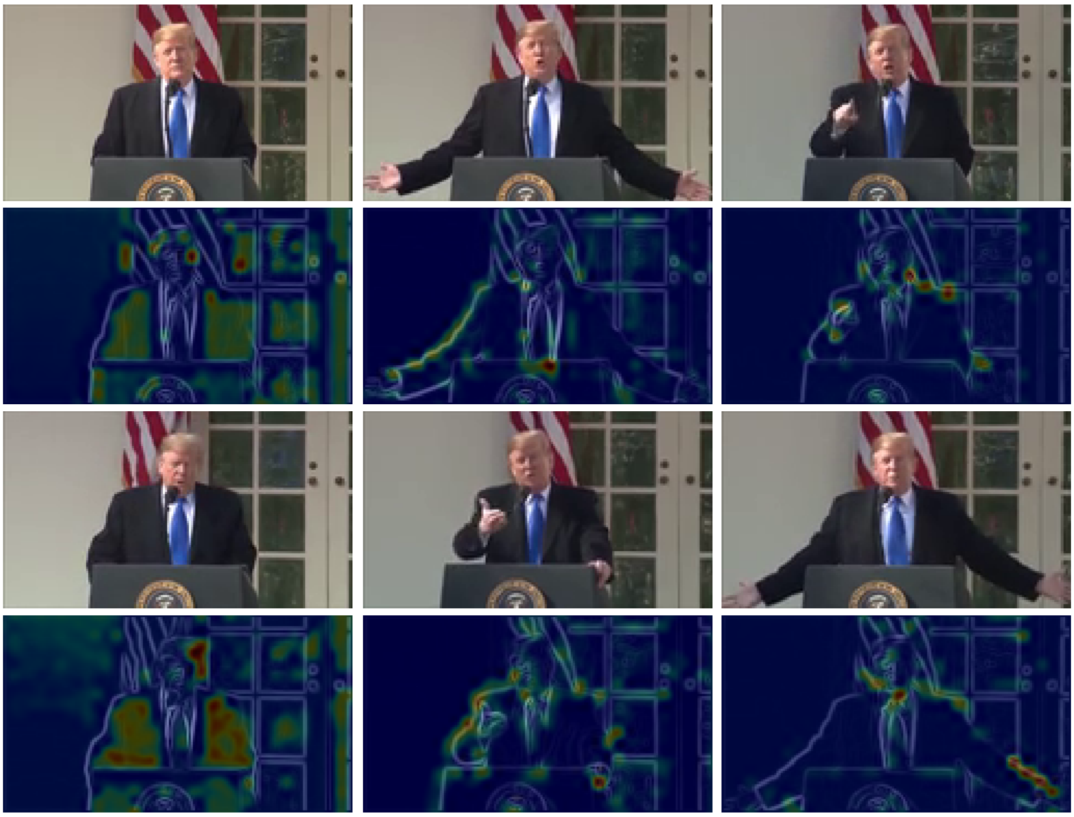


Figure 3.6: Example visualizations I. (FLTR): Neutral - Moderately Convincing - Very Convincing. The visualization shows that the neural network has learned to focus on the speaker's posture, hands, and contours to make its prediction. Due to the nature of our training data, the network hardly learned the person's features for barely annotated classes.

The saliency maps show that the network has learned to focus on the person, precisely on their postures and gestures. The background is primarily ignored and irrelevant for prediction (except for some background noise). In particular, the network follows the speaker's hands and face, which is consistent with the existing literature (cf. Sec. 2.3.1) as it states that gestures, gaze, and hand movements are significant indicators of perceived persuasiveness (Newman et al., 2016; J. Peters & Hoetjes, 2017). When predicting the *neutral* class, the network appears to look at every object in the frame (unlike the other two classes, where the network explicitly tracks the person's arms and hands). This is probably because the network cannot find any convincing markers, so every part of the frame is observed. These visualizations show a connection between the visual channel and subliminal persuasion, as well as the ability of neural networks to learn this connection, highlighting the effect of the persuasive power of nonverbal cues.

To examine the ability of the network to generalize (although it was trained on only one person), we also tested the prediction on several images of other politicians as visualized in Fig. 3.7.

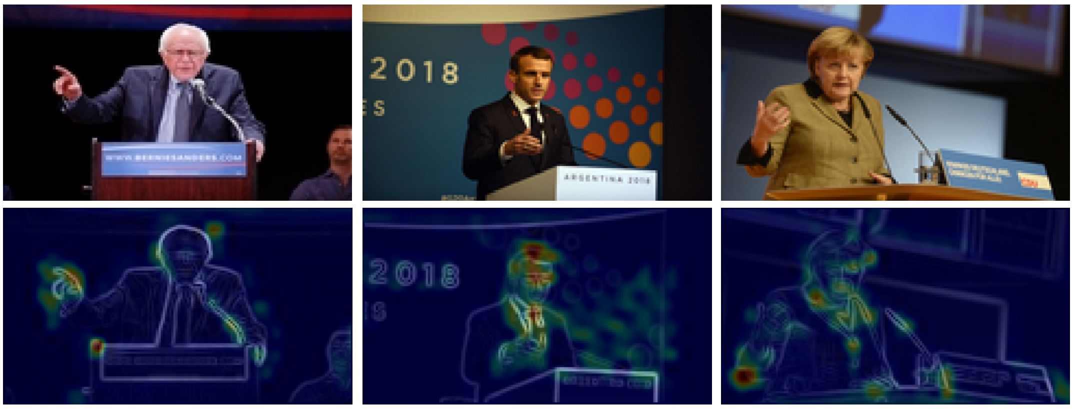


Figure 3.7: Example visualizations II. Visualizations of several other politicians with varying degrees of convincingness. (FLTR): Bernie Sanders [11] (predicted class: *very convincing*) - Emmanuel Macron [12] (predicted class: *very convincing*) - Angela Merkel [13] (predicted class: *moderately convincing*).

Despite variations in speakers and camera angles, the network consistently directs its attention toward hands and the general facial area. A closer look at Emmanuel Macron's picture reveals that the network appears to have learned to locate areas with skin-related colors (see background). The explanations show that the face contributed most to the decision *very convincing*. In addition, the tie played an inherent role in predicting the image as *very convincing*. Looking at the

[11] Modification of "*Election 2016: Bernie Sanders NYC Fundraiser Draws Campaign Supporters Who Are 'Feelin' The Bern*" by Michael Vadon. https://flickr.com/people/80038275@N00.

[12] Modification of "*Conferencia de Prensa - Presidente Emmanuel Macron - Día 2*" by G20 Argentina. https://www.flickr.com/photos/g20argentina.

[13] Modification of "*Speech by former German Chancellor Angela Merkel at the closing of the CDU party conference*" by CDU/CSU Bundestag Fraktion.

image of Bernie Sanders shows that the pointing finger was most important to classify the image as *very convincing*, while his face seems irrelevant. Looking at the image by Angela Merkel demonstrates a high relevance of the arm position, while other body parts, such as the face, seem to have only a low relevance.

3.2.3.2 Layer-wise-Relevance Propagation (LRP)

Next to Grad-CAM, we use LRP to analyze further the first convolutional layers of the network and what patterns they learned (see Sec. 2.1.2.2 for technical details). To create the LRP saliency maps for our model, we used iNNvestigate (Alber et al., 2019), a library that provides implementations of various analysis methods, including LRP. Fig. 3.8 shows example visualizations with red intensity indicating the input pixels' importance. Differently to Grad-CAM, LRP visualizations show more insights at a micro level, explicitly demonstrating a focus on the overall contours of the person. As before, we can see that the network has learned the person's spatial features, facial features, and gestures.

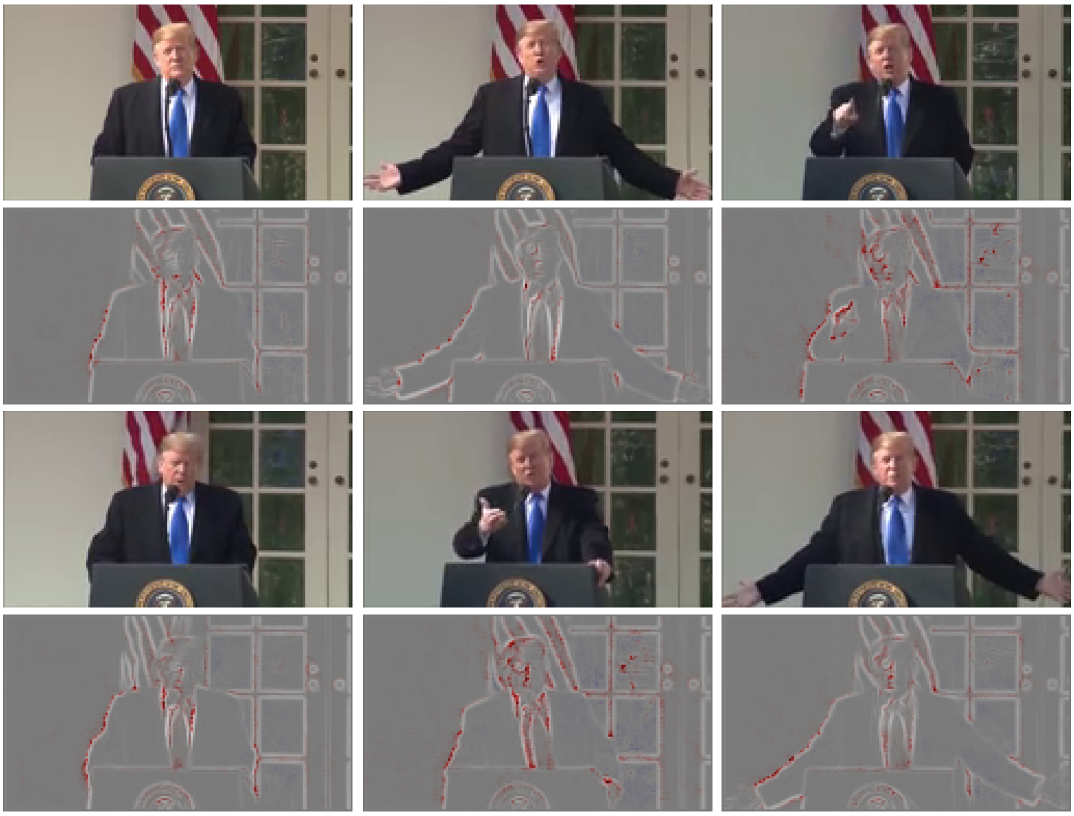


Figure 3.8: Example visualizations III. LRP visualizations (z+-rule) - (FLTR): Neutral - Moderately Convincing - Very Convincing.

3.2.4 Discussion and Conclusion

In the beginning, we argued that the literature primarily investigates *what* stimuli make a speaker persuasive but not *why* a speaker is persuasive, and thus, because people are usually unaware of this, a reflection bias is induced. To raise awareness of this, we analyzed an original political speech to highlight *why* the speaker is perceived as persuasive. We had annotators rate the speech based on their perceived persuasiveness by listening to and watching the video. We then trained a convolutional neural network on the visual input only to predict the degree of persuasiveness. We used XAI techniques, more specifically Gradient-weighted Class Activation Mapping (Grad-CAM) and Layer-wise Relevance Propagation (LRP), to highlight the most relevant sections at a *micro* and *macro* level. The results show that the network has learned to focus on the person, their gestures, and their contours.

Aligning with Alam et al. (2022), LRP produced more fine-granular saliency maps at a *micro* level with a focus on the person's contours. In contrast, Grad-CAM highlighted macro-level cues, exemplified by a pointing finger (Fig. 3.7) or exclusive focus on the overall speaker for the *neutral* class (Fig. 3.6) compared to LRP (Fig. 3.8). The macro-level emphasis on the overall person in Fig. 3.6 in the *neutral* class implies a broader focus that lacks specificity in identifying persuasive cues, showcasing a nuanced understanding by the network in capturing subtle gestures associated with persuasive communication.

In Sec. 1.3.1, we said analyzing the visualizations' fidelity is necessary. This refers to the correctness or faithfulness of the generated explanation (Huber et al., 2022; Mohseni et al., 2021). Within the context of this work, we argue that our network can be considered faithful because its focus aligns with findings from existing literature (e.g., Newman et al., 2016; J. Peters and Hoetjes, 2017), making it reasonable to assume that the explanations are correct and valid. This assumption is further supported by predictions from video frames featuring other politicians (Sanders, Macron, Merkel) which were not part of the training or validation set. With these, we can verify that the network has learned the person's gestures as persuasive indicators, as shown in Fig. 3.7 rather than just memorizing the data.

Some issues arise from the data distribution, comprising only three classes (*neutral*, *moderately convincing*, and *very convincing*). Therefore, the network could not learn any characteristics about what *not convincing* people look like. Using only one video, this outcome is not unexpected because, from a common-sense perspective, individuals may generally perceive another person as either more convincing or less convincing, but not both.

Further, regarding cultural implications, the current model's reliance on training data predominantly featuring white skin (see Fig. 3.7) suggests that one could explore cultural differences of perceived persuasiveness when extended to different cultures. In addition, while all annotators were German, the approach is extendable to annotators from different cultures to investigate cultural differences of perceived persuasiveness in detail.

To conclude, in this section, we investigated the research question **Q1.1**:

Q1.1 *"Can we effectively uncover behavior-based reflection bias in political speeches and provide satisfactory explanations?"*

The answer is: Yes, when keeping track of *overfitting*, we can train a neural network based on visual input only to predict perceived persuasiveness with sufficiently high accuracy effectively.

Using common XAI methods, we are able to generate explanations demonstrating a network's focus on persuasive cues, such as the speaker's contours, gestures, and hands. This aligns with known persuasive indicators from the existing literature (Newman et al., 2016; J. Peters & Hoetjes, 2017), proving the validity and fidelity of the generated explanations.

3.3 Highlighting Subjective Differences

This section answers research question **Q1.2**:

Q1.2: *"Can XAI contribute to highlighting and understanding subjective differences in persuasive cues in political speeches?"*

We train a single network for each annotator to allow for a nuanced comparison of what each annotator finds persuasive, emphasizing individual perspectives. This allows us to capture potential biases in the perception of persuasive cues to allow for highlighting *why* each annotator finds a speaker persuasive.

We extend the corpus to 30 videos to address the lack of data for *not convincing* data, providing a more comprehensive dataset for a better understanding and analysis of persuasive cues. Appx. A lists all video links. About 25% of the speakers were female.

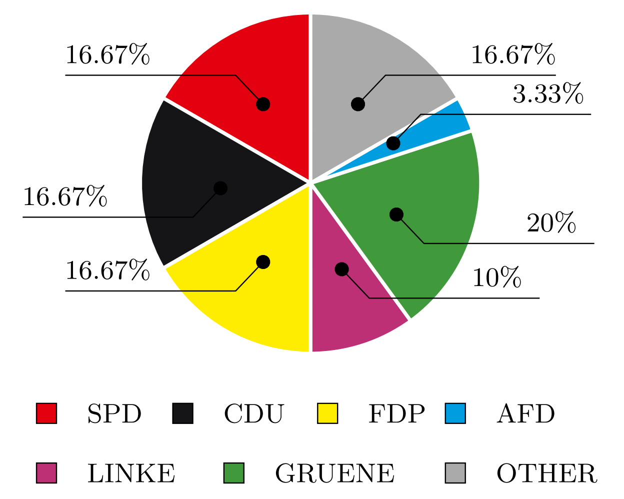


Figure 3.9: Speech distribution of party affiliations. Speeches were chosen to obtain a diverse data set.

The politicians were mainly from the German Bundestag. We tried to cover as many political directions as possible, and thus included speeches from the leading six German parties to increase the chance of obtaining *not convincing* data. Topics ranged from the COVID-19 pandemic and the German railway to agriculture and healthy food. Fig. 3.9 depicts the distribution of party affiliations.

Similar to 3.2.1, three annotators annotated the data with a sampling rate of 25 Hz. Only the first 3 - 5 minutes of all videos were taken to avoid over-representing certain speeches. The final employed data set consists of 150,000 video frames.

3.3.1 Video Frame Pre-Processing: Openpose

Further, we pre-process the video frames using *Openpose*. By pre-processing video frames with OpenPose, we filter out irrelevant information (such as background noise), offering a cleaner input that concentrates on the essential aspects of body language and facial expressions. *Openpose* is a multi-person detection tool for body, face, hand, and foot recognition on video frames and images (Cao et al., 2019; Simon et al., 2017). The tool uses *part affinity fields* defining vector fields encoding the limbs' position and orientation. For more technical details on *part affinity fields*, we refer to Cao et al. (2019).

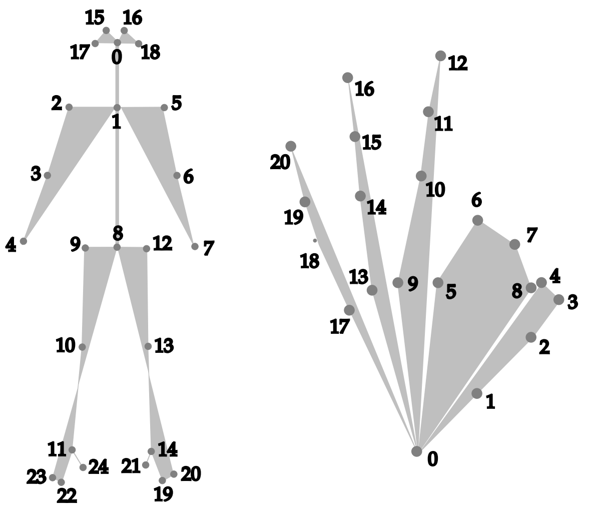


Figure 3.10: Openpose *Body_25* and hand key points. The *Body_25* format consists of 25 key points; there are 21 hand key points per hand (Figures adapted from Cao et al. (2019) and Simon et al. (2017)).

Openpose comes with two different pose key point mapping modes, which are *BODY_25* and *COCO*. While *COCO* allows for the detection of the joint positions of a person only based on 18 key points, *BODY_25* allows for detecting body rotation as well based on 25 key points with additional key points for the feet (+6) and the hips (+1). An extension of this model based on Simon et al. (2017) allows for also detecting 21 hand key points (see Fig. 3.10).

Each key point k is thereby defined as a triple $\left(u, v, c\right)$, where $\left(u, v\right) \in [0,1] \times [0,1]$ describes the image coordinates and $c \in [0,1]$ defines the confidence score. Consequently, the input signal $\boldsymbol{x}$ of the NN has 201 values of the following form:

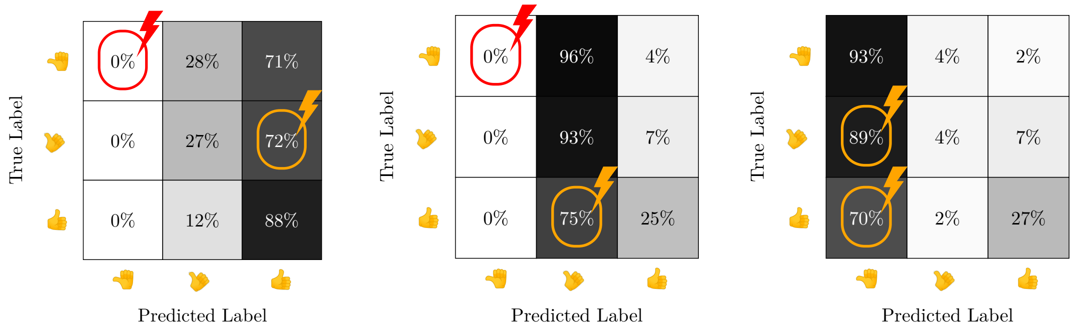
$$\boldsymbol{x} = \Big(\underbrace{\underbrace{u_0}_{x_0}\, , \underbrace{v_0}_{x_1}\, , \underbrace{c_0}_{x_2}}_{k_0}\, , \dots , \underbrace{\underbrace{u_1}_{x_3}\, , \underbrace{v_1}_{x_4}\, , \underbrace{c_1}_{x_5}}_{k_1}\, , \dots , \underbrace{\underbrace{u_{n-1}}_{x_{3(n-1)}}\, , \underbrace{v_{n-1}}_{x_{3(n-1)+1}}\, , \underbrace{c_{n-1}}_{x_{3(n-1)+2}}}_{k_{n-1}} \Big) \tag{3.2}$$

with n the number of key points. Because the speeches were held behind a podium, key points from leg and foot were not visible and omitted from the learning process. Therefore, the input vector has a total of $n = 54$. We kept the confidence values in the input to allow the network to ignore key points with lower confidence, which may be wrong with a very high chance, i.e., key points with low confidence value.

3.3.2 Network Architecture and Training

Contrary to 3.2.2, we use a standard fully-connected network with five layers because using Openpose does not rely on shape information (cf. Sec. 2.1.1.2). Each layer is succeeded by a *dropout* layer (Srivastava et al., 2014). *Dropout* is a simple technique to restrain the network from overfitting by simply dropping neurons, technically done by setting some weights to zero with a certain probability. This prevents the network from memorizing specific input vectors, i.e., overfitting.

The initial training results showed that the networks did not recognize certain classes at all (see Fig. 3.11), and primarily focused on one class only.



(a) Rater 1 (b) Rater 2 (c) Rater 3

Figure 3.11: Initial Training Results of each network showing that either classes were not recognized at all (red) or the classifications were wrong most of the time (orange)

Thus, we applied a hyper-parameter search based on the *random grid search* method for each of the three models, focusing on tuning the *number of neurons*, *dropout probability*, *number of hidden layers*, and *optimizer* (see Tab.3.3).

Table 3.3: Hyper-parameter range.

Parameter Name	*Min*	*Max*
Number of neurons	*128*	*8192*
Dropout probability	*0.0*	*0.5*
Number of hidden layers	*3*	*5*
Optimizer	SGD/ADAM	

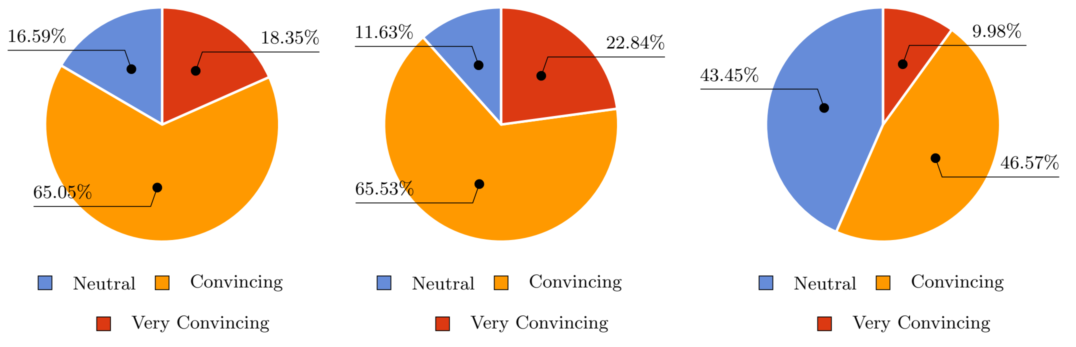
By utilizing a random grid search, we explored pre-defined parameter configurations. For example, we incremented the number of neurons by 128 rather than exhaustively testing every possible value. This approach was chosen based on the expectation that slight variations in these parameters are unlikely to significantly impact training results.

First, the ADAM optimizer outperformed every model trained on SGD (cf. Sec. 2.1.1.3). Dropout was found to be best at a level of 0.1. Concerning the neurons of the network, we found that rater one works best with inverted layer order compared to rater two and three as shown in Tab 3.4.

Table 3.4: Number of neurons per layer for each rater. (h_n) is short for the nth hidden layer.

	Neurons per Layer				
	h_1	h_2	h_3	h_4	h_5
`Rater 1`	*1024*	*1024*	*256*	*256*	*256*
`Rater 2`	*256*	*256*	*1024*	*1024*	*1024*
`Rater 3`	*256*	*256*	*1024*	*1024*	*1024*

A second step was taken to eliminate the unbalanced class distribution, preventing the network from focusing on only one class. As seen in Fig. 3.12, rater one and two tend to *convincing*, making learning from the two remaining classes *neutral* and *very convincing* difficult.



(a) Rater 1 (b) Rater 2 (c) Rater 3

Figure 3.12: The class distributions per annotator show an imbalance between classes.

To tackle this issue, we tested three imbalance learning techniques (Chawla et al., 2002; Lemaître et al., 2017):

- Undersampling
- Oversampling
- Synthetic Minority Oversampling Technique (SMOTE)

Undersampling is a technique that removes as many data items from all other classes so that all classes have equal cardinality, while *Oversampling* adds observations repeatedly. SMOTE, is a technique that synthetically generates samples by means of the *k-nearest* neighbors. *Undersampling* performed worst (accuracy 0.35), whereas SMOTE and *oversampling* achieved similar results (0.42). However, whether the generated data key points represent body postures using SMOTE remains questionable since the coordinates represent the body key points without any *part affinity* information that may be important for the synthetization process. Thus, we stuck to the basic *oversampling* technique. To tune the network independently of the data, we applied 10-Fold cross validation by splitting the video into twenty-seven training videos and three validation videos (see Tab. 3.5).

Table 3.5: 10-Fold cross validation scores of all ten folds for rater one.

Fold	1	2	3	4	5
Accuracy	*0.34*	*0.28*	*0.28*	*0.40*	*0.31*
Fold	**6**	**7**	**8**	**9**	**10**
Accuracy	*0.25*	*0.18*	*0.29*	*0.25*	*0.37*
Overall Accuracy	0.30				

As we can see, the models did not generalize well to data they were not exposed to during training. There are several explanations for this. On the one hand, some validation data are substantially biased towards a specific class, while some folds of the training data are uniformly distributed due to oversampling. On the other hand, the folds were predetermined: only three talks were used for validation [14], while the remaining 27 speeches were used for training. Within the validation data set, folds often contain triples $\left(u, v, c\right)$ that notably deviate from those encountered by the model during training due to different camera angles and speaker's position. A more extensive and varied data set would be required to

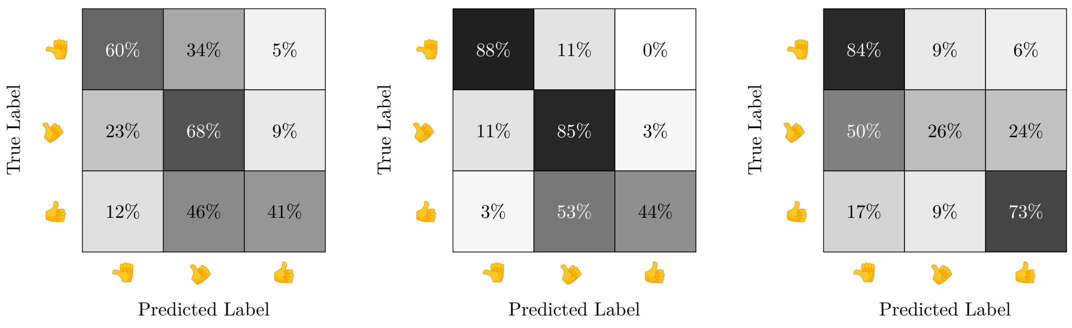
[14] They were not the same across different folds.

train models that determine persuasiveness in general. However, as our goal is to examine how individuals perceive persuasiveness rather than creating a model that can be used to predict persuasiveness generally, this is unnecessary.

For each rater $i \in \mathcal{I}$, a model $\mathcal{B}_i$ was trained to analyze the variations among them using the entire data set. Early stopping was applied after performing 10-fold cross-validation again to find the optimal number of training cycles (which turned out to be five), which prevented the models from overfitting and produced acceptable training results appropriate for the analysis (see Tab. 3.6 and Fig. 3.13).

Table 3.6: Training results of the final models for each rater.

		👎	👌	👍	Weighted Avg.
Rater 1	Precision	*0.28*	*0.75*	*0.60*	*0.66*
	Recall	*0.61*	*0.68*	*0.41*	*0.61*
	F1-Score	*0.39*	*0.72*	*0.49*	*0.63*
	Accuracy	0.61			
Rater 2	Precision	*0.65*	*0.82*	*0.77*	*0.79*
	Recall	*0.88*	*0.85*	*0.44*	*0.78*
	F1-Score	*0.75*	*0.84*	*0.56*	*0.77*
	Accuracy	0.78			
Rater 3	Precision	*0.59*	*0.71*	*0.34*	*0.62*
	Recall	*0.84*	*0.26*	*0.73*	*0.56*
	F1-Score	*0.70*	*0.38*	*0.47*	*0.53*
	Accuracy	0.56			



	👎	👌	👍
👎	60%	34%	5%
👌	23%	68%	9%
👍	12%	46%	41%

(a) Rater 1

	👎	👌	👍
👎	88%	11%	0%
👌	11%	85%	3%
👍	3%	53%	44%

(b) Rater 2

	👎	👌	👍
👎	84%	9%	6%
👌	50%	26%	24%
👍	17%	9%	73%

(c) Rater 3

Figure 3.13: Final training results of each network.

3.3.3 Inter-Rater Agreement Analysis

Next, we performed an inter-rater agreement analysis. For this, we compute Cronbach's *alpha* α_c based on Eq. 2.33 for each speech. Tab 3.7 summarizes the agreement scores, revealing a general disagreement with an average value of $\alpha_c = 0.473$ (=poor agreement).

Table 3.7: Inter-rater agreement of all 30 speeches.

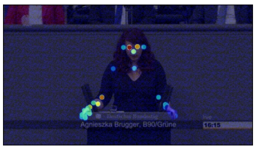
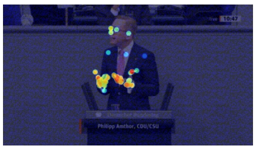
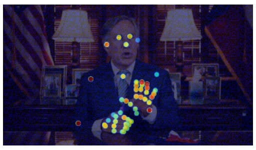
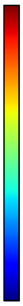
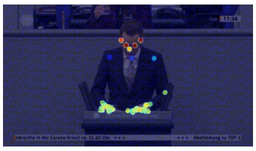
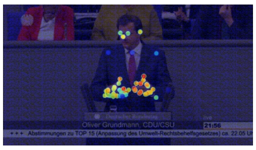
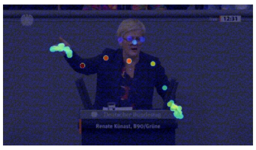
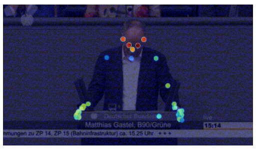
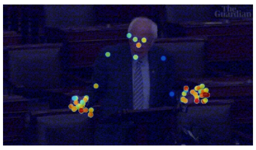
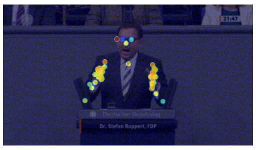
Video	Topic	α_c
1	Agriculture	*0.764*
2	Cellular Expansion	*0.432*
3-11	Covid-19	*0.471 (0.195 - 0.75)*
12	Social Code	*0.134*
23	Deutsche Bahn	*0.739*
14	Digital Pact	*0.155*
15	Eastern Partnership	*0.594*
16	Economy	*0.704*
17	F16-Bombers	*0.282*
18	Freedom of assembly	*0.224*
19	Freedom Protection	*0.26*
21	German Autobahn	*0.629*
24	German Bundestag	*0.623*
22-23	German Bundeswehr	*0.539 (0.358 - 0.719)*
24	Healthy Eating	*0.533*
25	Internet Filter	*0.381*
26	Innovation Principle	*0.586*
27	Kosovo	*0.497*
28	Online Education	*0.318*
29	Poverty Report	*0.508*
30	Veteran Support	*0.481*

Some speeches show a high agreement, such as speeches 1, 16 and 23, and others show a relatively low agreement, such as speeches 13 and 14. Overall, the perceived persuasiveness differs among the annotators. Compared to the first approach (Sec. 3.2), this is explainable by the fact that we had a variety of controversial topics that many people react differently to. The question is: Are there subliminal influences, and can we highlight these differences in the data using XAI, more precisely LRP? Again, we only visualized and analyzed video frames that were correctly classified, i.e., $\arg\max_k(\hat{y}_k) = \arg\max_k(y_k)$ and $\hat{y}_k \geq 0.50$ resulting in about 10,000 video frames for rater one and two, and 20,000 video frames for rater three.

To generate satisfactory explanation images $\hat{x}_{i,j} \in \hat{\mathcal{X}}_i$ for each rater $i \in \mathcal{I}$, we compute the relevance values for each key point using LRP. As each key point consists of three values $\left(\boxed{u}, \boxed{v}, \boxed{c}\right)$, we aggregate the relevance values $R_{\boxed{u_i}}, R_{\boxed{v_i}}$, and $R_{\boxed{c_i}}$ of each key point $\boxed{k_i}$ into a single relevance $\boxed{R_{k_i}}$ value by summing up and normalizing them. We consider a value of one most important, whereas a value near zero is considered least important.

Rater 1

Table 3.8: Images of rater one ($\hat{\mathcal{X}}_1$).

	✋	👋	👍	Relevance
1				most  least
2				
3				

First, we analyzed which cues rater one found persuasive based on the annotated data and identified trends. Then, we observed the saliency maps to see if we could also find those trends in the saliency maps. Example saliency maps for rater one are shown in Tab. 3.8.

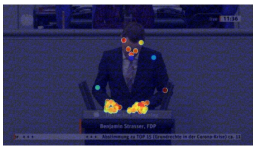
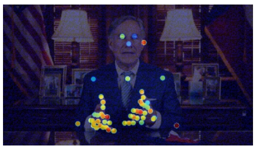
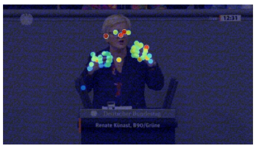
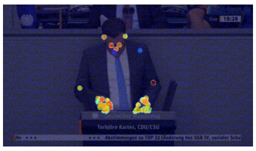
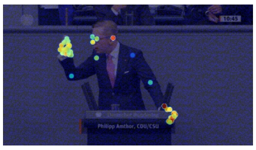
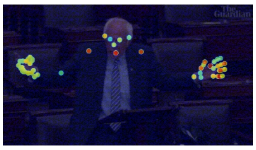
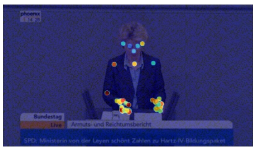
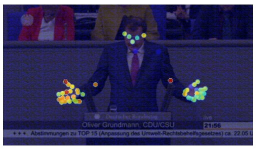
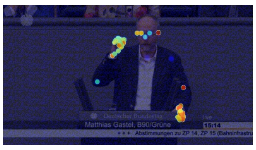
Trend 1: Although this pose was frequently seen in the *convincing* class, the network of rater one tends to classify frames with reading notes as *neutral*, suggesting that rater one might find reading notes as *neutral*. Looking at the saliency maps, we can see this trend that the network has learned to focus on the eyes for class *neutral*. Additionally, specific instances of clasped hands labeled as *neutral* can be found in the data, which implies an unbiased and objective view.

Trend 2: More dynamic gestures can be found within the class *convincing*. A speaker who makes active gestures and motions and does not constantly read from his notes appears prepared and confident to speak without overly relying on their notes. This is again highlighted in the saliency maps as we can see a stronger focus on the hands and less on the gaze for the *convincing* class compared to class *neutral*.

Trend 3: The *very convincing* class has a definite trend toward precise hand gestures and dynamic body language. Clear, dynamic body language makes a speaker seem very persuasive. Consequently, the network has learned to focus on shoulders and arms, indicating a big gesture.

Rater 2

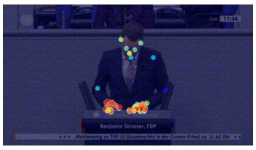
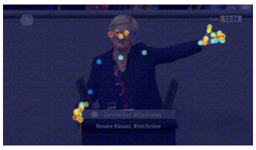
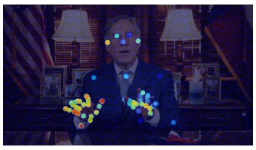
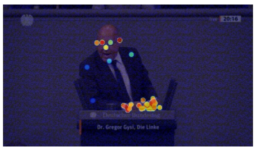
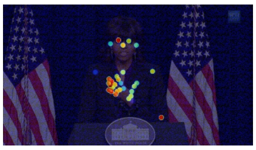
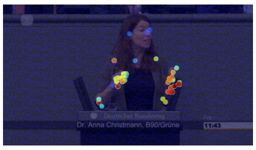
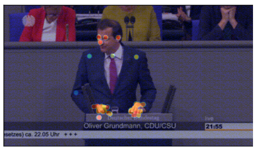
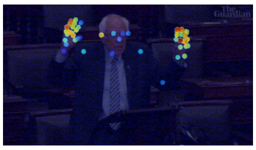
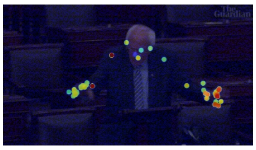
Table 3.9: Images of rater two ($\hat{X}_2$).

	🤚	👋	👍	Relevance
1				most
2				
3				least

The same applies to rater two. Note-reading with a bowed head was often labeled as *neutral* or *convincing*, which is yet again (**Trend 1**) highlighted in the saliency maps by a network's focus on the eyes. The *very convincing* class tends toward more dynamic, energetic poses and movements, while both the classes *convincing* and *very convincing* contain images with hand motions in general (**Trend 2**). Both raters one and two concur that engaging and active speakers are more convincing than those relying heavily on reading their speeches from notes. See Fig. 3.9 for examples of rater two's saliency maps.

Rater 3

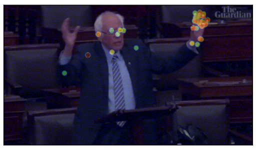
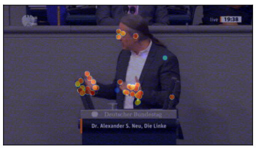
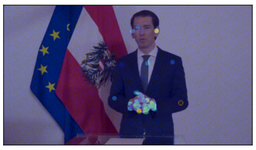
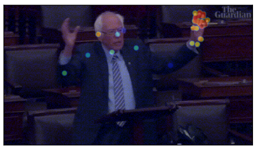
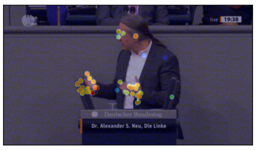
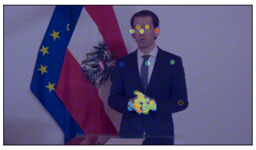
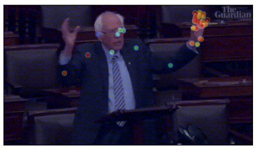
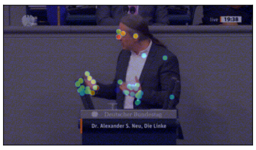
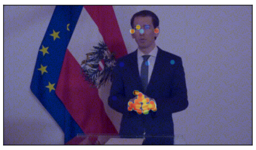
Table 3.10: Images of rater three ($\hat{X}_3$).

	🤚	👋	👍	Relevance
1				most
2				
3				least

Rater three (see Tab. 3.10) disagrees with the other raters. The reading notes and bowed-head poses can be seen once more in the *neutral* class (**Trend 1**). However, compared to raters one and two, these poses are far less common in the class *convincing*. While we can again identify a focus on the eyes for class *neutral*, there seems to be a specific new focus on the hands lying on the notes, which is revealed by the salience maps. The *convincing* class primarily consists of poses with overt body language. Nearly all images of the class *very convincing* contain expressive gestures and lively body language (**Trend 2**); thus, we can often see a focus on the hands in saliency maps. It seems that rater three was more influenced by body language than other elements, such as tone and speech content.

Rater Comparison We then run an automatic analysis over all video frames to find all triple pairs $(\hat{x}_{1,i}, \hat{x}_{2,j}, \hat{x}_{3,k})$ with $\hat{x}_{1,i} \in \hat{\mathcal{X}}_1, \hat{x}_{2,j} \in \hat{\mathcal{X}}_2, \hat{x}_{3,k} \in \hat{\mathcal{X}}_3$ and $i = j = k$ resulting in a total set of 4,539 triples. In other words, we selected all correctly classified video frames among all three annotators to compare them. Video frames that were only correctly classified for one or two annotators were omitted.

Table 3.11: Comparison of raters. FLTR: Very convincing, convincing, neutral.

	✋	👋	👍	Relevance
Rater 1				most
Rater 2				
Rater 3				least

Comparing the saliency maps of the raters (see Table 3.11 for examples), we can identify one trend again: The third rater's network puts a lot more focus on the hands for the *neutral* class, which is not surprising as previously we generally identified a stronger focus on gaze for rater one and two. This suggests that the gestures, the pose, and the perceived persuasiveness are related. As analyzed earlier, this trend is weaker for raters one and two. While we can again see the highlighted hands in the saliency maps for class *very convincing* across all raters in the salience maps, the trend that rater three pays more attention to the hands is somewhat reversed in the example above for class *convincing*. This is likely because of the particular body posture (gaze to the left), which suggests that rater three paid more attention to specific body language, such as directly facing someone, which the network seems to have learned.

3.3.4 Conclusion and Limitations

In this section, we explored research question **Q1.2**:

> Q1.2 *"Can XAI contribute to highlighting and understanding subjective differences in persuasive cues in political speeches?"*

Again, the answer is yes. While the training turned out to be difficult, the explanation results again show a clear tendency towards the importance of big gestures and hand movements. We highlighted a correlation between rater three's perceived persuasiveness and body language, which we identified as a clear difference between rater three and the others. Again, the networks' focus on the hands explicitly demonstrates the importance hand gestures play in persuasive contexts, aligning with existing research findings proving the validity of the presented approach.

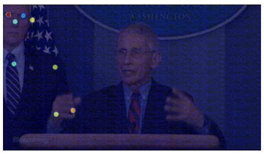
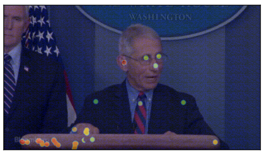
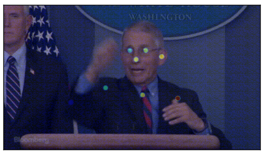
Thus, the approach presented here is a useful and promising tool for highlighting and raising awareness of these subliminal persuasive cues, which is one step closer to a deeper understanding of the subliminal effects of non-verbal cues. In contrast to studies that merely vary stimuli, our explanation-based method provides a more intuitive way to comprehend the tangible influence of these subtle cues, enabling people to directly witness the highlighted effects they have and become aware of it.

To conclude, we highlight some limitations worth mentioning for future research below.

Table 3.12: Limitations of the presented methodology and analysis.

Limitation	Description
Time-consuming annotation process	To highlight and compare reflection bias, subjective annotations from each person are required, which demands significant effort in data collection and annotation.
Time-consuming training process	Network optimization for each annotator is necessary due to unsatisfactory results from baseline networks, requiring substantial training and fine-tuning time. Therefore, automation through research into autonomous network architecture generation tools is recommended in future work.

…continued

Limitation	Description
Data set still not satisfactory	Despite including diverse political directions, the data set only consists of talks perceived as persuasive, hindering an analysis of non-persuasive cues. Future research should, therefore, focus on improving data diversity.
Pre-processing of input data generates errors	OpenPose's detection of additional people complicates the training, especially in speeches with numerous bystanders as seen below: **Wrong person:**  **Object detected:**  **Missing hands:**  Thus, further pre- and post-processing filters and manual filtering of irrelevant key points are essential for accurate training results. Considering the better performance of the initial CNN-based method described in Sec. 3.2, it is recommended to employ CNNs without pre-processing of input data in future work.
Time-consuming analysis process	The analysis requires significant time, as each output image must be manually reviewed to identify data trends. Employing XAI techniques can aid in highlighting the network's decision-making process, but individual examination of output images remains necessary for identifying and comparing specific persuasive cues.

3.4 Key Points & Summary

Within this chapter...

- ...we demonstrated that we can generate satisfactory explanations using XAI methods to uncover the behavior-based reflection bias.
- ...we showed that XAI can contribute to highlighting subjective persuasive cues (gestures) and comparing the differences among individuals.

This chapter explored an approach to highlighting persuasive public speech indicators using XAI techniques and addressed research questions **Q1.1** and **Q1.2**:

> Q1.1 *"Can we effectively uncover behavior-based reflection bias in political speeches and provide satisfactory explanations?"*

> Q2.2 *"Can XAI contribute to highlighting and understanding subjective differences in persuasive cues in political speeches?"*

To answer **Q1.1**, we trained a convolutional neural network to predict perceived persuasiveness based on the visual input of a single-video data set annotated concerning the person's persuasiveness. We then applied explainable AI techniques, namely Gradient-weighted Class Activation Mapping (Grad-CAM) and Layer-wise Relevance Propagation (LRP), to highlight relevant areas of the video frame used by the network for predicting the degree of persuasiveness to raise awareness of the stated effect of subliminal persuasive cues and show *why* a speaker is perceived as persuasive. The results show that the network has learned to focus on the person, their contours, and the face and hands. The latter is particularly interesting as it shows the importance of hand movements, in line with existing literature (Newman et al., 2016; J. Peters & Hoetjes, 2017), which demonstrates that we are able to create satisfactory explanations and make them visible using XAI.

In the second part, we explored whether or not these explanations can contribute to highlighting subjective persuasive cues. In the first part, we found that the networks sometimes focused on the background, so we pre-processed the images and computed OpenPose features to be used as input data. We extended the data set to 30 videos to obtain data for *not convincing*, which was not present in the first. To answer **Q1.2**, we trained a fully-connected neural network for every annotator and generated explanations of what the network looked at using LRP. While the training turned out to be difficult, the results of the explanations again show a clear tendency towards the importance of big gestures and hand movements. We also highlighted the strong correlation between rater three's perceived persuasiveness and body language, which identified a clear difference between rater three and the others.

As the networks have successfully learned to prioritize hands as a significant indicator of perceived persuasiveness, aligning with existing research findings, and we found differences among annotators, our approach presented here is a feasible tool for raising awareness of the subliminal influence of persuasive cues. In contrast to studies that merely investigate *what* stimuli make someone persuasive, our explanation-based method shows *why* someone is perceived as persuasive and provides a more intuitive way to comprehend the tangible influence of these subtle cues on an individual level, enabling individuals to witness the highlighted effects they have directly.

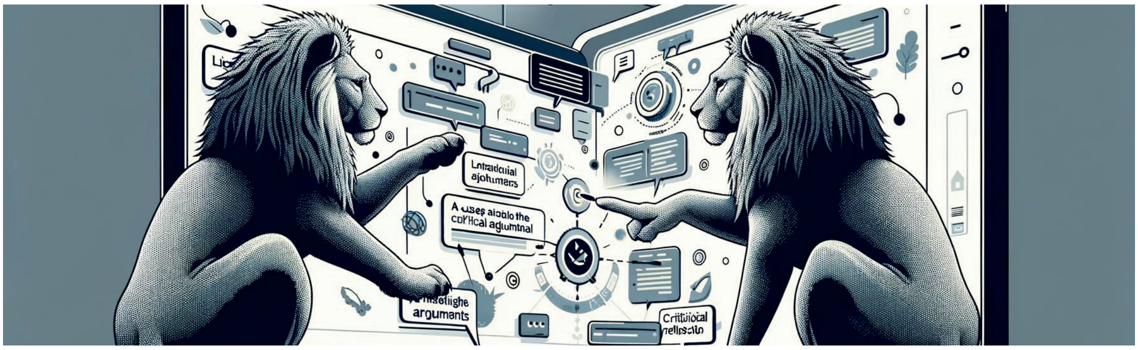
3.5 Relevant Publications

- Weber, K., Tinnes, L., Huber, T., Heimerl, A., Pohlen, E., Reinecker, M.-L., & Andé, E. (2020c). Towards Demystifying Subliminal Persuasiveness: Using XAI-Techniques to Highlight Persuasive Markers of Public Speeches. *Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS)*, 113–128

- Weber, K., Tinnes, L., Huber, T., & André, E. (2023b). Exploring the Effect of Visual-Based Subliminal Persuasion in Public Speeches Using Explainable AI techniques. *Proceedings of the 25th International Conference on Human-Computer Interaction (HCII)*, 381–397

Chapter 4

A System to Engage Users in the Critical Reflection of Arguments



"The mind is its own place, and in itself can make a heaven of hell, a hell of heaven.." (John Milton)

For notation of arguments, read Sec. 2.1.3

This chapter describes the system architecture and interface of the intelligent agent.

The system architecture presented in this chapter was previously published by the author in a similar form in peer-reviewed papers (Aicher et al., 2023, 2024; Weber et al., 2023a, 2024).

In this chapter, we describe the interface and architecture along with the components of our intelligent agent, namely Building Engaging Argumentation

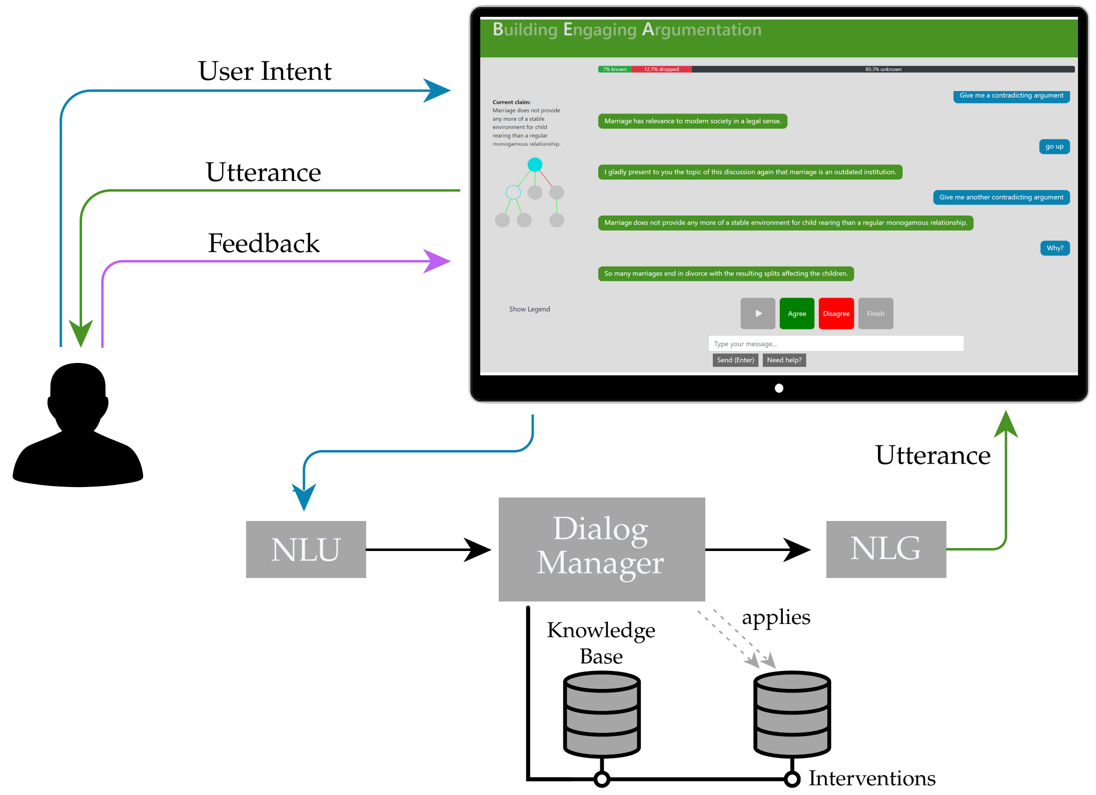
(BEA), which allows a human user to explore pro and con arguments on a predefined topic. The user can request the *pro* and *con* arguments one after the other by navigating *up* and *down* in the argument graph of a particular topic. In the scope of this thesis, this is *Marriage is an outdated institution* (Langhammer, 2018) based on the dataset by idebate (2022) encoded as acyclic graph (Langhammer, 2018). The choice of the topic is intentionally provocative to encourage participants to engage deeply and critically with the available arguments. This topic was selected because its dataset is sufficiently large, balanced in terms of argument stance (pro/con), of high quality, and has depth in arguments. Depth means that arguments have multiple layers of support and attack arguments, enhancing the engagement and complexity of the interaction.

To enable seamless communication between users and the system, we employ an integrated Natural Language Understanding (NLU) framework (Abro et al. (2022), see Sec. 4.1.2) to map the *user intent* to one out of nine available *speech acts* (see Sec. 4.1.1). The *dialogue manager* (see Sec. 4.1.1) processes the speech act, employing the arguments graph from the *knowledge base* to select the next argument based on the user intent, parsed by the NLU component. The *intervention strategy* is employed by the dialogue manager to generate an intervention utterance in case of monitored non-reflective argument exploration. The Natural Language Generation (NLG) component (see Sec. 4.1.3) generates a *textual* or *speech* response.

When interacting with conversational agents, the user's engagement and motivation are important factors that highly influence the success or failure of such a mixed team. Likewise to human-human interaction, the way arguments are presented may influence the user's willingness to engage in a critical reflection. To maintain the users' trust and satisfaction, the users' perception of the respective system is an important factor. Virtual agents have been proven to enhance users' interest, enjoyment, and intrinsic motivation aspects in various contexts (Miao et al., 2022; Qiu et al., 2021). Therefore, we equip the agent with two different presentation modalities (*chat* vs. *embodied*) to analyze if this co-variate affects the success of the intervention strategies.

4.1 System Components

The system's architecture (Fig. 4.1) consists of 1) an argumentative dialogue manager, 2) an NLU component, and 3) an NLG component. Below, we give a more detailed overview of the components.



Current claim: Marriage does not provide any more of a stable environment for child rearing than a regular monogamous relationship.

Give me a contradicting argument

Marriage has relevance to modern society in a legal sense.

go up

I gladly present to you the topic of this discussion again that marriage is an outdated institution.

Give me another contradicting argument

Marriage does not provide any more of a stable environment for child rearing than a regular monogamous relationship.

Why?

So many marriages end in divorce with the resulting splits affecting the children.

Figure 4.1: Overview of the system BEA: A user interacting with the intelligent agent via chat input/output. The user input (blue) is displayed on the right; the agent output is displayed on the left side of the chat view.

4.1.1 Argumentative dialogue Manager

The argumentative dialogue manager makes use the system's knowledge base consisting of the argument tree (argument components L_t, see Sec. 2.1.3) . Langhammer (2018) encoded the topic as ontology, which comprises 73 argument components ($\varphi_i \in L_t$) along with the relation ($\rightarrow$) between them to build an acyclic directed graph as elaborated in Sec.2.1.3 (see Fig. 4.2).

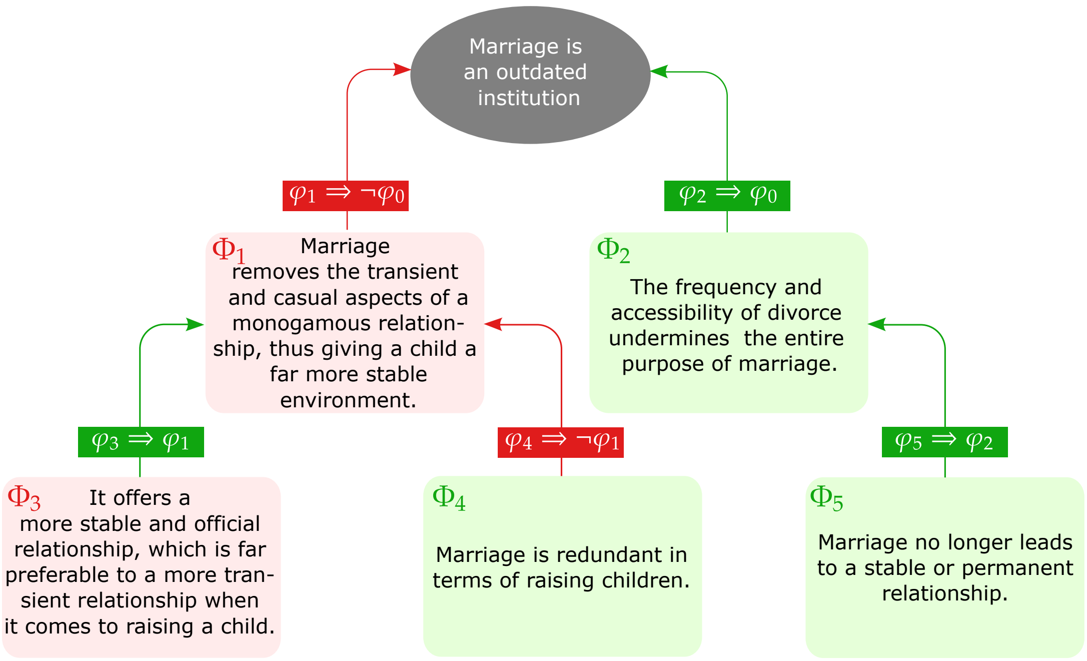

Marriage is an outdated institution
$\varphi_1 \Rightarrow \neg\varphi_0$
$\varphi_2 \Rightarrow \varphi_0$
Φ_1 Marriage removes the transient and casual aspects of a monogamous relationship, thus giving a child a far more stable environment.
Φ_2 The frequency and accessibility of divorce undermines the entire purpose of marriage.
$\varphi_3 \Rightarrow \varphi_1$
$\varphi_4 \Rightarrow \neg\varphi_1$
$\varphi_5 \Rightarrow \varphi_2$
Φ_3 It offers a more stable and official relationship, which is far preferable to a more transient relationship when it comes to raising a child.
Φ_4 Marriage is redundant in terms of raising children.
Φ_5 Marriage no longer leads to a stable or permanent relationship.

Figure 4.2: Excerpt of the dataset *Marriage is an outdated institution* based on the ontology by Langhammer (2018) showing atomic argument components φ_i, φ_j (pro vs. con) and their relation (support vs. attack) $\varphi_i \Rightarrow \_\varphi_j$.

The dialog manager handles the dialogue flow between the agent and the user using a communication language L_c and ensures logical consistency throughout the dialog. Similar to dialogue games (see Sec. 2.1.3), the communication language L_c consists of speech acts that the agent and the user can use to express their intents and communicate. Within this chapter, we define the set speech acts as shown in Tab. 4.1:

Table 4.1: Communication language L_c consisting of nine speech acts.

Speech Act	Description
Agent	
$argue(\varphi_i \rightarrow \varphi_j)$	*Present argument* $\varphi_i \rightarrow \varphi_j$
$jump\_to(\varphi_i)$	*Jump to argument component* φ_i
$intervene$	*Ask for a challenger argument*
User	
$why_{pro}(\varphi_i)$	*Ask for a supporting component* φ_j *with* $\varphi_j \rightarrow \varphi_i$

$why_{con}(\varphi_i)$	*Ask for an attacking component* φ_j *with* $\varphi_j \rightarrow \neg\varphi_i$
$level_{up}$	*Move level up*
$agree(\varphi_i)$	*Feedback to agree with a statement* φ_i
$disagree(\varphi_i)$	*Feedback to disagree with a statement* φ_i
$confirm/reject$	*Confirm/Reject intervention*

To ensure proper communication, the argumentative dialogue manager sets up pre-defined *protocol rules* that ensure *turn taking*, *allowed speech acts* (e.g., if an argument Φ_i is a leaf node, $why_{pro}(\varphi_i)$ is not allowed; if the user requests a new argument (why_x), the dialogue manager selects a random argument from all legal arguments of the *knowledge base* fitting the requested relation $x \in \{pro, con\}$). An argument $\Phi_i \in Args$ is considered legal if its conclusion supports the evidence of any argument that has been presented to the user before, e.g., if $\Phi_2 \Rightarrow \Phi_1$ was presented, then any argument Φ_i with $\Phi_i \Rightarrow \Phi_2$ or $\Phi_i \Rightarrow \neg\Phi_2$ is legal.

4.1.2 Natural Language Understanding (NLU)

The system has a chat-based input field where users can freely type in their requests for a natural conversation. An integrated natural language understanding framework (NLU) (Abro et al., 2022) is used to map this user input to the available speech acts (see Tab. 4.1). The employed NLU is based on an intent classifier model consisting of a Bidirectional Encoder Representations from Transformers (BERT) Transformer Encoder and a bidirectional Long Short-Term Memory (LSTM) classifier (Abro et al., 2022). BERT uses a technique based on Transformers introduced by Vaswani et al. (2017). Refer to Abro et al. (2022), Reimers and Gurevych (2019), and Vaswani et al. (2017) for technical details.

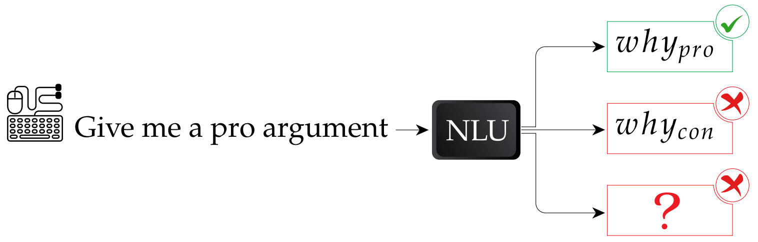


Figure 4.3: Sketch of NLU: The input sentence *Give me a pro argument* is mapped to one of the available speech acts (here: why_{pro}).

The NLU framework identifies and returns the most probable speech act based on the user input, which is then passed to the argumentative dialogue manager (see Sec. 4.1.1). Fig. 4.3 illustrates the mapping process.

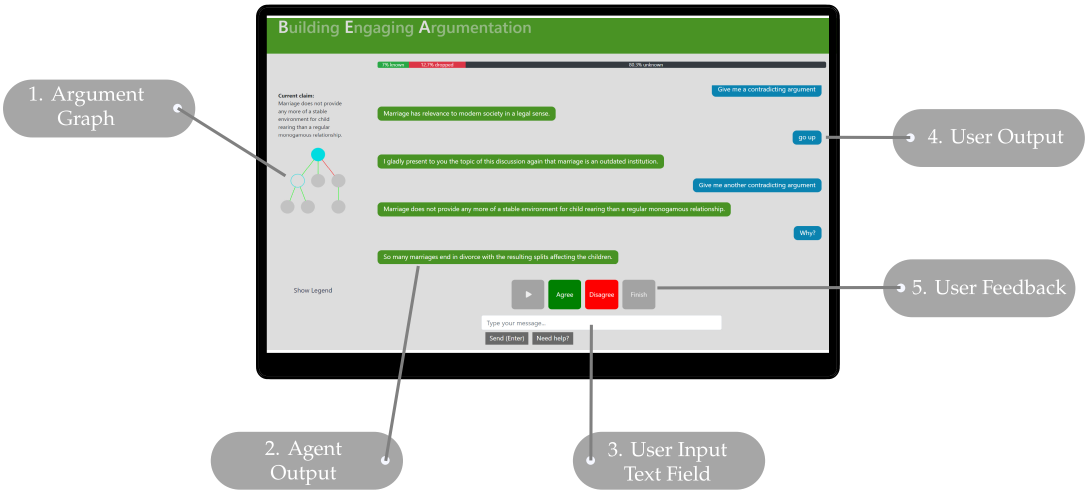
4.1.3 Natural Language Generation (NLG)

The system provides two output modalities: a *textual* or *spoken response*. The NLG is based on the textual surface text of the argument components $\varphi_i \in L_t$. The sentences were manually modified regarding grammatical syntax to form stand-alone utterances, which serve as a template for the respective system response. A list of natural language templates for each speech act is also defined. The explicit formulation is chosen randomly from this list during the response generation. To assist users in structuring the argumentative discourse, the system employs strategies such as resuming previous dialogue threads with statements like *Let us return to the previous argument, that . . .* and explicitly verbalizing the connections between arguments, as in *This claim is supported by the argument that. . .* These utterances do not express a specific stance but rather clarify the argumentative structure of the discourse.

Note that this NLG section only refers to the presentation of the argument components and not to the intervention mechanism and strategies of its linguistic style, which is described in detail in Ch. 6.

4.2 Graphical User Interface

The user interface consists of four components: 1) the graphically displayed argument structure, 2) the agent output, 3) the user input text field, 4) the user output, and 5) the user feedback buttons. (see Fig. 4.4).



Building Engaging Argumentation

7% known 12.7% dropped 80.3% unknown

Current claim:
Marriage does not provide any more of a stable environment for child rearing than a regular monogamous relationship.

Give me a contradicting argument

Marriage has relevance to modern society in a legal sense.

go up

I gladly present to you the topic of this discussion again that marriage is an outdated institution.

Give me another contradicting argument

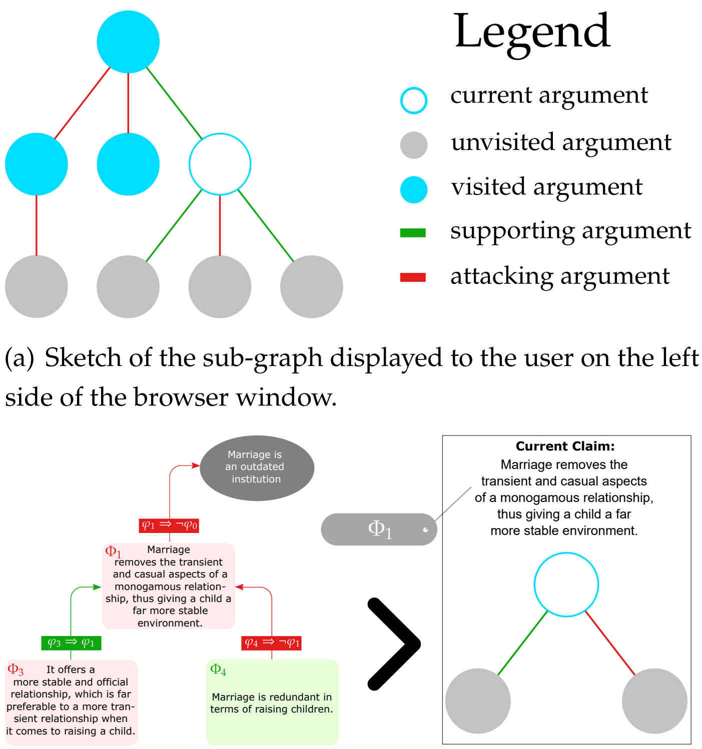
Marriage does not provide any more of a stable environment for child rearing than a regular monogamous relationship.

Why?

So many marriages end in divorce with the resulting splits affecting the children.

Figure 4.4: User interface elements: 1) the argument graph, 2) the agent output, 3) the user input, 4) the user output, and 5) the user feedback buttons.

The interface provides a *text* (chat) input where users can freely type in their requests to express their intent, initiating conversations with the agent. To gauge user agreement or disagreement with the presented arguments, the system includes two buttons (*Agree* and *Disagree*), which users can press at any point during the interaction.


Legend
○ current argument
● unvisited argument
● visited argument
🟩 supporting argument
🟥 attacking argument
Marriage is an outdated institution
$\varphi_1 \Rightarrow \neg\varphi_0$
Φ_1 Marriage removes the transient and casual aspects of a monogamous relationship, thus giving a child a far more stable environment.
$\varphi_3 \Rightarrow \varphi_1$
$\varphi_4 \Rightarrow \neg\varphi_1$
Φ_3 It offers a more stable and official relationship, which is far preferable to a more transient relationship when it comes to raising a child.
Φ_4 Marriage is redundant in terms of raising children.
Φ_1
Current Claim:
Marriage removes the transient and casual aspects of a monogamous relationship, thus giving a child a far more stable environment.

(a) Sketch of the sub-graph displayed to the user on the left side of the browser window.

(b) Transformation of the graph with claim Φ_1 as root from the argument graph excerpt from Fig. 4.2 to the displayed sub-graph with the text of claim Φ_1 written above it.

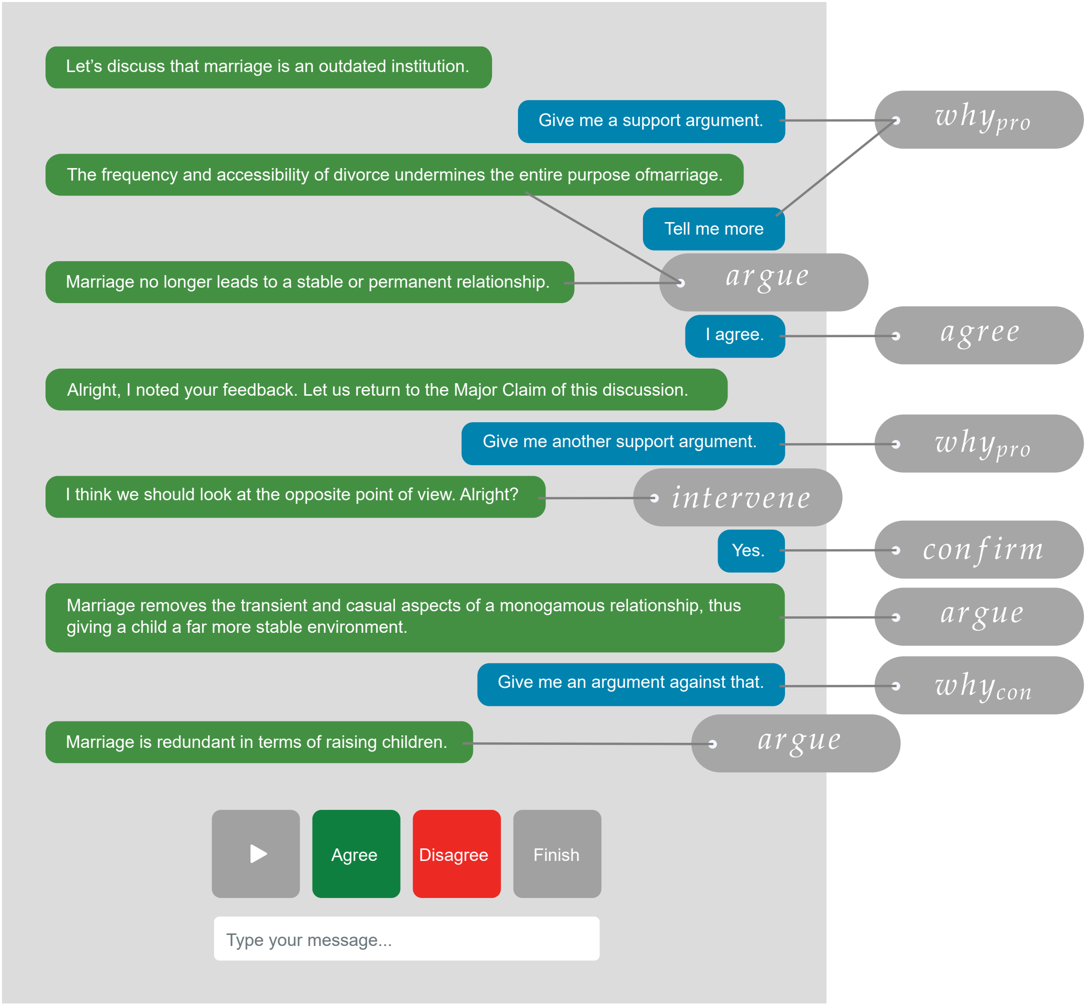
Figure 4.5: Sub graph: Sketch and legend (**Fig. a**), and example transformation (**Fig. b**).

Information about the current claim φ_i with $\varphi_i \Rightarrow \_\varphi_0$, and its underlying argument structure is visually depicted to users through a graph located on the left side of the browser window, aiding them in understanding the context and relations between different arguments, as illustrated in Fig. 4.5(a). The user's current position is depicted with an *outlined* turquoise node, the already discussed arguments are shown in *solid* turquoise, and unheard arguments in *gray*. The edges between the nodes show a *supporting* relation in *green* and an *attacking* relation in *red* color. The sub-graph is generated based on the ontology and argument graph representation. As an example, Fig 4.5(b) sketches the transformation of the graph with claim Φ_1 as root from the argument graph excerpt seen previously in Fig. 4.2 to the displayed sub-graph.

As the allowed user speech acts strongly depend on the user's position in the argument graph, a help button at the bottom is displayed where the user can get information about possible input statements.

Users can choose whether to ask for a pro or con argument or how they want to navigate through the argument tree. It is noteworthy that intervention (see

Ch. 6) only takes place if the user input is mapped to why_{pro} or why_{con}. Fig. 4.6 sketches a sample dialogue between the user and agent based on the argument graph excerpt in Fig. 4.2.



Let's discuss that marriage is an outdated institution.

Give me a support argument.

The frequency and accessibility of divorce undermines the entire purpose ofmarriage.

Tell me more

Marriage no longer leads to a stable or permanent relationship.

I agree.

Alright, I noted your feedback. Let us return to the Major Claim of this discussion.

Give me another support argument.

I think we should look at the opposite point of view. Alright?

Yes.

Marriage removes the transient and casual aspects of a monogamous relationship, thus giving a child a far more stable environment.

Give me an argument against that.

Marriage is redundant in terms of raising children.

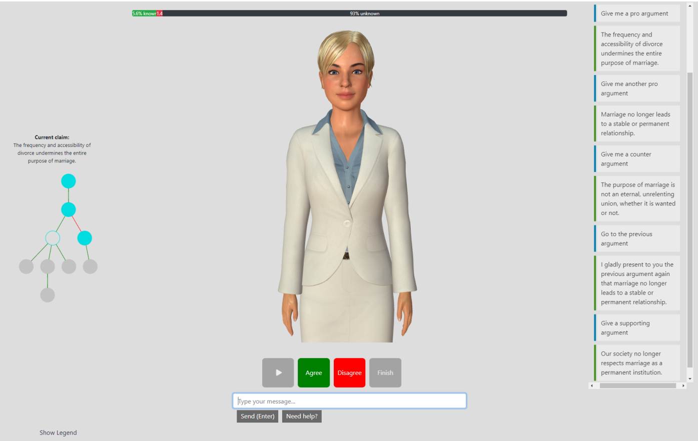
Figure 4.6: Conversation between the agent (left) and the user (right) about the topic *"Marriage is an outdated institution"* along with the *speech acts*.

4.3 Embodied Virtual Agent

Following challenge **C2.1** and to answer the research questions **Q2.3** and **Q2.4**, we equip the system with an embodied virtual agent [a] (see Fig. 4.7) using the Charamel 3D character rendering engine [b].

[a] **Embodied virtual agents** are "*computer-generated characters built to replicate the likeness of humans*" (Lloyd et al., 2020)

[b] http://www.charamel.com



The frequency and accessibility of divorce undermines the entire purpose of marriage.

Give me a pro argument

The frequency and accessibility of divorce undermines the entire purpose of marriage.

Give me another pro argument

Marriage no longer leads to a stable or permanent relationship.

Give me a counter argument

The purpose of marriage is not an eternal, unrelenting union, whether it is wanted or not.

Go to the previous argument

I gladly present to you the previous argument again that marriage no longer leads to a stable or permanent relationship.

Give a supporting argument

Our society no longer respects marriage as a permanent institution.

Figure 4.7: Embodied virtual agent interface. Instead of a chat-based agent (Fig. 4.4), the embodied virtual agent displays an embodied virtual agent. In addition, a dialogue history is displayed on the right-hand side to allow for re-reading previous arguments if needed. The rest of the interface, including the underlying system architecture, is the same as described in this chapter.

Research in marketing (Miao et al., 2022) or games has demonstrated the effectiveness of player avatars in enhancing interest, enjoyment, and various intrinsic motivational aspects (Qiu et al., 2021). Qiu et al. (2021) argue that conversational agents offer advantages over traditional graphical user interfaces due to their capacity for more human-like interactions. For example, Rebolledo-Mendez et al. (2008) showed that using avatars in Computer-Aided Instruction significantly boosts learner motivation. Moreover, the choice of investigating the impact of an embodied agent in our system is supported by the Persona Effect, which states the positive impact of employing life-like characters on user experience and interaction enjoyment (Lester et al., 1997).

Even though avatars display great potential, their effectiveness varies significantly (Miao et al., 2022). Lin et al. (2021) point out that discrepancies between online customer reviews and the purchase recommendations offered by the virtual salesperson affect customers' trust and their willingness to follow the avatar's recommendations. Furthermore, as the literature in this domain is very fragmented (Miao et al., 2022), it is unclear whether avatars have an impact on the course of argumentative debates (Blount et al., 2015). This raises the question of whether using an embodied agent as a counterpart in an argumentative discussion is perceived as motivating and engaging. This is especially important as we aim for the user to scrutinize arguments thoroughly to get a well-founded opinion.

In Ch. 3, we investigated the impact of non-verbal signals, such as gestures, on persuasiveness. The findings suggest that omitting gestures in our embodied agent could negatively affect its persuasiveness. Further, research indicates that gestures can increase the positive perception of embodied agents (Krämer et al., 2007), potentially making interactions more engaging. Thus, to enhance the realism and lifelikeness of our embodied agent, we incorporated beat gestures (see Fig. 4.8) synchronized with its speech.

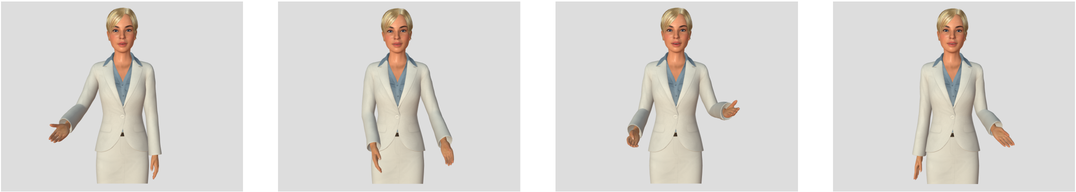


Figure 4.8: Example speak animations employed while speaking.

The character engine comes with a preset list of 25 beat gesture speak animations we can employ while the agent is speaking. From this list, an appropriate gesture is selected for every output to ensure a natural interaction. Once the character engine sends a speak stop event, all active animations are halted immediately, transitioning the agent smoothly into an idle animation.

4.4 Key Points & Summary

Within this chapter...

- ... we described the architecture of our argumentative dialogue system capable of fostering reflective thinking by means of intervention strategies.
- ... we described the system components and user interface in detail.

In this chapter, we described the interface and architecture of our developed chat-based agent, which allows a human user to explore pro and con arguments on a pre-defined topic. Specifically, we chose the topic *Marriage is an outdated institution* encompassing 73 arguments organized in a tree-like structure. This topic was chosen due to its size, balanced ratio of pro and con arguments, high quality, and argument depth. Depth means that arguments have multiple layers of support and attack arguments, enhancing the engagement and complexity of the interaction.

We employed a customized NLU component to facilitate user interaction via free-text prompts. This component maps user input to predefined speech acts, such as why_{pro}, indicating a request for a supporting argument. Based on the identified speech act, the agent selects an appropriate argument from the knowledge base and presents it through a chat-like interface.

As the user's enjoyment and engagement play an important role in increasing the interest and motivation to interact with the agent more frequently, we also implemented an embodied agent. According to the literature, the embodied agent can enhance user engagement by providing a more natural interaction by presenting the arguments via speech and gestures. This interaction is expected to make the conversation more enjoyable, potentially increasing user interest and motivation. While we have yet to verify these effects within our specific context, the integration of the embodied agent aims to create a seamless and interactive experience, potentially encouraging users to delve deeper into the topic.

4.5 Relevant Publications

- Weber, K., Aicher, A., Minker, W., Ultes, S., & André, E. (2023a). Fostering User Engagement in the Critical Reflection of Arguments. *Proceedings of the 13th International Workshop on Spoken Dialogue Systems (IWSDS)*, 1–16
- Aicher, A., Weber, K., Minker, W., André, E., & Ultes, S. (2023). The Influence of Avatar Interfaces on Argumentative Dialogues. *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents (IVA)*, 1–8
- Aicher, A., Weber, K., André, E., Minker, W., & Stefan, U. (2024). BEA: Building Engaging Argumentation. *Proceedings of the 1st International Conference on Robust Argumentation Machines (RATIO)*, 1–17
- Weber, K., Hogh, N., Conati, C., & André, E. (2024). A Gaze into Argumentative Chatbots: Exploring the Influence of Challenger Arguments on Reflection and Attention. *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents (IVA)*, 1–10

Chapter 5

A Computational Metric for Reflective Engagement (RE)

"The greatest enemy of knowledge is not ignorance, it is the illusion of knowledge." (Stephen Hawking)

! For notation of arguments, read Sec. 2.1.3

ℹ This chapter answers research question **Q2.1**:

Q2.1: *"How can we formulate a computational metric for RE that is operationally feasible and can be programmatically implemented allowing the agent to guide the user's argument visitation focus?"*

In order to allow the agent to guide the user's focus towards *challenger arguments*, it requires a metric that indicates whether the user's focus is too confirmation-biased. In this section, we define a computational metric, namely Argument

Visitation Quotient (AVQ), which gauges the extent to which users explore arguments that challenge their own position (referred to as *challenger arguments*). This is done in three steps:

1. We calculate the user's focus as the ratio between visited pro and con arguments.
2. We then relate this ratio to the user's stance (i.e., pre-existing opinion) to obtain a score in the range [0,1], describing how focused the user is on *challenger arguments*. A score close to 0 indicates that the user is too biased and not focused on *challenger arguments*.
3. We develop a prediction model considering the user's agreement and disagreement towards the arguments to estimate the user's stance over time during interaction. This has three main reasons:
   - When exposed to *challenger arguments*, the user might change their opinion. Therefore, if the agent detects a change in the user's stance, it allows the user also to explore arguments supporting their previous opinion again, preventing reinforcement of the new *challenger arguments* in the same way.
   - In practical applications, asking for the user's stance can be disruptive and hinder the interaction flow. Thus, estimating the user's stance and tracking opinion changes without direct queries is beneficial.
   - Third, biases are often unconscious, making them difficult for users to self-report accurately.

5.1 Metric for Challenging Argument Exploration

ℹ This section was previously published by the author in a similar form in peer-reviewed papers (Aicher et al., 2024; Weber et al., 2023a, 2024).

We first define the user's visitation focus and then introduce the metric AVQ for challenging argument exploration, which takes into account the inverse proportion between the user's stance and their aggregated visitation focus of every argument $\Phi_i \in Args$. A set of arguments with same target argument Φ_i is denoted as $Args_{\Phi_{*\Rightarrow\_i}} \subseteq Args(\Phi_i)$. If it is in favor of stance +, it is denoted as $Args_{\Phi_{*\Rightarrow i}}$ and $Args_{\Phi_{*\Rightarrow\neg i}}$ elsewise. The set of all visited arguments Φ_j with target argument Φ_i is denoted as $Args^v_{\Phi_{*\Rightarrow\_i}} \subseteq Args_{\Phi_{*\Rightarrow\_i}}$.

We define the user's visitation focus for argument $\Phi_{*\Rightarrow\_i}$ based on visited pro and con arguments as follows:

$$focus_{\Phi_i} = \frac{\left|Args^{v}_{\Phi_{*\Rightarrow i}}\right| - \left|Args^{v}_{\Phi_{*\Rightarrow \neg i}}\right|}{\left|Args^{v}_{\Phi_{*\Rightarrow\_i}}\right|} \in [-1, 1] \tag{5.1}$$

It is easy to verify that if the user selects arguments of a certain stance, the focus shifts in the direction of the respective stance.

The deeper the user descends in the argument graph, the more evidence supporting or attacking a certain *Claim* is provided. This is a direct consequence of how the argument graph structure was generated by Langhammer (2018). To take that into account, we employ two weights introduced by Aicher et al. (2021a), that are $\omega_{d,k}$ at argument graph level k on the one hand, and a weight ω_{n,Φ_i} for the size of the sub-graph $Args(\Phi_i)$ of argument Φ_i on the other.

Thus, the hierarchical weight $\omega_{d,k}$ incorporates the argument depths into the metric and ensures that exploring lower levels results in a higher score (Aicher et al., 2021a). It is defined as (Eq. 5.2):

$$\omega_{d,k} = \frac{k}{\sum_{j=1}^{l_{\max}} j} \tag{5.2}$$

where k is the argument graph level that the weight is assigned to and $l_{\max}$ is the maximum level of the argument graph.

Example 5.1: Hierarchical weight

Take the previously seen graph excerpt from Fig. 4.2 as an example. The max depth in this graph is $l_{max} = 2$. With this, we can compute the weights for level one and two as follows:

$$\omega_{d,1} = \frac{1}{1+2} = \frac{1}{3} \qquad \omega_{d,2} = \frac{2}{1+2} = \frac{2}{3} \tag{5.3}$$

The other weight ω_{n,Φ_i} accounts for level size variations in the sub-graph beneath argument Φ_i. This weight is determined by calculating the ratio of the number of direct child descendants of argument Φ_i, i.e., $Args_{\Phi_{*\Rightarrow\_i}}$, to the total number of descendants in the sub-graph, i.e., $Args(\Phi_i)$, i.e., (Eq.5.4)

$$\omega_{n,\Phi_i} = \frac{\left|Args_{\Phi_{*\Rightarrow\_i}}\right|}{\left|Args(\Phi_i)\right|} \tag{5.4}$$

This weight ensures that visiting two arguments at the same level of two distinct sub-graphs $Args(\Phi_i)$ and $Args(\Phi_j)$ with $Args(\Phi_i) \cap Args(\Phi_j) = \emptyset$ are weighted differently if the sub-graph sizes differ substantially at lower levels. Fig. 5.1 illustrates this based on two sub-graphs $Args(\Phi_1)$ and $Args(\Phi_2)$ with different sizes at level three, where sub-graph $Args(\Phi_1)$ has twenty-two, $Args(\Phi_2)$ has five children. As a result, $\omega_{n,\Phi_1} < \omega_{n,\Phi_2}$. As before, because leaf nodes have no descendants, we set $\omega_{n,\Phi_i} = 0$ if Φ_i is a leaf node.

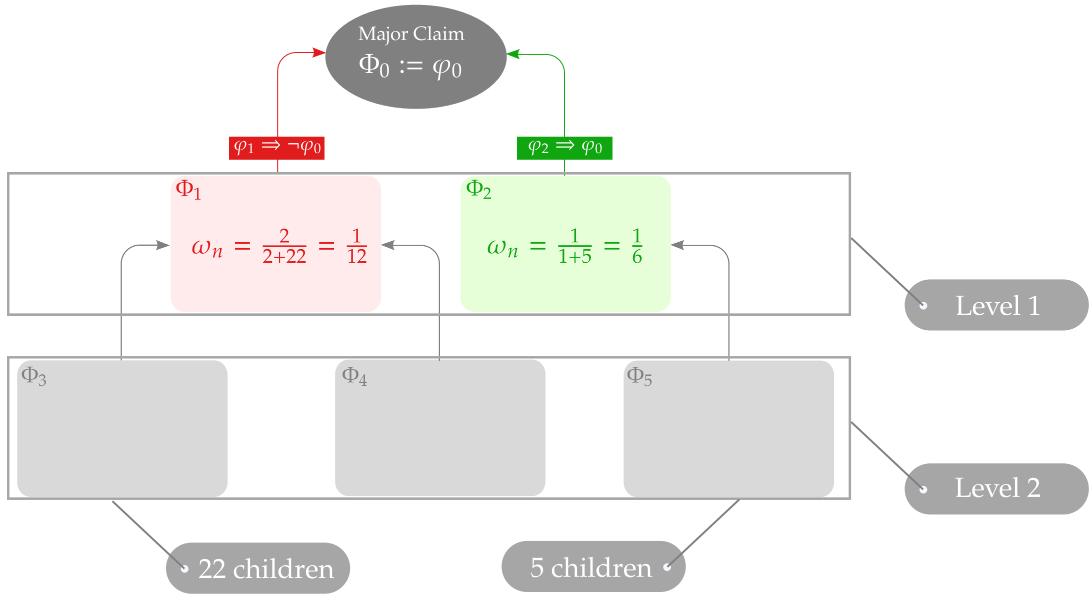

Major Claim
$\Phi_0 := \varphi_0$
$\varphi_1 \Rightarrow \neg\varphi_0$
$\varphi_2 \Rightarrow \varphi_0$
Φ_1
$\omega_n = \frac{2}{2+22} = \frac{1}{12}$
Φ_2
$\omega_n = \frac{1}{1+5} = \frac{1}{6}$
Level 1
Φ_3
Φ_4
Φ_5
Level 2
22 children
5 children

Figure 5.1: Sketch of the sub-graph weight: Two distinct sub-graphs $Args(\Phi_1)$ and $Args(\Phi_2)$ with unequal level sizes at level three (22 vs. 5 children). Consequently, $\omega_{n,\Phi_1} = \frac{1}{12}$, and $\omega_{n,\Phi_2} = = \frac{1}{6}$.

Let $depth : Args \Rightarrow \mathbb{N}$ be the respective depth of an argument in the argumentation structure. The total normalized user focus $F \in [0, 1]$ is defined by incorporating the defined weights with $W_{\Phi_k} = \omega_{d,depth(\Phi_k)+1}\, \omega_{n,\Phi_k}$ as follows:

$$F := \frac{\left(\sum_{\Phi_k \in Args} focus_{\Phi_k} \cdot W_{\Phi_k}\right) + 1}{2 \cdot \sum_{\Phi_k \in Args} W_{\Phi_k}} \tag{5.5}$$

The user's stance z_{Φ_0} and focus F should be inversely proportional. This is based on the assumption that users with a particular stance are likely to focus more on arguments that are in line with their stance. Users with a higher level of reflection tend to look at claims that support an opposite view as well following Paul (1990). Thus, the metric for challenging argument exploration, i.e., *challenger arguments*, denoted as AVQ, is defined as:

$$\mathrm{AVQ} = 1 - \left| z_{\Phi_0} - \left(1 - F\right) \right| \tag{5.6}$$

After inverting the total user focus, the difference between user stance and focus is taken to compute AVQ, i.e., the more the focus aligns with the user's stance, the lower the AVQ and vice versa. This approach ensures that if the user stance is positive (+), the agent suggests that the user choose more con arguments (challenger arguments of pro arguments) and vice versa.

From a psychological perspective, any model or metric should be validated to ensure they accurately capture and reflect the complex cognitive processes. Our metric aligns with established psychological theories, such as strong sense reflection (Paul, 1990), and other existing literature (Gelter, 2003; Mason, 2007). The metric is intended to be a mathematical description of the user's focus on opposing ideas. It yields a higher value when users choose more *challenger arguments*, indicating increased engagement with perspectives that contradict their opinions. Given its purposeful design and applicability within the context of this thesis, we are foregoing a separate validation. Additionally, reflection involves monitoring how deeply someone delves into the details of arguments. This includes considering *challenger arguments* and thoroughly examining its supporting arguments at lower levels in the graph, which is considered using higher weights for arguments at lower levels.

It is important to note that the proposed score is a simple approximation of RE. Ideally, users should also present their own arguments that the agent challenges. However, this aspect is not considered in this work, as only arguments given by the agent itself are utilized. Additionally, the metric does not take into account the extent to which users engage with the provided arguments or whether a deep understanding of the underlying arguments is achieved.

5.2 A fine-grained Model for User-Stance Estimation

ℹ This section was previously published by the author in a similar form in peer-reviewed papers (Weber et al., 2020b, 2020a).

The second author mainly developed the underlying dialogue model at Ulm University, while the prediction model was developed at Augsburg University.

To allow for a dynamic stance employed in the computational metric AVQ, we derive a model to predict the user stance during the interaction. While asking the user for their stance at the beginning (as we also do) provides a starting point, it results in a static metric that does not account for the changing nature of opinions. Frequently querying the user about their stance can disrupt the interaction flow. Additionally, biases are often unconscious, making them difficult for users to self-report accurately.

Therefore, we explore if we can reasonably predict the *user stance* by aggregating user feedback on whether an argument is *convincing* or *not convincing* throughout the interaction. To validate the feasibility of this approach, we conducted a user study with 48 participants to verify the predictive capability of the stance model.

5.2.1 Derivation of the Estimation Model

In order to define an estimation model for the user stance, referred to as *persuasive effectiveness*, we took inspiration from the work of Aicher et al. (2021b). Their recent research introduced an interactive system to aid users in their opinion-building process, allowing them to express their preferences and rejections towards arguments. The system then computes the user's preferences using Bipolar Weighted Argument Graphs (BWAGs) and a linear Euler-based restricted semantics (Amgoud and Ben-Naim, 2018; see Eq. 5.7). BWAGs are typically employed to compute the strength Ψ_{Φ_i} of arguments Φ_i in an acyclic directed argument graph (see Sec. 2.1.3), taking into account their own weight $\left(\omega_{\Phi_i}\right)$ and the strengths $\left(\Psi_{\Phi_{j\Rightarrow i}}, \Psi_{\Phi_{j\Rightarrow\neg i}}\right)$ of their supporting child arguments $\Phi_{j\Rightarrow i}$ and attacking child arguments $\Phi_{j\Rightarrow\neg i}$:

$$\Psi_{\Phi_i} = 1 - \frac{1 - {\omega_{\Phi_i}}^2}{1 + \omega_{\Phi_i} \cdot e^{\left(\sum_j \Psi_{\Phi_{j\Rightarrow i}} - \Psi_{\Phi_{j\Rightarrow\neg i}}\right)}} \tag{5.7}$$

However, for our research, the Euler-based restricted semantics is unsuitable because it defines arguments with a zero weight as invalid, causing them to have no impact on the strength of their target argument (Amgoud & Ben-Naim, 2018).

This is a result of the *Neutrality principle* stating that "*Worthless attackers/supporters do not affect their target*" (**Issue 1**), which is one of the twelve principles established by Amgoud and Ben-Naim (2018). Irrespective of the strengths assigned to its child nodes, an argument's strength remains constant if its weight is one (**Issue 2**). That means every argument found *convincing* (1.0) would always have a persuasive effectiveness of 1.0 as the feedback of all supporting and attacking arguments is eliminated. Thinking that further, all arguments from node level 3 and below would be useless. Thus, the computed strength would only depend on the arguments having a strength of 1.0 and the sub-graphs with a strength of 0.5 at node level 2, while eliminating the whole sub-graphs with a strength of 0.0 as illustrated in Fig. 5.2 (summarizing issue 1 and 2).

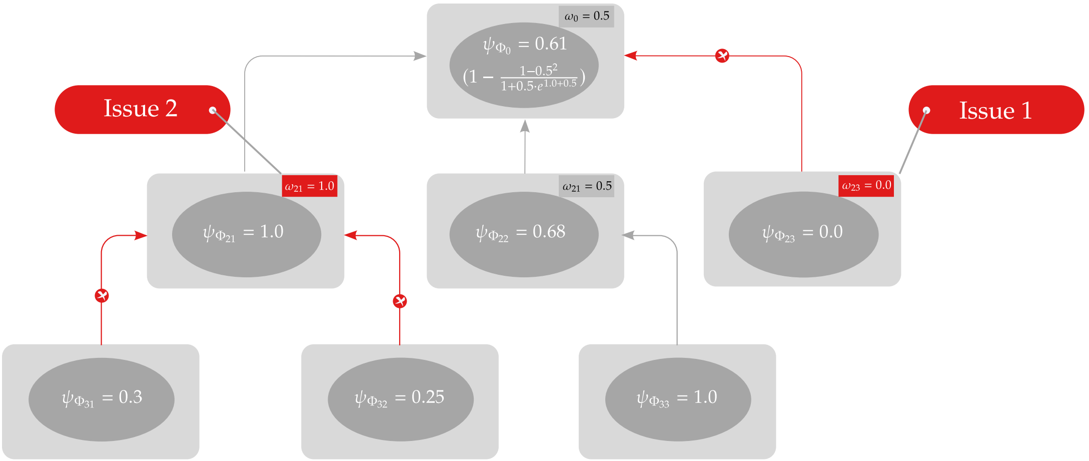


Figure 5.2: Sketch of limitations of Euler-based Restricted Semantics, making it unsuitable for our approach. (**Issue 1**) Since weight $\omega_{\Phi_{23}} = 0.0$, it is not included in the exponent and, thus, does not affect the target's strength Ψ_{Φ_0}. (**Issue 2**) Since $\omega_{\Phi_{21}} = 1.0$, the child nodes do not affect the strength $\Psi_{\Phi_{21}}$

Using this information to predict the user's stance correctly, we are still interested in arguments with zero weight as this affects the overall user stance of a topic negatively (or positively, depending on the argument's stance). Hence, we estimate the user stance by computing the argument's *persuasive effectiveness* Ψ_{Φ_i} recursively (Def. 5.1) based on how convincing each argument is found by assigning a weight ω_{Φ_i} to each argument Φ_i ranging from 0.0 (*not convincing*) to 1.0 (*convincing*). The intuition behind this estimation model is that a user who tends to agree more often with arguments that support the *Major Claim* is likely to have that stance in line with the *Major Claim*. If no feedback is given for an argument Φ_i, the default value of feedback ω_{Φ_i} is 0.5. For the sake of simplicity, $z_{\Phi_i} := z_{\varphi_i}$ throughout this thesis (see Def. 2.2).

Definition 5.1: Persuasive Effectiveness

Let $\omega_{\varphi_{i \Rightarrow j}} \in [0, 1]$ be the user feedback how convincing argument $\Phi_i \in Args$ is, and let n be the number of direct child arguments of Φ_i, then the persuasive effectiveness z_{φ_i} for the argumentative component φ_i is computed by its own weight and the strength values of its evidences as follows:

$$z_{\varphi_i} = \frac{\omega_{\varphi_i} + \sum_{\varphi_{j \Rightarrow \_i} \in L_t} \iota^{-1}\left(z_{\varphi_j}\right)}{1 + n} \tag{5.8}$$

where $\iota^{-1} : [0, 1] \Rightarrow [0, 1]$ defines the inverse function of z_{φ_j} as

$$\iota^{-1}(z_{\varphi_j}) = \begin{cases} z_{\varphi_j} & \text{if } \varphi_j \longrightarrow \varphi_i \\ 1 - z_{\varphi_j} & \text{else} \end{cases} \tag{5.9}$$

Using the effectiveness, we predict the user's current stance as follows:

$$user\_stance = \begin{cases} + & z_{\varphi_0} \geq 0.5 \\ \text{-} & else \end{cases} \tag{5.10}$$

It is worth mentioning that no interval is defined for an *unknown* stance. Instead, the effectiveness z_{φ_0} can be used to determine the confidence value of how sure the system is about the prediction by looking at how close the effectiveness is to the criterion value 0.5.

5.2.2 Evaluation Prototype

To evaluate the model, we designed a simple interface with an embodied virtual agent of the Charamel 3D character rendering engine [15] (see Fig. 5.3):

The agent's task is to talk about the topic *This hotel is worth a visit* (*Major Claim*, see Sec. 5.2.3) by giving pieces of evidence of arguments (components) that are either *for* or *against* the topic. We refrained from using the topic *Marriage is an outdated institution* from the overall system in Sec. 4 because we aimed for a topic that the user could not have an opinion about beforehand (see Sec. 5.2.3).

[15] http://www.charamel.com

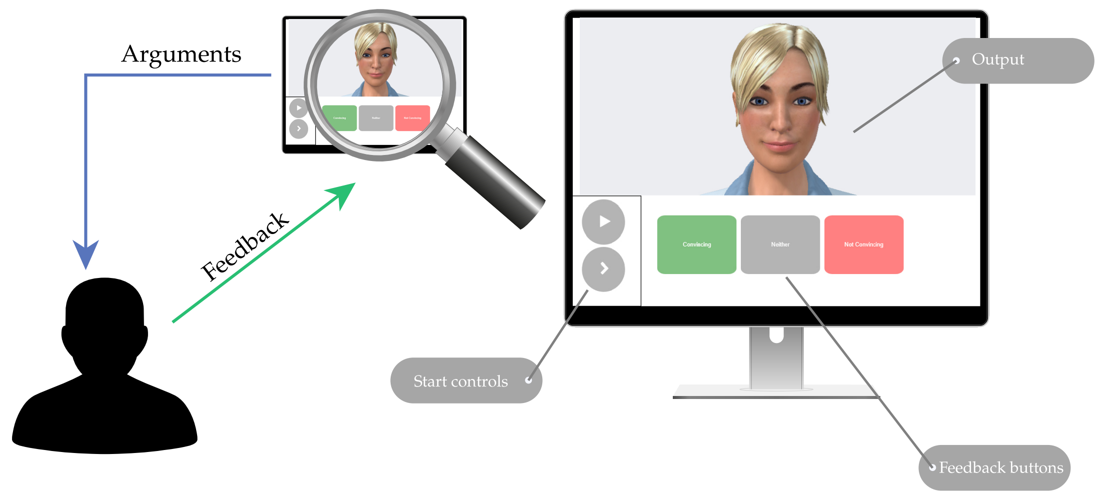


Figure 5.3: Prototype of our web interface consisting of an embodied virtual agent presenting her arguments to a user. The user gives feedback (*convincing*, *neutral*, *not convincing*) about the persuasive effectiveness, which is used to estimate the user's stance.

During the interaction, the agent presents pro- and counter-arguments about the topic that are *legal*. An argument $\Phi_{i\Rightarrow\_j}$ or its evidence component φ_i is called legal if the following yields: 1) $\Phi_{i\Rightarrow\_j} \notin Args_t$, and 2) $\Phi_j \in Args_t$, with $Args_t \subseteq Args$ the presented arguments at time step t. In other words, an argument is *legal* if its conclusion supports the evidence of any argument that has been presented to the user before, e.g., if $\Phi_2 \Rightarrow \Phi_1$ was presented, then any argument Φ_i with $\Phi_i \Rightarrow \Phi_2$ or $\Phi_i \Rightarrow \neg\Phi_2$ is legal. Thus, at every interaction step t, a random legal argument $\Phi_i \in Args$ is selected and presented to the users. After each argument Φ_i, the user provides the agent with explicit feedback $f(\Phi_i)$ with $f : Args \Rightarrow \{1.0, 0.5, 0.0\}$ for every argument $\Phi_i \in Args$ by using the feedback buttons (*convincing*, *neutral*, *not convincing*) as illustrated in Fig. 5.3, which translates to:

- *Convincing*, i.e., positive feedback ($f = 1.0$)
- *Neutral* ($f = 0.5$)
- *Not convincing*, i.e., negative feedback ($f = 0.0$)

The feedback is used to determine the weight ω_{Φ_i} as follows

$$\omega_{\Phi_i} = \begin{cases} f(\Phi_i) & \text{argument used by agent} \\ 0.5 & \text{else} \end{cases} \tag{5.11}$$

This feedback is subsequently used to determine the argument's persuasive effectiveness (Definition 5.1) and to predict the user's stance.

5.2.3 Argument Acquisition and NLG

ℹ The section was done in cooperation with the University of Ulm and was partly published in a similar form in Rach (2022), Rach et al. (2021), and Weber et al. (2020a). The argument acquisition and NLG part, excluding the study, was done by co-workers at Ulm University. They are included in this thesis for completeness.

This section describes the acquisition of the underlying argument structure. We identified three requirements in order to ensure a fair and reasonable evaluation (Rach, 2022):

Req. I. As our objective is to examine the predictive power of our model utilizing user feedback, we aim for topics that are *non-opinion-based*, meaning that we can assume a minimal bias from the user.

To fulfill the first requirement, we utilize extracted arguments from hotel reviews for a specific hotel that is assumed to be unknown to the system's users. With *non-opinion-based*, we refer to topics a user generally cannot have an opinion about beforehand as the user does not know anything about this specific hotel. Thus, whether or not to visit the hotel can only be based on the *facts* that the system presents to the user.

Req. II. The argument structure should not be biased in one direction based on one-sided information.

The second requirement is addressed by comparing the number of arguments in favor and against a topic.

Req. III. The structure must exclude any argument that outweighs the other ones, such as an excessively positive or negative aspect that defines the opinion towards the topic on its own for most users.

Meeting the third requirement is challenging as finding the existence of a so-called *defining argument* without an experiment with a representative number of participants is difficult. To mitigate the likelihood of such an argument, we consider the following factors within our domain: In actual reviews, a defining argument corresponds to a highly positive or negative aspect of a hotel that multiple people recognize in the same manner (such as bugs in beds). Such an aspect is typically the primary focus of most reviews since it is extraordinary. Consequently, if a representative set of reviews is examined, the overall opinion conveyed in the reviews is inclined towards this extraordinary aspect. Hence, a balanced argument structure based on real-world reviews excludes opinion-defining arguments.

Thus, based on the assumptions and requirements above, we employ hotel reviews from the annotated *SemEval-2015 Task 12* Test Data set (Pontiki et al., 2015).

For each hotel, we create a template that comprises the *Major Claim*, which is *"This hotel is worth a visit"*. We infer the argument structure for each hotel from the labels, utilizing a procedure adapted from the argument mining approach outlined in Cocarascu and Toni (2016). This selection was made because the corpus contains extensive, high-quality annotations on a substantial data set of genuine reviews, which includes all the essential information needed:

1. An aspect **category** $E \longrightarrow A$ consisting of an **entity** E (e.g., Hotel, Service, Location) and an **attribute** A (e.g., Price, Quality).

In Ex. 5.2, the **entity** is `facilities`, while the **attribute** is `general`. Each identified **entity** E is added as *Claim* to the argument structure.

Example 5.2: Annotations

The annotation is illustrated by the following example that includes the original sentence as well as the annotated labels:

Vending machines were out of everything except in the lobby.

Category	Polarity	Target
facilities $\longrightarrow$ *general*	*negative*	*Vending machines*

2. A **polarity** (`positive`, `negative`, `neutral`) of the category.

The **polarity** of each *Claim* is determined by comparing the number of `positive` and `negative` sentences related to that specific **entity**. For instance, as there are more negative sentences than positive ones for the **entity** `facilities`, the corresponding *Claim* is defined as *"The facilities are bad"*.

3. An **opinion target expression** within the annotated sentence that explicitly refers to the **entity** (if present).

We include all sentences that exhibit consistent **polarity** annotations for at least one **entity** and contain an **opinion target expression** (e.g., `Vending machines` in Ex. 5.2). We assume that sentences within a single review with the same entity label are interconnected, with the initial sentence addressing the corresponding *Claim* and subsequent sentences forming a chain of arguments. For other argument components, we presume a direct link to the relevant *Claim* unless they share the same target expression as a previous component, in which case they are directly linked to the same.

The relation (support vs. attack) between components is determined by the polarities. E.g., since "*The facilities are bad*" and "*Vending machines were out of everything except in the lobby*" share the same polarity, the relation between them is support .

The final corpus consists of 43 argument components ($\varphi_i \in L_t$) along with the relation ($\longrightarrow$) between them to build an acyclic-directed graph G as elaborated in Sec.2.1.3 (see Fig. 5.4 for an excerpt).

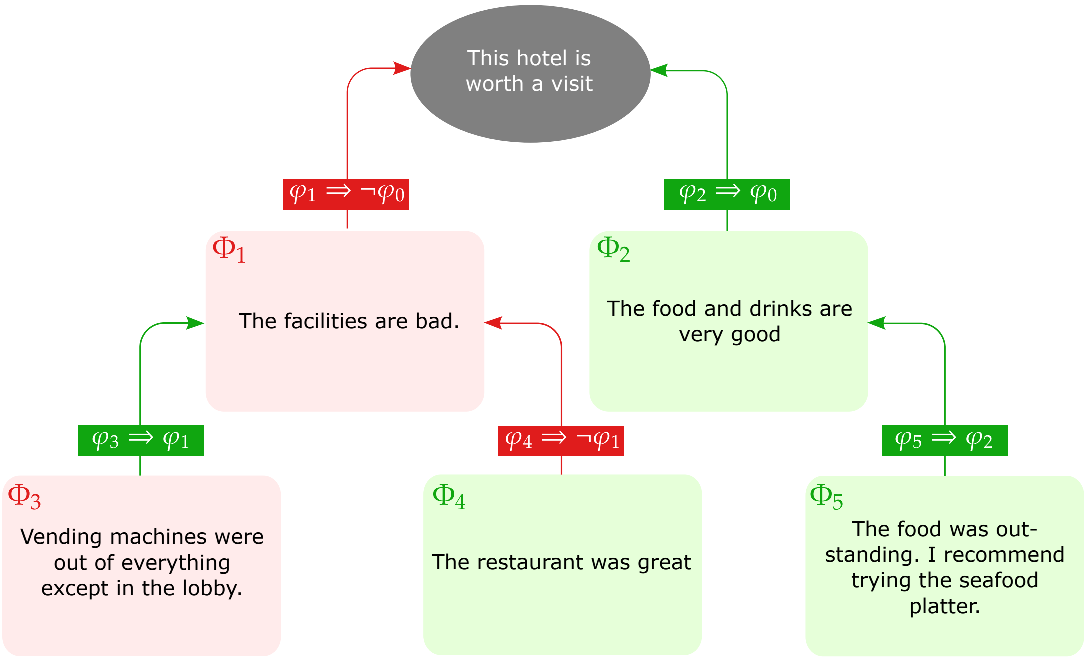

This hotel is worth a visit
$\varphi_1 \Rightarrow \neg\varphi_0$
$\varphi_2 \Rightarrow \varphi_0$
Φ_1
The facilities are bad.
Φ_2
The food and drinks are very good
$\varphi_3 \Rightarrow \varphi_1$
$\varphi_4 \Rightarrow \neg\varphi_1$
$\varphi_5 \Rightarrow \varphi_2$
Φ_3
Vending machines were out of everything except in the lobby.
Φ_4
The restaurant was great
Φ_5
The food was out-standing. I recommend trying the seafood platter.

Figure 5.4: Excerpt of the dataset *This hotel is worth a visit* based on the *SemEval-2015 Task 12* Test Data set (Pontiki et al., 2015) showing atomic argument components φ_i, φ_j (pro vs. con) and their relation (support vs. attack).

The annotated sentences also serve as a basis for the system's Natural Language Generation (NLG) component. To ensure grammatical completeness, all incomplete arguments were manually revised to form a stand-alone sentence, and any repeated arguments were merged into a single one. Moreover, phrases linking arguments were integrated into a separate template to ensure a fluent interaction. These statements contain details regarding the position taken within the argument (particularly when it attacks its target), a notification if the argument does not relate to the immediate prior argument, and both an introduction and a closing statement (see Ex. 5.3). The template encompasses various versions for each scenario, from which the system randomly chooses one for each response.

Example 5.3: Generated NLG output

The following example utterance includes a topic switch (**ts1**), the referenced argument (Φ_1), a notification about the stance (s1), and the new argument (Φ_2): *The next argument is related to something I mentioned earlier. I said (**ts1**): All in all, it is a nice and affordable spot for sightseeing in the area (Φ_1). I also found an opinion that disagrees with this aspect. The respective author wrote (s1): I think all in all the price was way too high for such a poor accommodation (Φ_2).*

To ensure that the data set does not contain outliers as required by **Req. III.**, we conducted an online survey using *LimeSurvey*. 105 participants between 18 and 56 (mean: 32.886) were asked to rate the strength of the arguments on a 5-Likert Scale based on the question: *"To what extent would the given review have an impact on your decision (not) to visit the hotel?"*

In order to prevent cheating, we added two test questions to the survey. Out of 391 participants, only 105 answered the test questions correctly and were subsequently included in the analysis. To derive the strength of each argument, we calculated the mean value of all responses and normalized them within the range of $[0, 1]$. The final strength Ψ_i of each argument $\Phi_i, 1 \leq i \leq 43$ is summarized in Fig. 5.5 (see Appendix B.1.1 and B.1.2 for full data overview).

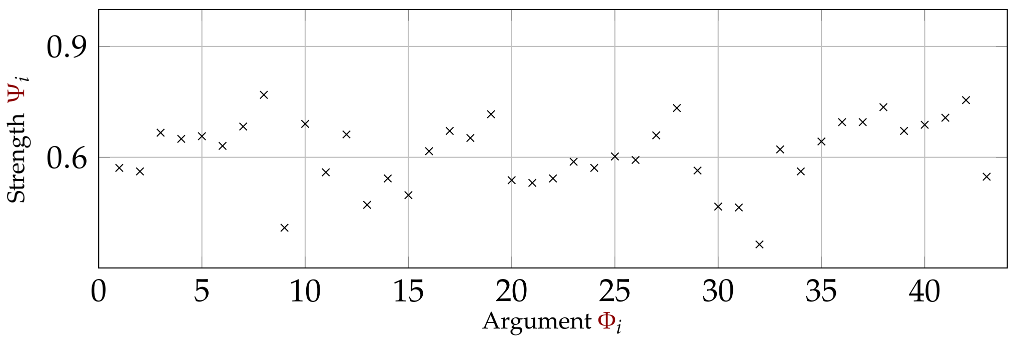


Figure 5.5: Argument Strengths for the SemEval-2015 Test Dateset

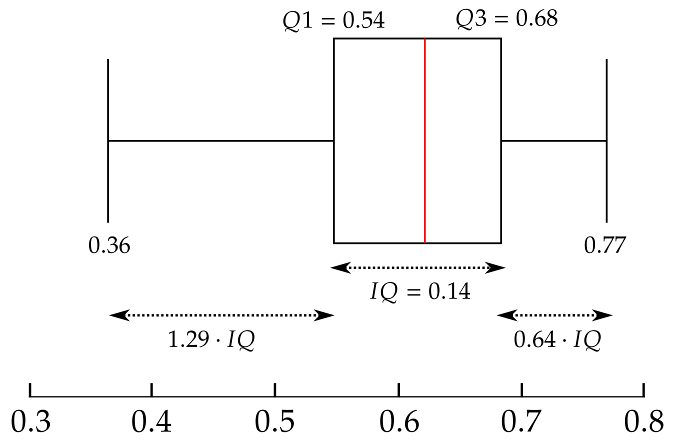


Figure 5.6: Box plot of Argument Strengths for the SemEval-2015 Test Dateset.

We can check whether there are exceptionally strong arguments by using Box plots (see Fig. 5.6). We define an argument as exceptionally strong if it is an outlier of the distribution (description of **Req. III.**), i.e., it is outside the *upper whisker*, which is calculated by adding the third quartile and 1.5 times the interquartile range (IQ). As of the analysis, no arguments lie beyond the *upper whisker*, leading us to conclude that no exceptionally strong arguments outweigh the others.

5.2.4 A Study to Evaluate the Stance Model's Predictive Power

ⓘ The study aims to evaluate the effectiveness of the stance estimation model, with a specific focus on its predictive power and accuracy.

For this study, we defined the following research question to be examined:

(RQ1) Does the stance estimation model accurately reflect the user's opinion?

5.2.4.1 Participants, Apparatus, and Procedure

The study was conducted in-person in our lab. We recruited 48 participants (32 male, 16 female, 18-30 years old) from a the University of Augsburg. All participants were students. At the start of the study, they were informed about the general procedure and asked to provide the agent with feedback whether or not they find an argument convincing to (not) visit the hotel. After the session, they were asked whether they would like to visit the hotel. To avoid bias effects beforehand, they were not told about the system's overall goal in predicting their decision but asked to provide feedback on whether or not they found an argument convincing. Fig. 5.7 depicts the general study setup showing a participant interacting with the agent.

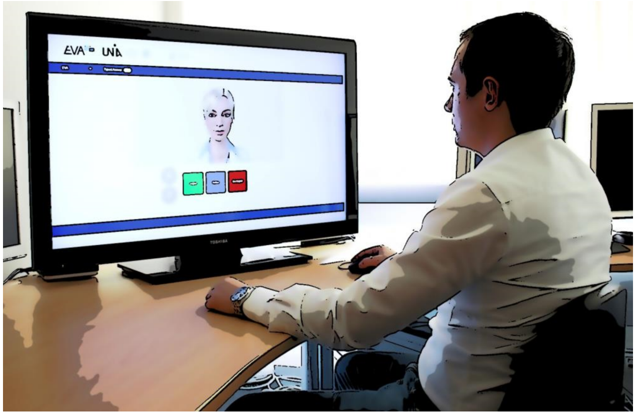


Figure 5.7: Evaluation setup with an interactive agent.

The agent presented each of the 43 arguments for and against the hotel in each session, which took 10-15 minutes. After each argument the user gave the required feedback, which the agent used to compute the effectiveness z_{φ_0}. The agent's *assigned stance* was counter-balanced, i.e., half of the participants interacted with an agent who favored visiting the hotel and vice versa. An interaction based on the argument graph excerpt seen previously in Fig. 5.4 would like like as follows:

Agent: This hotel is not worth a visit.
Agent: The facilities are bad.
User: * *gives feedback* *
Agent: I also found an opinion that disagrees with this aspect. The respective author wrote: The restaurant was great
User: * *gives feedback* *
Agent In contrast to that I also found the following opinion: Vending machines were out of everything except in the lobby.
User: * *gives feedback* *
Agent Okay, let's get back to the initial claim. The next argument contradicts the initial claim by saying: The food and drinks a very good.
User: * *gives feedback* *
Agent The food was outstanding. I recommend trying the seafood platter.
User: * *gives feedback* *

During the study, we collected the following data:

1. Directly given user feedback $f(\Phi_i), \forall \Phi_i \in Args$.
2. Computed effectiveness z_{φ_0} using the given feedback.
3. Subjective decision if users like to visit the hotel (post-study).

5.2.4.2 Analysis

We plot collected data to explore trends and present statistical tests in the following.

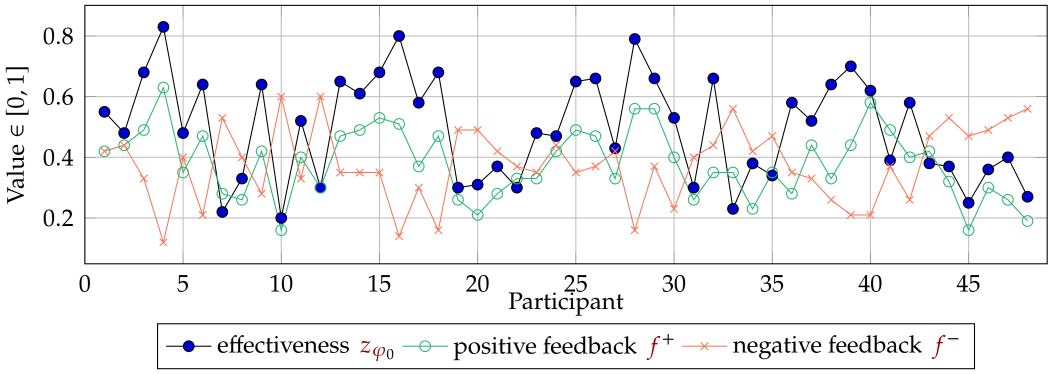


Figure 5.8: Collected data: Effectiveness z_{φ_0}, percentage of pos. (neg.) feedback f^+ (f^-) with respect to the assigned stance (see Appendix B.2).

General trends Fig. 5.8 summarizes the results for all participants depicting the agent's final *effectiveness* z_{φ_0}, the percentage of *user feedback* in favor of the agent's assigned stance f^+ and the percentage of *user feedback* not in favor of the agent's assigned stance f^-. Neutral feedback is not depicted as it does not affect the effectiveness (see Def. 5.1). First, we notice two trends:

1. The higher (lower) the positive feedback, the higher (lower) the effectiveness z_{φ_0}.
2. The lower (higher) the negative feedback, the higher (lower) the effectiveness z_{φ_0}.

Thus, the positive feedback f^+ seems to correlate with the effectiveness positively, and the negative feedback f^- seems to negatively correlate with the effectiveness z_{φ_0}. As stated, the general idea of the effectiveness z_{φ_0} is to predict the user's current stance. So, positive feedback increases the effectiveness score, while negative feedback decreases the effectiveness score. The trends, therefore, are in line with our expectations.

Statistical Analysis To verify the trends statistically, we computed the correlation between feedback and effectiveness, showing a strong and significant correlation (*positive correlation for positive feedback, negative correlation for negative feedback*, see Tab. 5.1).

Table 5.1: Correlation between feedback and effectiveness.

	n	r	p	
Pos. Feedback & effectiveness - Pearson correlation	48	0.92	<.001	✓
Neg. Feedback & effectiveness - Pearson correlation	48	-0.83	<.001	✓

Prediction Accuracy We then evaluated to what degree the predicted user's stance $user\_stance$ (see Eq. 5.10) and the subjective user's decision to visit the hotel match. We computed the agent's confidence in the predictions based on their proximity to the criterion value ($z_{\varphi_0} = 0.5$). To this end, we utilize a modified sigmoid function to a) ensure that the extreme values of 0 and 1 correspond to a confidence of 100% and b) obtain a more fine-grained prediction for the most common interval $[0.3, 0.7]$ (see Def. 5.2). Consequently, a confidence ≥ 80% means $z_{\varphi_0} \geq 0.64$ or $z_{\varphi_0} \leq 0.36$.

Definition 5.2: Confidence

Let z_{φ_0} be the computed persuasive effectiveness; then the confidence is computed as

$$\frac{1}{1 + e^{-10\cdot|z_{\varphi_0} - 0.5|}} \tag{5.12}$$

The results in Fig. 5.9 show that the objective system's prediction is very accurate even for low confidence values, proving the practical potential of the model.

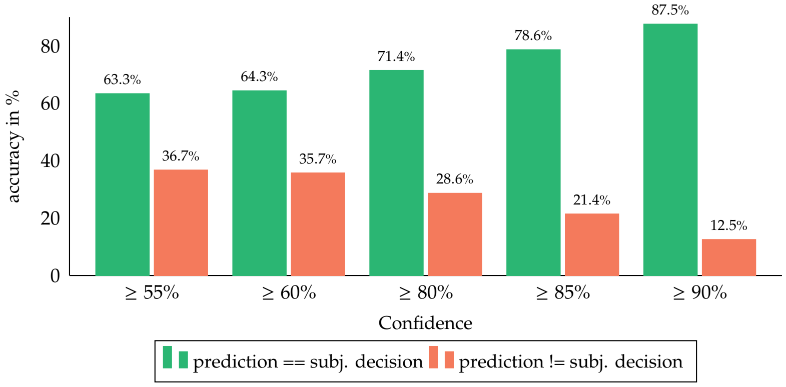


Figure 5.9: Model accuracy of predicted user's stance depending on different confidence values.

Table 5.2: F1 score for different prediction confidences.

	Confidence			
Stance	≥ 60%	≥ 80%	≥ 85%	≥ 90%
+	*0.67*	*0.73*	*0.77*	*0.86*
-	*0.62*	*0.70*	*0.80*	*0.89*

To verify the sensitivity and precision of the predictions, we computed the F1 score for both the positive and negative stances depending on the prediction confidence as summarized in Tab. 5.2. The results show that the F1 score increases with higher confidence, thus, proving both the sensitivity and precision of the prediction.

5.2.5 Discussion and Limitations

We presented a user stance prediction model, which is needed for our computational metric to predict the user stance based on direct user feedback. We conducted a user study to evaluate the validity and practicability of the underlying prediction model. The validation of the underlying model showed a significant correlation between feedback and computed effectiveness level.

5.2.5.1 Argument Structure as Information Source

Despite the observed correlation between overall feedback and effectiveness at the end of the interaction, using bipolar argument graphs as a prediction model bears several advantages compared to models based solely on feedback statistics: (i) The graph allows different feedback weighting for different arguments, resulting in more fine-grained user stance estimation. (ii) Additional argument-specific or structure-specific information can be integrated to provide more detailed information for learning. (iii) The user stance estimation can be used for behavior learning, as outlined in Weber et al. (2020b) and Rach et al. (2021). (iv) Behavior learning can be combined with fine-grained logical strategies (Rach et al., 2018a; Rosenfeld & Kraus, 2016).

5.2.5.2 Predictive Power of the Model

It is important to note that the feedback provided by the users during the study may not necessarily reflect their final decision regarding the hotel. However, we generally assume that users who agree with arguments supporting the *Major Claim* are more likely to have the same stance, which we confirmed within the user study. The ability to accurately predict a user's current stance based on their feedback during an agent-user interaction makes it highly effective for our approach and other scenarios. For example, the predicted stance can be used to determine when the user is likely to be convinced, allowing a persuasive system to cease the persuasion process. Additionally, in persuasive debates involving multiple agents or humans, the predicted stance can be used to determine the overall success of the debate during the interaction, enabling the agents to adopt strategies employed by the more successful agent, such as proposed in Rach et al. (2021) and Weber et al. (2020b).

5.2.5.3 Limitations

Despite our approach's validation and high predictive sensitivity and precision, it is essential to acknowledge its limitations. The effectiveness of the user's feedback depends on the entire argument structure and, more importantly, on the number of arguments directed towards a single argument. However, in the current version, arguments targeting the same parent argument (siblings) are still equally weighted, unlike arguments targeting different arguments. Therefore, it would necessary to investigate whether this approach is practical even when there is a significant imbalance between argument strengths. A plausible solution could be to allow users to provide additional information about the weight of their feedback relative to sibling arguments or employ interval-scaled feedback, similar to Aicher et al. (2021b).

5.3 Key Points & Summary

Within this chapter...

- ... we presented and evaluated a stance estimation model to allow for dynamic stance estimation within our proposed computational metric.
- ... we derived the computational metric AVQ for RE.

This chapter addressed research question **Q2.1**:

> Q2.1 *"How can we formulate a computational metric for RE that is operationally feasible and can be programmatically implemented allowing the agent to guide the user's argument visitation focus?"*

In this chapter, we introduced a computational metric, namely Argument Visitation Quotient (AVQ), for RE, designed to measure the extent to which users explore arguments that challenge their position. Thereby, our metric first measures the user's argument focus by considering the ratio of visited pro and con arguments. It then computes AVQ based on the inverse proportion between the user's stance and their argument focus.

As the user could change their opinion when exposed to *challenger arguments*, we presented a user stance estimation model that avoids disrupting the interaction by directly asking for their stance. Instead, this model infers stance changes by considering the user's agreement and disagreement towards the arguments.

We conducted a user study to evaluate the validity and practicability of the underlying prediction model. The validation of the model demonstrated a significant correlation between user feedback and the computed effectiveness level, showing that the system can predict the user's stance with high accuracy, even for low confidence values. This makes the model a powerful tool for our computational metric AVQ.

5.4 Relevant Publications

- Weber, K., Janowski, K., Rach, N., Weitz, K., Minker, W., Ultes, S., & André, E. (2020a). Predicting Persuasive Effectiveness for Multimodal Behavior Adaptation using Bipolar Weighted Argument Graphs. *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 1476–1484

- Weber, K., Rach, N., Minker, W., & André, E. (2020b). How to Win Arguments: Empowering Virtual Agents to Improve Their Persuasiveness. *Datenbank-Spektrum*, *20*, 161–169

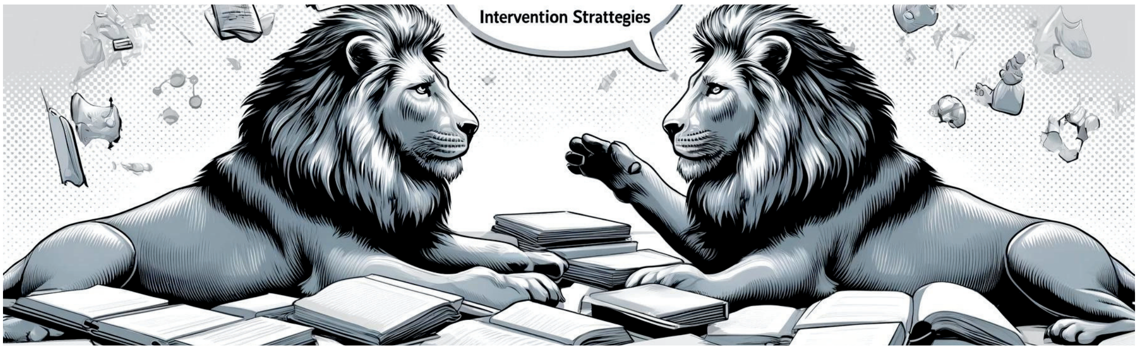
- Weber, K., Aicher, A., Minker, W., Ultes, S., & André, E. (2023a). Fostering User Engagement in the Critical Reflection of Arguments. *Proceedings of the 13th International Workshop on Spoken Dialogue Systems (IWSDS)*, 1–16

- Aicher, A., Weber, K., André, E., Minker, W., & Stefan, U. (2024). BEA: Building Engaging Argumentation. *Proceedings of the 1st International Conference on Robust Argumentation Machines (RATIO)*, 1–17

- Weber, K., Hogh, N., Conati, C., & André, E. (2024). A Gaze into Argumentative Chatbots: Exploring the Influence of Challenger Arguments on Reflection and Attention. *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents (IVA)*, 1–10

CHAPTER 6

FOSTERING REFLECTION: INTERVENTION STRATEGIES



"The roots of education are bitter, but the fruit is sweet." (Aristotle)

❗ For background on RL, read Sec. 2.1.4

ℹ This chapter describes the intervention strategies the agent can employ to promote a less biased argument exploration if users stick to arguments supporting their own opinions.

The non-adaptive base strategy (Sec. 6.1) was partly previously published by the author in a similar form in a peer-reviewed paper (Weber et al., 2023a, 2024).

The RL adaptation approach (Sec. 6.2) is based on previous work of the author (Rach et al., 2021; Weber et al., 2018a, 2020b, 2020a).

In Ch. 5, we presented a computational metric that gauges the extent to which users predominantly focus on arguments aligning with their viewpoint. Based on this metric, the agent uses interventions to promote less biased argument exploration if users consistently focus on arguments supporting their own opinions. Since focusing solely on one's own arguments can reinforce existing views rather than encourage reflective thinking, our intervention algorithm aims to increase the focus on *challenger arguments*, in line with relevant literature (Paul, 1990; U. Peters, 2022). The most critical question here is when the agent should intervene. While setting a threshold (e.g., intervening if AVQ < threshold) might be an option, determining the optimal threshold is quite challenging. Therefore, we opted for the most conservative approach and decided to intervene whenever the agent identifies an argument that leads to an increase in engagement with *challenger arguments*, regardless of the current AVQ value.

In our studies, we aim to investigate the effects of different intervention strategies. Therefore, we introduce and describe three types of strategies:

- *Base* strategy (non-adaptive, Sec. 6.1): This strategy defines the conditions for intervention. It triggers an intervention whenever an argument is identified that could increase the engagement with *challenger arguments*.
- *Gamification* strategy (non-adaptive, Sec. 6.1): This strategy involve using game-like elements to motivate users to engage with *challenger arguments*.
- The agent's *linguistic* style (adaptive, Sec. 6.2): This strategy focuses on how the agent's language can be adapted to encourage users to consider *challenger arguments*.

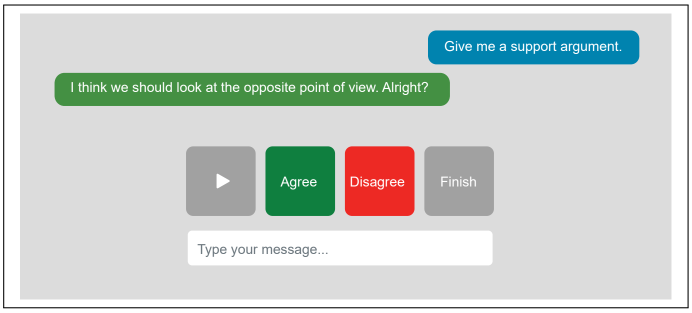
As previously mentioned, we will also investigate the effect of the agent's *embodiment* later on. However, this is considered a co-variate and not a separate type of intervention strategy. Therefore, it is not detailed further in this chapter.

6.1 Non-Adaptive Base Strategy

The intelligent agent keeps track of the user's AVQ (see Sec. 5.1) and intervenes if necessary, i.e., it suggests considering an opposing viewpoint. Given a user request why_x, let $Args_x \in Args$ be the set of all valid arguments the agent can present matching the user request why_x, and $Args_c \in Args$ all other valid arguments not matching the request. The agent simulates the AVQ for all arguments and returns the argument that maximizes it. The intervention takes place if the simulated AVQ for any argument in $Args_c$ is greater than the maximum possible AVQ for any argument in $Args_x$, i.e.,

$$\max_{\Phi_i \in Args_c} (sim\_cavq(\Phi_i)) > \max_{\Phi_i \in Args_x} (sim\_cavq(\Phi_i)) \tag{6.1}$$

and, if so, returns $\Phi_i = \arg\max_{\Phi_i \in Args_c}(sim\_cavq(\Phi_i))$, and presents it to the user if the user accepts the interventions.


Give me a support argument.

I think we should look at the opposite point of view. Alright?

Figure 6.1: Intervention example. The user (right) requests a supporting argument, and the agent (left) intervenes, suggesting the opposing view.

Fig. 6.1 shows an example intervention by the agent. In case of denial, the agent proceeds with the initial user request why_x and presents an argument $\Phi_i \in Args_x$.

Algorithm 3 sketches the overall intervention algorithm. The algorithm uses two parameters ϵ and c. ϵ defines the minimal difference of AVQ before intervention is applied, and c defines a fixed constant that defines how many arguments the user can ask for before any intervention is applied. The function `generateIntervention()` generates the intervention utterance.

Algorithm 3: Base Intervention Strategy (BIS)

```
Data: $\epsilon > 0, c > 0, c' = 0$
foreach $k = 1, \ldots, n$ do
    BIS($\epsilon, c, c'$)
Function BIS($\epsilon, c, c'$):
    $intent \leftarrow$ extract user intent from user move $m_k$
    if $intent \in \{why_{pro}, why_{con}\}$ then
        \\Check visitation quotient
        $avqUser \leftarrow \max_{\Phi_i \in Args_x}(sim\_cavq(\Phi_i))$          // Eq. 6.1
        $avqIdeal \leftarrow \max_{\Phi_i \in Args_c}(sim\_cavq(\Phi_i))$         // Eq. 6.1
        if $avqIdeal - avqUser > \epsilon$ and $c' \geq c$ then
            \\Trigger and apply intervention strategy
            $utterance \leftarrow$ generateIntervention($intent$)
        else
            $utterance \leftarrow$ getArgument($intent$)
            $c' \leftarrow c' + 1$
        apply($utterance$)
    \\Process other speech acts
```

6.1.1 Non-Gamified Intervention Strategy

After triggering the intervention, the system shows a pop-up window that prompts the user to approve or deny the suggestion (see Fig. 6.2).

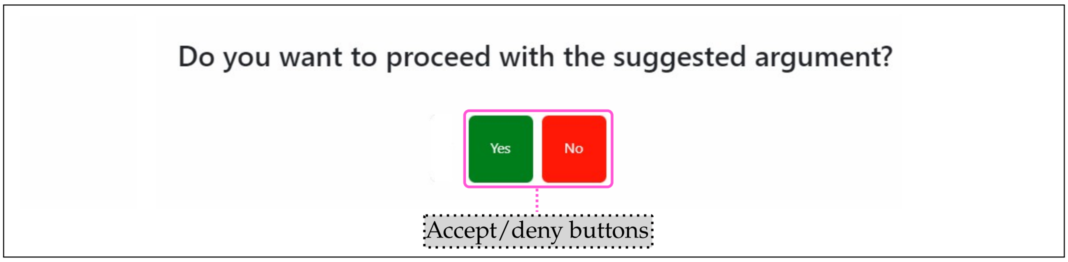

Do you want to proceed with the suggested argument?

Figure 6.2: Intervention pop-up prompting the user to accept or deny the agent's suggestion.

This step is crucial, as it allows users to decide whether they wish to view the diverging viewpoint. The agent encourages reflection and slow thinking by providing users with this decision-making option. Users are prompted to consider the potential benefits of exploring alternative perspectives before making their decision. Without this option for users to engage in the cognitive task of deciding whether to explore alternative viewpoints, the potential for meaningful cognitive development could be limited. Further, a forceful intervention could lead to unhappiness and dissatisfaction with the agent.

6.1.2 Gamified Intervention Strategy

Since gamification can increase users' motivation and engagement by making the interaction more enjoyable and rewarding (Deterding et al., 2011), we further added a *gamified* intervention strategy to our system. Gamification applies game mechanics to non-game contexts. The most common used gamification strategies encompass 1) *challenge*, 2) *badge*, and 3) *points*, followed by *leader-board* and *level* (Silpasuwanchai et al., 2016). Tab. 6.1 shows an example of employed gamification strategies in popular games.

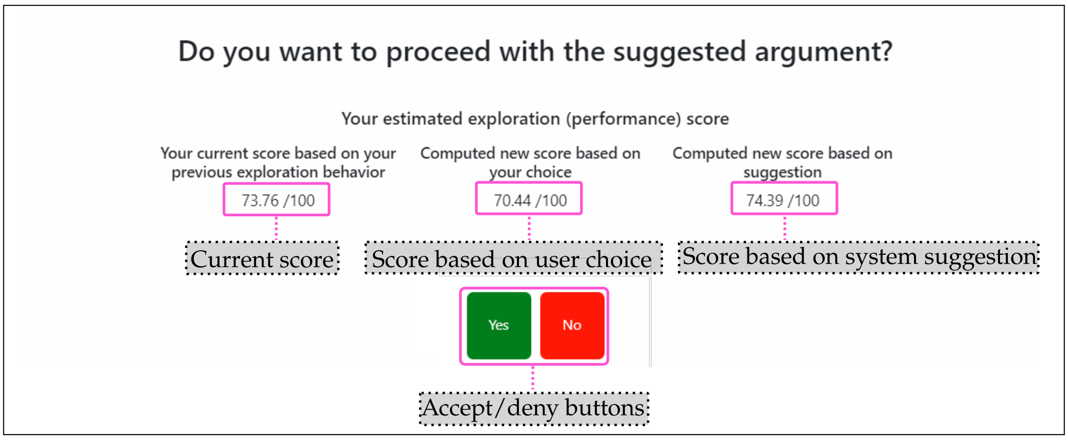
Table 6.1: Examples of gamification strategies in popular games.

Strategy	Game	Description
Challenge	*Super Mario*	In *Super Mario*, challenges involve completing levels by navigating obstacles, defeating enemies, and reaching the goal flag.
Badge	*Need for Speed*	The *Need for Speed* series utilizes badges in the form of achievements or trophies, rewarding players for their racing accomplishments.
Points	*Angry Birds*	*Angry Birds* incorporates a points system where players earn points by successfully launching birds to destroy structures and defeat pigs.
Leaderboard	*Fortnite*	*Fortnite* features leaderboards that display players' rankings based on their performance in matches, fostering competition among players.
Level	*World of Warcraft*	In *World of Warcraft*, players progress through levels by completing quests, defeating enemies, and gaining experience points within a comprehensive leveling system.

As seen before, the *non-gamified* strategy (Sec. 6.1.1) also contains a sort of *challenge*, as users are challenged to consider alternative perspectives by deciding whether to accept the agent's suggestion. It is noteworthy that this is a cognitive challenge that is necessary within the reflection process and not a game strategy due to the missing incentive of managing this challenge.

Within the *gamified* strategy, referred to as *"with gamification"*, we opted for the gamification strategy *points* as this can be easily reflected using the calculated metric AVQ during interaction unlike *badges*, *level*, and *leader-board* which would apply after the interaction took place. As a scoring system, we display the computed metric AVQ, which serves as a form of points that users can earn based on their exploration behavior within the argumentative dialogue system.

Fig. 6.3 shows the gamified intervention strategy employing the *points* strategy displaying three scores: One reflecting the user's current points, one reflecting the user's choice and another indicating the score they would receive by accepting the agent's suggestion. This creates a point differential that incentivizes users to strive for higher scores by accepting the agent's recommendations. Users can see the direct impact of accepting or rejecting the agent's suggestion on their score, which helps them understand the consequences of their actions and encourages reflection on their decision-making process.



Do you want to proceed with the suggested argument?

Your estimated exploration (performance) score

Your current score based on your previous exploration behavior: 73.76 /100

Computed new score based on your choice: 70.44 /100

Computed new score based on suggestion: 74.39 /100

Figure 6.3: Gamified intervention displaying the *performance* scores (high score) of the exploration.

6.2 Adaptive Strategies using RL

In this section, we present a conceptual approach to adapt the linguistic style of the interventions. The focus is on verbalizing the interventions, i.e., how the user is prompted to view challenger arguments.

While the *base strategy* only intervenes based on the *intervention condition* without adjusting the intervention text (Eq. 6.1), there is a chance that users might not be willing to accept the intervention. In such a case, it might be helpful to adapt the linguistic style to the user, enhancing the success rate. Therefore, we employ the politeness theory by Brown and Levinson (2009) because it can effectively reduce the risk of offending or alienating the user.

Ritschel et al. (2019) used the politeness theory to adapt the linguistic style of an assistive robotic health companion. Rather than encouraging people to look into challenger arguments, they encouraged activities to enhance mental and physical well-being. Different verbal strategies can have varying impacts concerning their perceived persuasiveness (Hammer et al., 2016), e.g., a direct request might be perceived as too forceful, while an indirect suggestion might be ignored.

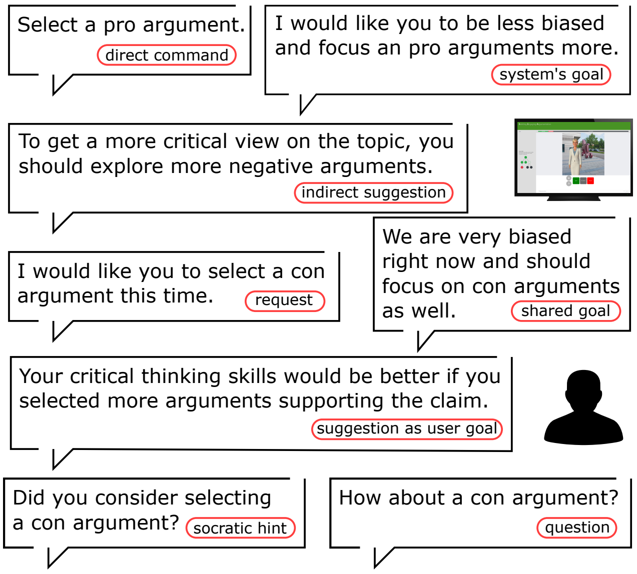

Select a pro argument.
direct command
I would like you to be less biased and focus an pro arguments more.
system's goal
To get a more critical view on the topic, you should explore more negative arguments.
indirect suggestion
I would like you to select a con argument this time.
request
We are very biased right now and should focus on con arguments as well.
shared goal
Your critical thinking skills would be better if you selected more arguments supporting the claim.
suggestion as user goal
Did you consider selecting a con argument?
socratic hint
How about a con argument?
question

Figure 6.4: Example expression of each of the employed politeness strategies adapted from Ritschel et al. (2019)

Figure 6.4 illustrates exemplary utterances for each strategy. Given the challenges, an adaptive approach using RL becomes the best option. Using RL, the agent can dynamically tailor its interventions based on the user's responses and behavior, i.e., whether the user accepts the intervention. RL allows the agent to learn from interactions and optimize the verbal strategy to maximize the user's acceptance rate of the intervention. This adaptive method ensures that the intervention is neither forceful nor subtle but appropriately tailored to the user's preferences.

The RL approach is formulated conceptually as follows with modifications compared to own previous works:

1. We establish the action space $\mathcal{A}$ (Def. 6.2), consisting of various linguistic style types. This approach differs from our previous work, where the action space included:
   - Various joke categories within a robotic humor adaptation approach (Weber et al., 2018a).
   - A set of arguments with different emotional overtones within a persuasive argumentative dialogue setup. This involved two agents adapting their emotional expressions (verbal or non-verbal) based on the user's perceived persuasiveness (Rach et al., 2021; Weber et al., 2020b, 2020a).
2. We define the state space $\mathcal{S}$ (Def. 6.1), which includes user behavior and feedback. In our previous works, we incorporated user behavior in the state space with the following differences:
   - The user's social signals, such as smiles and laughter (Weber et al., 2018a).
   - The emotions conveyed by the arguments (Weber et al., 2020b).

- The user's affective state (Rach et al., 2021).

3. We define an appropriate reward function $\mathcal{R}$ (Def. 6.3) that allows the agent to optimize its verbal strategies. In our previous work:
   - The agent relied on social signal analysis (smiles and laughter) as a reward signal, which introduced noise into the reward signal (Weber et al., 2018a).
   - The agent relied on the user's affective state (Rach et al., 2021).

   In contrast, we rely solely on the user's exploration behavior, i.e., the number of pro and con arguments selected.

6.2.1 RL Model

A state $s_t \in \mathcal{S}$ (see Def. 6.1) contains the AVQ score AVQ_t, the user's focus F_t and the user's stance $z_{\Phi_0,t}$. This allows the agent to learn *interaction effects* between AVQ score, user focus, and user stance.

Definition 6.1: State Space $\mathcal{S}$

Let $\text{AVQ}_t \in [0,1]$ be the current visitation quotient at time step t. Further let $F_t \in [-1,1]$ be the user focus and $z_{\Phi_0,t} \in [0,1]$ be the estimated user stance. Then, a state $s_t \in \mathcal{S}$ is defined as a triple:

$$s_t := (\text{AVQ}_t, F_t, z_{\Phi_0,t}) \tag{6.2}$$

The action space (see Def. 6.2) incorporates the politeness theory by Brown and Levinson (2009) and eight verbal strategies proposed in Johnson et al. (2005).

Definition 6.2: Action Space $\mathcal{A}$

Let $a_t \in \mathcal{A}$ be an action at time step t. The action space the agent can make use of consists of eight verbal strategies (Johnson et al., 2005):

- **Direct command**
- **Indirect Suggestion**
- **Requests**
- **Suggestion as agent's goal**
- **Suggestion as shared goal**
- **Questions**
- **Suggestion expressed as user goal**
- **Socratic Hints**

It is important for every RL problem to use a suitable reward function. Wiewiora (2010) indicated that most problems provide intuitive rewards. Such intuitive rewards can be laughter in a joke-telling scenario (Weber et al., 2018a), engagement in a story-telling scenario (Ritschel et al., 2017), the difference between a agent's and user's affective state (Rach et al., 2021), or any other measure indicating the distance between the current agent's state and the agent's intended goal.

The agent aims to motivate the user to engage in specific exploration behavior, particularly to view challenger arguments, which is reflected by the metric AVQ. Thus, using AVQ as a reward signal is reasonable. An agent's action is successful if the user's AVQ has increased and unsuccessful if it has decreased. Thus, we assign a positive reward of 1 if AVQ increases, and -1 otherwise (see Def. 6.3).

Definition 6.3: Reward Function $\mathcal{R}$

Let $s_t \in \mathcal{S}$ be the current state and $a_t \in \mathcal{A}$ an action at RL time step t, and let AVQ_{t+1} be the current visitation quotient after performing action a_t as well as AVQ_t the previous one, then the reward $\mathcal{R}_{t+1}(s_t, a_t)$ is defined as:

$$\mathcal{R}_{t+1}(s_t, a_t) := \begin{cases} 1 & if\ \mathrm{AVQ}_{t+1} > \mathrm{AVQ}_t \\ -1 & \text{else} \end{cases} \tag{6.3}$$

6.2.2 Algorithm for Intervention Strategy Adaptation

This section shows the intervention and adaptation process (see Alg. 4). The agent does not learn after every dialogue interaction step k but only if the intervention was applied, which is if $avqIdeal - avqUser > \epsilon$. After that, the agent checks the success or failure of the strategy (see Def. 6.3). As RL algorithm, we used *Q-Learning with Function Approximation* as previously introduced in Alg. 2.

Algorithm 4: Adaptive Intervention Strategy (AIS)

```
Data: $\epsilon > 0, t = 0, c > 0, c' = 0, s_t = s_0$
foreach $k = 1, \ldots, n$ do
    AIS($\epsilon, c, c'$)
Function AIS($\epsilon, c, c'$):
    $intent \leftarrow$ extract user intent from user move $m_k$
    if $intent \in \{why_{pro}, why_{con}\}$ then
        \\Check visitation quotient
        $avqUser \leftarrow \max_{\Phi_i \in Args_x}(sim\_cavq(\Phi_i))$        // Eq. 6.1
        $avqIdeal \leftarrow \max_{\Phi_i \in Args_c}(sim\_cavq(\Phi_i))$       // Eq. 6.1
        if $avqIdeal - avqUser > \epsilon$ and $c' \geq c$ then
            \\Trigger and apply intervention strategy
            $s_t \leftarrow$ observe state.                                    // Def. 6.1
            $a_t \leftarrow$ according to strategy $\pi$.
            $utterance \leftarrow$ generateIntervention($intent, a_t$)
        else
            $utterance \leftarrow$ getArgument($intent$)
            $c' \leftarrow c' + 1$
        apply($utterance$)
    else if $intent \in \{accept, deny\}$ then
        $utterance \leftarrow$ getArgument($intent$)
        apply($utterance$)
        \\Learning
        $s_{t+1} \leftarrow$ observe state.                                    // Def. 6.1
        $\mathcal{R}_{t+1} \leftarrow$ computeReward($s_t, s_{t+1}$).              // Def. 6.3
        $Q_\omega(s_t, a_t) \leftarrow$ update($s_t, \mathcal{R}_{t+1}, s_{t+1}$)    // Learning, Eq. 2.54
        $t \leftarrow t + 1$
    \\Process other speech acts
```

6.2.3 Experiments with Simulated Users

We conducted several experiments (simulations) as a first proof-of-principle evaluation of the adaptation approach. This is necessary to check

1. . . . if the agent can adapt a personalized strategy based on the user's initial stance and their AVQ.
2. . . . if the prototype is robust enough to cope with ***non-deterministic*** and ***non-stationary*** user reactions as it mostly occurs in human-computer interactions.

Non-deterministic reactions in our context mean that even if the chosen action $a \in \mathcal{A}$ considered the best option based on the user's previous responses, the intervention strategy may not be successful. This means there is a probability that users respond differently with a probability $\beta \in [0, 1]$. This scenario could occur for two main reasons: 1) Users might want to hear a specific argument and thus *deny* the agent's suggestion. 2) Misunderstandings caused by *speech recognition* errors when interpreting the user's response.

To simulate *non-deterministic* behavior, we apply a ***noise simulation***, involving the addition of random variations or uncertainties into the system. This approach mimics the unpredictability and variability present in real-world scenarios.

Non-stationary reactions imply that the optimal strategy Q^* can change over time meaning that the user changes their behavior over time and no longer responds to a particular politeness intervention strategy as they did before. The question is how quickly the agent can adapt its strategy Q_π (see Eq. 2.49) to such changes (Sutton & Barto, 2018, p.30).

To simulate *non-stationary* behavior, we apply a ***shuffling*** mechanism. *Shuffling* refers to the rearrangement of the action probabilities. After assigning initial success rate probabilities (see below), a shuffling mechanism is applied during interaction to rearrange these probabilities among the different actions, simulating changes or fluctuations that may occur over time. This helps evaluate how well the agent can adjust its strategy when faced with variations in the ***success rates*** of different actions.

When conducting simulation experiments in RL environments, accurately mapping real-world users to simulated users is crucial for meaningful analysis. In most scenarios, strategies likely have a ***success rate*** greater than zero. To model this, ***success rates*** for each action were randomly assigned within the [0,1] range. Subsequently, the highest probability was set to 1.0, while the lowest was set to 0.0. This choice was made to create a controlled environment for assessing the impact of noise and shuffling. By designating extreme probabilities, we ensure that each simulation always has a known correct and a known incorrect action. This added certainty in outcomes allows for a systematic examination of the effects of randomness (noise) and rearrangements of action probabilities (shuffling) on the overall performance of the RL agent.

We simulated 100 users over 250 time steps. Each user was assigned a random stance and a random distribution of the success rates of the available actions. Tab. 6.2 shows the success rates of three example users.

Table 6.2: Action success rate distributions for three example users.

	Success Probability		
Action $a \in \mathcal{A}$	*User 1*	*User 2*	*User 3*
Direct Command	*0.85*	*0.00*	*0.39*
Agent's Goal	*1.00*	*0.37*	*0.42*
Indirect Suggestion	*0.83*	*0.48*	*0.00*
Request	*0.27*	*0.32*	*1.00*
Shared Goal	*0.12*	*0.17*	*0.37*
Suggestion as a User Goal	*0.06*	*0.20*	*0.69*
Socratic Hint	*0.00*	*1.00*	*0.41*
Question	*0.17*	*0.24*	*0.70*

In these examples, User 1 exhibits strong success rates in *direct commands*, *agent goal*, and *indirect suggestions*, and has relatively low success rates with other actions. Conversely, User 2 responds positively to *Socratic hints* but is not receptive to *direct commands*, with some other actions having a medium success rate. User 3 has more mixed success rates across different actions, with varying preferences and responses to agent interventions.

At the RL time steps $t = 50$ and $t = 150$, the distribution was shuffled by **inverting** the success rate probabilities to evaluate how well the learning approach copes with non-stationary success rates of strategies and to verify how quickly the agent can adapt to it.

Alg. 5 shows the user simulation algorithm. We made two simplifications compared to real-world scenarios:

1. We assumed that simulated users always request an argument that would reduce their AVQ, and, as a result, the intervention was applied at every interaction step. This simplification aimed at speeding up the simulation since our primary interest lies in determining whether the agent can learn the

optimal intervention strategy. Therefore, simulating interactions without RL is unnecessary as they do not influence the outcome of the learned strategy.

2. Additionally, we assumed a flat hierarchy of the argument structure, meaning all arguments directly target the root argument Φ_0. This simplification is made because including information about the specific argument and its level selected by a real-world user would unnecessarily complicate the simulation process as the intervention strategy does not depend on it.

Algorithm 5: User Simulation

```
Data: Noise probability $\beta \geq 0$
foreach User $u_1, \ldots, u_{100}$ do
    Data: User stance $z_{\Phi_0} \in [0,1]$, Focus $F = 0$, compute AVQ
    \\Init success rates for all actions
    foreach $a \in \mathcal{A}$ do
        $\mathcal{P}_{success}(a) \leftarrow$ random(0,1)
    foreach $t = 1, \ldots, n$ do
        if $t == 50$ or $t == 150$ then
            shuffleSuccessRates()      // Assign new success rates
        $s_t \leftarrow$ observe state.                          // Def. 6.1
        $a_t \leftarrow$ according to strategy $\pi$.
        $success \leftarrow$ random(0,1).              // Success simulation
        $noise \leftarrow$ random(0,1).                  // Noise simulation
        if $success \leq \mathcal{P}_{success}(a_t)$ and $noise > \beta$ then
            $\mathcal{R}_{t+1} \leftarrow 1$.                            // Def. 6.3
        else
            $\mathcal{R}_{t+1} \leftarrow -1$.                           // Def. 6.3
        $F \leftarrow$ updateFocus()          // Considering $z_{\Phi_0}$ and $\mathcal{R}_{t+1}$
        AVQ $\leftarrow$updateAVQ()                           // Using $F$
        $s_{t+1} \leftarrow$ observe state.                      // Def. 6.1
        $Q_\omega(s_t, a_t) \leftarrow$ update($s_t, \mathcal{R}_{t+1}, s_{t+1}$)       // Eq. 2.54
```

Since we use continuous values in the state space, we employ Linear Function Approximation (see Sec. 2.1.4.2) along with a $7th$-order Fourier Basis transformation as described in Sec. 2.1.4.3 with coupling = 2. The parameters were chosen not too high not to slow down learning, but high enough to learn possible indirect proportional correlations among state features (Weber, 2017).

Following Def. 6.1, it is easy to verify that $\forall \phi_i(s) \in \boldsymbol{\phi}(s) : \phi_i(s) \in [-1, 1]$, thus the convergence criteria of Eq. 2.70 is fulfilled. The learning rate α is set to $dim\overline{\boldsymbol{\phi}}(s)^{-1}$ following Eq. 2.73.

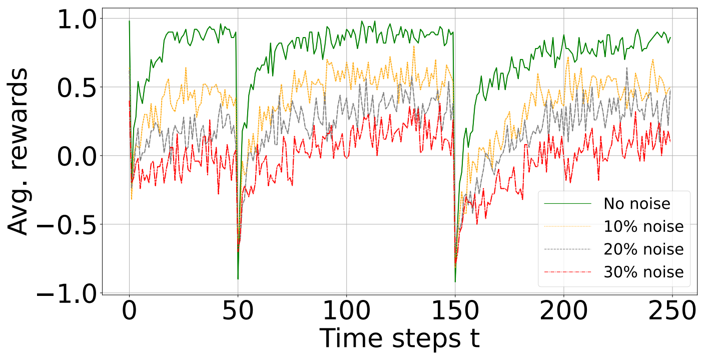


Figure 6.5: Experimental results.

Figure 6.5 shows the average reward over all 100 users. At time steps 50 and 150, we see the impact of the changed non-stationary success rates showing an immediate drop. Noise (0%, 10%, 20%, 30%) simulates random (*non-deterministic*) failure rates of the chosen actions (Ritschel et al., 2017).

We can see that even with high noise (30%), the agent can re-learn its optimal strategy. We also see that learning is robust without noise and reaches its maximum. Negative rewards are due to exploration. Increased noise leads to more unsuccessful actions and, thus, lower average rewards. However, we can see that even with high noise (30%), the average reward hardly falls below 0, which means that the agent's average reward is positive (= indicating success).

This suggests that the agent's intervention strategy is resilient to a certain degree of unpredictability in user behavior. The ability of the agent to maintain positive performance despite high noise levels means it can effectively handle scenarios with high uncertainty, such as varying user preferences or misunderstandings due to speech recognition errors.

At time steps 50 and 150, where the success rates change, we see an immediate drop in the average reward as expected and can see how the agent easily adapts to the new situation. In reality, this worst-case scenario hardly happens as success rates are more likely to change gradually; thus, it only demonstrates the robustness of the adaptation approach in worst-case situations.

6.3 Key Points & Summary

Within this chapter...

- ... we described the (non-)gamified intervention strategies of the agent.
- ... we described an adaptation approach of the linguistic style.
- ... we conducted a user simulation with *non-deterministic* and *non-stationary* user behavior to evaluate the adaptations' robustness and feasibility in dynamic and realistic scenarios.

In this chapter, we described the intervention strategies designed to encourage users to explore arguments that challenge their own positions, fostering a more comprehensive understanding of the topic. By leveraging the AVQ metric, the agent can identify when users primarily focus on arguments that align with their existing views and intervene to redirect their attention toward more challenging arguments.

We specifically developed three types of intervention strategies: 1) a non-adaptive base strategy without *gamification*, 2) a non-adaptive base strategy with *gamification*, and 3) an adaptive strategy that adjusts the agent's *linguistic* style. The *base strategy* intervenes based on an intervention condition and prompts the intervention without further context.

As the base strategy can lack intrinsic motivation for users to follow, we implemented a *gamification* strategy. Gamified elements can increase users' motivation and engagement by making the interaction more enjoyable and rewarding. We described several gamification strategies and employed the *points* strategy, as this can be easily integrated with our metric AVQ during the interaction, unlike other strategies such as *badges*, *levels*, and *leaderboards*. We display three scores: one reflecting the user's current points, one reflecting the user's choice, and another indicating the score they would receive by accepting the agent's suggestion.

While the *base strategy* intervenes based on an intervention condition without modifying the intervention text, there is a chance that users might not accept the intervention due to personal preferences. Thus, we implemented an adaptive approach based on the linguistic style, employing the politeness theory. Research suggests that varying verbal strategies can impact perceived persuasiveness, with more polite suggestions potentially being more effective than direct commands, depending on the user's preferences. Given these uncertainties, an adaptive approach using RL was implemented to tailor interventions dynamically.

We tested our adaptation approach in an experimental setup with simulated users. During this simulation, we considered *non-deterministic* and *non-stationary* user reactions to account for unpredictable user behavior that varies over time, allowing us to test how well the adaptation performs in dynamic and realistic scenarios.

6.4 Relevant Publications

- Weber, K., Ritschel, H., Aslan, I., Lingenfelser, F., & André, E. (2018a). How to Shape the Humor of a Robot - Social Behavior Adaptation Based on Reinforcement Learning. *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI)*, 154–162
- Weber, K., Janowski, K., Rach, N., Weitz, K., Minker, W., Ultes, S., & André, E. (2020a). Predicting Persuasive Effectiveness for Multimodal Behavior Adaptation using Bipolar Weighted Argument Graphs. *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 1476–1484
- Weber, K., Rach, N., Minker, W., & André, E. (2020b). How to Win Arguments: Empowering Virtual Agents to Improve Their Persuasiveness. *Datenbank-Spektrum*, *20*, 161–169
- Rach, N., Weber, K., Yang, Y., Ultes, S., André, E., & Minker, W. (2021). EVA 2.0: Emotional and Rational Multimodal Argumentation between Virtual Agents. *it - Information Technology*, *63*(1), 17–30
- Weber, K., Aicher, A., Minker, W., Ultes, S., & André, E. (2023a). Fostering User Engagement in the Critical Reflection of Arguments. *Proceedings of the 13th International Workshop on Spoken Dialogue Systems (IWSDS)*, 1–16
- Weber, K., Hogh, N., Conati, C., & André, E. (2024). A Gaze into Argumentative Chatbots: Exploring the Influence of Challenger Arguments on Reflection and Attention. *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents (IVA)*, 1–10

Chapter 7

User Studies

“The mind once enlightened cannot again become dark.” (Thomas Paine)

This chapter answers the research questions **Q2.2** - **Q2.7**:

Q2.2: *“Does the intervention mechanism impact the user’s engagement with challenger arguments, i.e., leads to an increase of AVQ?”*

Q2.3: *“Do the gamification mechanism and agent embodiment affect intervention success positively?”*

Q2.4: *“Does agent embodiment affect system perception, trust, and the user’s CE positively?”*

Q2.5: *“Do interventions affect user trust negatively?”*

Q2.6: *“Do interventions impact users’ eye gaze behavior (attention to arguments)?”*

Q2.7: *“Is there an interaction effect of User Characteristics (UCs) on the exploration behavior, and agent interactions?”*

In this chapter, we answer the research questions **Q2.2** - **Q2.7**. We conducted three studies to investigate the main and interaction effects as summarized in Tab. 7.1:

Table 7.1: Overview of investigated main (+), no main (−), and interaction (∗) effects derived from the research questions.

Study	Effects Type	Hyp.	RQs	Chapter
First	+ *Intervention on AVQ*	***H1*** [✓]	**Q2.2**	*7.1.2*
	∗ *Embodiment × intervention on AVQ* ∗ *Gamification × intervention on AVQ*	***H2*** [✗]	**Q2.3**	
	+ *Embodiment on system perception* + *Embodiment on trust* + *Embodiment on CE*	***H3*** [✓]	**Q2.4**	*7.1.3*
Second	+ *Intervention on AVQ*	***H1*** [✓]	**Q2.2**	*7.2*
	+ *Intervention on argument visitation*	***H4*** [✓]		
	− *Intervention on user trust*	***H5*** [✓]	**Q2.5**	
Follow-Up	+ *Intervention on eye gaze*	***E*** [16] [✓]	**Q2.6**	*7.3*
	∗ *UCs × intervention on AVQ* ∗ *UCs × intervention on* eye gaze		**Q2.7**	

In the first study, we investigate the effect of the agent's *embodiment* on CE and, as an additional variable, on system perception and trust. We also examine the effect of *intervention* on AVQ, including the interaction effects of *embodiment* × intervention and *gamification* × intervention on AVQ. This is done with fewer participants in a preliminary setup.

Based on the preliminary findings, we conduct a second study to investigate the effect of *intervention* on AVQ and argument visitation focus, i.e., engagement with *challenger arguments*. We also examine the impact of *intervention* on trust, as a decline in user trust could reduce the user's willingness to interact with the intelligent agent.

In the third study (follow-up), we explore the effect of *intervention* on eye gaze and investigate the interaction effects of UCs on AVQ and eye gaze. This helps us

[16] The follow-up study was conducted as an exploratory study. Thus, no hypotheses were defined.

gain a deeper understanding of which user personality traits most influence the success of the intervention.

Based on the research questions, five hypotheses were formulated:

H1 [✓] There is a *main effect* of *intervention* on AVQ.

H2 [✗] There are *interaction effects* between *gamification/embodiment* and the *intervention strategy* on AVQ.

H3 [✓] The *embodied virtual agent* leads to a more natural and engaging interaction experience and affects user trust positively.

H4 [✓] The intervention mechanism leads to increased engagement with *challenger arguments*.

H5 [✓] The intervention mechanism has no negative impact on user trust in the system, i.e., there is no *main effect* of *intervention* on *user trust*.

7.1 Study 1: Effects of Gamification and Embodiment

The first study aimed. . .

- . . . to assess potential existing effects of the research questions **Q2.2** and **Q2.3**, and the derived hypotheses. . .
  - **H1** (main effect of *intervention* on AVQ).
  - **H2** (interaction effect of *gamification/embodiment* on AVQ).
- . . . to answer the research question **Q2.4** and the derived hypothesis. . .
  - **H3** (main effect of *embodiment* on *system perception/trust/*CE).

7.1.1 Participants, Apparatus, and Procedure

The *first* study was conducted in cooperation with Ulm University online (31st May -23rd June 2022) via the crowd-sourcing platform "Crowdee" [17] with 51 English native speakers from the UK, US, and Australia (aged 18-65, $\mu = 34.1$, $\sigma = 8.6$; 34 female, 17 male) without a topic-specific background. The participants were divided into six groups (at least six people per condition, see Tab. 7.2).

[17] https://www.crowdee.com/

Table 7.2: Conditions: 1) Experimental (Intervention = Yes) and 2) Control (Intervention = No). The experimental condition is divided into sub-conditions with the co-variates *gamification* and *embodiment*.

	Intervention		
	Yes		No
Embodied	**w/o gamification**	**w/ gamification**	
No	G1	G2	G0
Yes	G4	G5	G3

All participants were given an introductory text explaining how to interact with the system, i.e., how they can ask for *pro* and *con* arguments, how the displayed argument graph is read, how the feedback (*agree*, *disagree*) buttons are used to express their opinion if they have one. The participants were not told anything about the underlying metric but only to select at least ten arguments to build a well-founded opinion on the topic *Marriage is an outdated institution*. To ensure they understood the interaction, we had them type a test command to request a pro argument, which the system validated before proceeding.

During the study, we collected the following data (see Fig. 7.4) anonymously [18]:

1. Calculated metric score AVQ (Fig. 7.4(a)).
2. Trust questionnaire by Körber (2019) (Fig. 7.4(b), see Appx. B.4.2).
3. User stance z_{Φ_0} (Fig. 7.4(c))
4. Set of visited arguments $Args_v^+$ and $Args_v^-$ (Fig. 7.4(d)).

To assess the initial user stance z_{Φ_0}, they were asked to rate their opinion on the topic on a 5-point Likert scale. In addition, demographic data was collected. After the conversation, the participants rated statements on different Likert-Scales concerning the interaction taken from a questionnaire according to ITU-T Recommendation P.851 [19] (Möller, 2003) (see Appx. B.4.1). To further investigate the CE for co-variate *embodiment* (see Sec. 7.1.3), the questionnaire by O'Brien et al. (2018) was employed (see Appx. B.4.3), a survey that measures CE as the "*quality of user experience [...] when interacting with a digital system*" (O'Brien et al., 2018). We further collected general statistics, the dialogue history (utterances), and the

[18] In line with the applicable privacy policy each user has voluntarily agreed to.

[19] Such questionnaires can be used to evaluate the quality of speech-based services.

user's argument agreement/disagreement (indicated by a click on the respective button).

7.1.2 Analyzing Hyp. H1 and H2: Effects on AVQ

For the *first study*, we investigated AVQ as dependent variable following the research questions **Q2.2** and **Q2.3** and derived hypotheses. . .

- **H1** (main effect of *intervention* on AVQ)
- **H2** (interaction effect of *gamification*/*embodiment* on AVQ)

Table 7.3: First study: The AVQ score, the number of participants, and the standard deviation in brackets per condition.

	Intervention		
	Yes		**No**
Embodied	**w/o gamification**	**w/ gamification**	
No	*0.95 (6, 0.05)*	*0.93 (9, 0.05)*	*0.80 (6, 0.12)*
Yes	*0.94 (13, 0.06)*	*0.93 (10, 0.04)*	*0.87 (7, 0.08)*

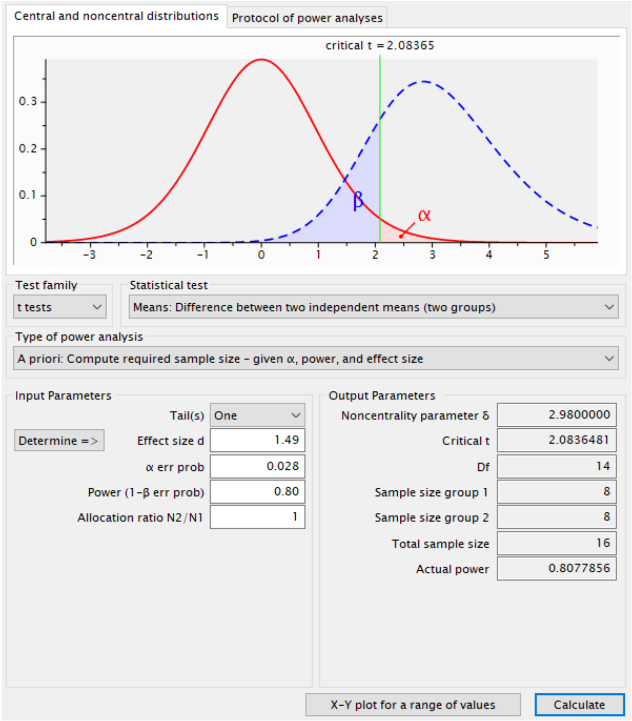
Tab. 7.3 summarizes the collected AVQ scores (see Appx. B.3.1 for data). We did an Analysis of Variances (ANOVA) with *intervention* as the dependent variable and *gamification*/*embodiment* as co-variates (see Tab. 7.4).

Table 7.4: Multi-factor ANOVA with main effect *intervention* on AVQ and interaction effects *intervention* × *gamification* and *intervention* × *embodiment*.

	df	Squares' sum	Mean squares	F	p	
intervention	*1*	*0.102*	*0.102*	*19.71*	*<.001*	✓
interv. × gami.	*1*	*0.003*	*0.003*	*0.58*	*.44*	✗
interv. × embod.	*1*	*0.01*	*0.01*	*2.41*	*.12*	✗
Residual	*46*	*0.239*	*0.005*	-	-	
η_p^2	0.28					

The ANOVA analysis revealed a significant main effect of *intervention* on AVQ ($p < .001$). However, no significant interaction effect between the *gamification* covariate and AVQ was observed ($p = .44$). Also, no significant interaction effect between the *embodiment* covariate and AVQ was found ($p = .12$).

To further explore the main effect of the *intervention* condition (post-hoc), we conducted a Student's t-test without checking for the assumptions of homogeneity of variances and normality in the first study data (see Tab. 7.3). The applied t-test revealed a strong [20] significant main effect of *intervention* on AVQ ($p = .028$; $d = 1.49$).


Central and noncentral distributions
Protocol of power analyses
critical t = 2.08365
0.3
0.2
0.1
0
-3
-2
-1
0
1
2
3
4
5
β
α
Determine =>
X-Y plot for a range of values
Calculate

Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Tail(s)	One
Effect size d	1.49
α err prob	0.028
Power (1-β err prob)	0.80
Allocation ratio N2/N1	1

Output Parameters

Noncentrality parameter δ	2.9800000
Critical t	2.0836481
Df	14
Sample size group 1	8
Sample size group 2	8
Total sample size	16
Actual power	0.8077856

Figure 7.1: Power analysis using the effect size d and the p-value.

Due to the absence of the user's speech input, which would introduce an additional source of errors, an asymmetry between *agent speech* and *chat input* would be the consequence. Consequently, we have chosen to employ the *chat-based* agent to evaluate the effectiveness of *intervention* on AVQ in the second study rather than using the *embodied* agent. Using an embodied agent can lead to negative gender effects (Siegel et al., 2009; Wessler et al., 2022), which the chat-based agent mitigates. Thus and further based on the first study results, we decided to omit conditions G2 - G5 and performed a power analysis for condition G1 using the *G*Power 3* tool (Faul et al. (2007), see Fig. 7.1). The analysis showed that a minimum sample size of 16 is required to achieve a minimal power of 0.80.

[20] Following J. Cohen (2013), Field (2013), and Tomczak and Tomczak (2014), we consider the effect size as **small** for $d, r < .3, \eta_p^2 \leq .06$, as **moderate** for $d, r < .5, \eta_p^2 \leq .14$ and as **large** for $d, r \geq .5, \eta_p^2 = \geq .14$.

7.1.3 Analyzing Hyp. H3: System Perception, Trust, and CE

Using the data of the *first study*, we conducted a deeper analysis of co-variate *embodiment* on system perception, trust, and CE addressing research question **Q2.4** and the derived hypothesis. . .

- **H3** (main effect of *embodiment* on *system perception*/*trust*/CE).

This section was previously published by the author in a similar form in peer-reviewed papers (Aicher et al., 2023, 2024) in cooperation with the University of Ulm.

In this section, we show the analysis of the co-variate *embodiment* concerning the following data:

1. Dialogue interaction metrics, such as time.
2. System Perception questionnaire (Möller, 2003).
3. Trust questionnaire (Körber, 2019).
4. User Engagement questionnaire (O'Brien et al., 2018).

The analysis was done with the 51 participants from Sec. 7.1 and 33 additional participants (aged 18-57 ($\mu = 37.6$, $\sigma = 11.2$); 16 female, 17 male) that were collected at the same time during the first study using the same system but with a different intervention metric (not reported in this thesis). The participants were divided into two groups based on the co-variate *embodiment*:

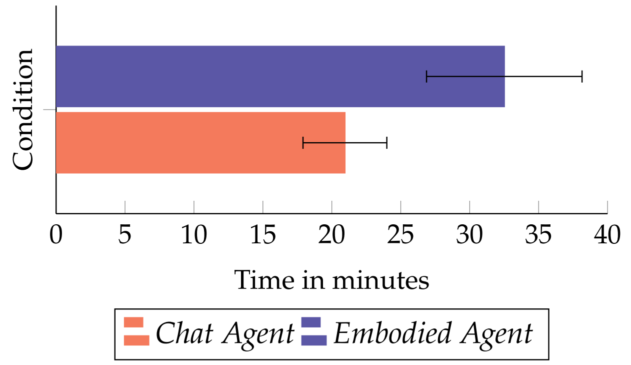


Figure 7.2: Interaction time of participants between both groups *chat* vs. *embodied*. Error bars denote 95% confidence interval.

46 participants interacted with an *embodied agent* interface further on called *embodied condition*, and 38 participants with a *chat-based* agent further on called *chat condition*.

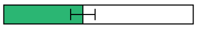
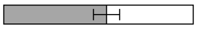
On average the participants interacted significantly longer (p = .013, r = .34) with the dialogue system for 32.52 minutes (σ = 19.52) in the *embodied* and for 20.95 minutes (σ = 9.57) in the *chat* condition (see Fig. 7.2).

Results of the questionnaires (ITU-T Tab. 7.5, CE Tab. 7.7, Trust Tab. 7.6) are displayed in the following tables. For readability, we only show excerpts of the most interesting data. A complete overview is provided in Appx. B.4.

Table 7.5: ITU-T - Excerpt. Means ($\mu_{embodied}$, μ_{chat}) and 95% confidence intervals denoted by bars of the questionnaire items regarding the user's perception of the system (Möller, 2003) (1 = *strongly disagree*, 5 = *strongly agree*, (**) 1 = *Bad*, 2 = *Poor*, 3 = *Good*, 4 = *Fair*, 5 = *Excellent*) grouped by the following categories: information provided by the system (IPS), communication with the system (COM), system behavior (SB), dialogue (DI), user's impression of the system (UIS), acceptability (ACC), argumentation (ARG), and overall quality (QLT). See Appx. B.4.1 for the full table. Significant values are check-marked [✓]. (*) Items have to be inverted.

Cat.	Question	Embod.	$\mu_{embodied}/\mu_{chat}$	p	r	
SB	*5. The system reacted naturally.*	*Yes*	**3.27**	**.037**	.227	✓
		No	2.78			
	*8. The system reacted too slowly.**	*Yes*	3.29	**<.001**	.485	✓
		No	**2.19**			
	*10. The system's responses were too long.**	*Yes*	2.58	**.016**	.263	✓
		No	**2.03**			
DI	*1. You perceived the dialogue as natural.*	*Yes*	**3.52**	**.032**	.234	✓
		No	3.03			
	*3. The dialogue was too long.**	*Yes*	2.42	0.093		✗
		No	**2.11**			
	*6. You would have expected more help from the system.**	*Yes*	3.75	**.014**	.269	✓
		No	**3.19**			
UIS	*5. You felt relaxed during the dialogue.*	*Yes*	3.42	0.146		✗
		No	**3.69**			
ACC	*2. You would recommend the system to a friend.*	*Yes*	**3.21**	0.067		✗
		No	2.75			
ARG	*1. I felt motivated by the system to discuss the topic.*	*Yes*	**3.64**	0.068		✗
		No	2.94			
	5. I felt engaged in the conversation with the system.	*Yes*	**3.40**	**.039**	.226	✓
		No	2.83			
	*7. I do not like that the arguments are provided incrementally.**	*Yes*	3.04	0.111		✗

. . . continued

		No	 ***2.67***			
QLT**	*1. What is the overall impression of the system?*	*Yes* *No*	***3.67***  *3.17*	*.047*	*.216*	✓

For the evaluation of the self-assessment questionnaires, the means μ_{chat} and $\mu_{embodied}$ and confidence interval were determined for each single item and condition. Regarding all items, the assumption of a normal distribution based on the Shapiro-Wilk-Test had to be discarded ($W = 0.696 - 0.913$, $p < .001$). Thus, we again used the non-parametric *Mann-Whitney U* test (McKnight & Najab, 2010) for two independent samples with no specific distribution. We conducted an exploratory study and thus refrained from employing multiple test correction methods. Concerning the overall perceived quality (QLT), our analysis shows a statistically significant difference ($p = .047$) between the two conditions. The *embodied* condition received an averaged rating of 3.67 ($\sigma = 0.930$), outperformed the *chat* condition with 3.17 ($\sigma = 1.082$). This difference is considered moderate, as indicated by the effect size $r = .216$.

As shown in Tab. 7.5 (for a complete overview see Sec. B.4.1), the single item analysis between both groups does not show any significant differences regarding the categories *information provided by the system* (IPS), *communication with the system* (COM), *user's impression of the system* (UIS) and *acceptability* (ACC).

A significant difference is notable regarding the category *system behavior* (SB) in three single items (*natural system reaction* (SB 5 [21]), *system's response speed* (SB 7) and *length* (SB 10)). Concerning the category *Dialogue*, two single items (*naturalness of the dialogue* (DI 1) and *expected help* (DI 6)) showed a significant difference between the two groups (effect sizes: $r_{DI\ 1} = .234$, $r_{DI\ 6} = .269$). Concerning the category *argumentation* (ARG), we observed a significant difference ($r_{ARG\ 5} = .226$) in the single item *engagement induced by the system* (ARG 5).

When the single items or their inverted counterparts marked with (*) are summarized in their associated categories, there is no significant difference (p= .290- .993) perceivable for any of these merged categories.

In Table 7.6, the single item results for the trust questionnaire (Körber, 2019) are shown to analyze the user trust regarding the argumentative dialogue system. In three single items, a significant difference with a moderate effect size ($r_{PT\ 3} = .244$, $r_{PT\ 1} = .250$, $r_{TA\ 1} = .244$) between the *embodied* and the *chat* condition is perceived. Merging the single items (inverted counterparts respectively) into their associated

[21] The numbers following the abbreviations correspond to the respective question numbers.

four categories (UP, F, PT, TA) results in a significant difference for PT ($p = .015$, $r = .266$).

Table 7.6: Trust - Excerpt. Means ($\mu_{embodied}$, μ_{chat}) and 95% confidence intervals denoted by bars of the questionnaire items regarding user trust (Körber, 2019) (1 = *strongly disagree*, 5 = *strongly agree*) grouped by the following categories: understanding/predictability (UP), familiarity (F), propensity to trust (PT) and trust in automation (TA). See Appx. B.4.2 for the full table. Significant values are check-marked [✓]. (*) Items have to be inverted.

Cat.	Question	Embod.	$\mu_{embodied}/\mu_{chat}$	p	r	
PT	*1. One should be careful with unfamiliar automated systems.**	*Yes*	**3.46**	***.025***	*0.250*	✓
		No	*3.97*			
	3. Automated systems generally work well.	*Yes*	**3.25**	***.022***	*.244*	✓
		No	*2.72*			
TA	*1. I trust the system.*	*Yes*	**3.44**	***.039***	*.244*	✓
		No	*2.97*			

Table 7.7 displays the results of the short form of the user engagement scale introduced by O'Brien et al. (2018). Even though only one of the single items did show a significant difference with a moderate effect size ($r_{\texttt{RW 1}} = .250$), the *embodied* condition was rated better than the *chat* condition in every item. When merging the items (inverted counterparts respectively) into their associated four categories (FA, PU, AE, RW), no significant difference ($p = .061 - 0.358$) was perceived.

Table 7.7: CE - Excerpt. Means ($\mu_{embodied}$, μ_{chat}) and 95% confidence intervals denoted by bars of the items of the short user engagement questionnaire (O'Brien et al., 2018) (1 = *strongly disagree*, 5 = *strongly agree*) grouped by the following categories: Focused attention (FA), perceived usability (PU), aesthetic appeal (AE) and reward factor (RW). See Appx. B.4.3 for the full table. Significant values are check-marked [✓].

Cat.	Question	Embod.	$\mu_{embodied}/\mu_{chat}$	p	r	
FA	*1. I lost myself in this experience.*	*Yes*	**2.63**	*.121*		✗
		No	*2.19*			
	3. I was absorbed in this experience.	*Yes*	**3.08**	*.144*		✗
		No	*2.69*			

. . . continued

Cat.	Question	Embod.	$\mu_{embodied}/\mu_{chat}$	p	r	
AE	*2. The application was aesthetically appealing.*	*Yes*	**3.25**	.174		✗
		No	3.06			
RW	*1. Using the application was worthwhile.*	*Yes*	**3.38**	**.022**	.250	✓
		No	2.86			
	2. My experience was rewarding.	*Yes*	**3.40**	.076		✗
		No	2.94			

7.1.4 Discussions

7.1.4.1 Impact of Embodiment on System Perception, Trust and CE

Participants generally interacted significantly longer with the *embodied* condition than the *chat* condition. This can be explained by the fact that listening to the spoken utterance of the embodied agent and re-reading the respective response in the dialogue history takes longer than just reading the displayed answer. Furthermore, sometimes the reaction time in the *embodied* condition was longer compared to the *chat* condition due to the longer processing time of the Text-To-Speech System (TTS) system.

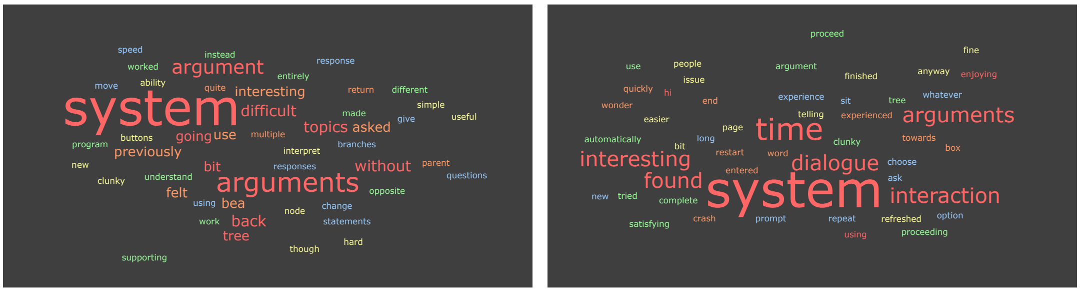
Even though no main effect of the aggregated values (FA, PU, AE and RW, see Tab. 7.7 and B.4.3) on CE was found, we can observe a strong tendency towards the *embodied* condition as it was rated better in every item. The overall engaging main effect (CE) is supported by ARG 5 which states that *the users felt significantly more engaged* in the *embodied* condition, and (RW 1), which is significant for *the impression that using the system was worthwhile*.

Still, the rating of *perceived usability* (PU) indicates there is a need for enhancement, especially regarding the errors and the explanation of the system's reaction if the user is not understood correctly. This is also observable in Tab. 7.5 where the system's reaction time (SB 8) and responses (SB 10) were rated significantly worse in the *embodied* condition. Unfortunately, this delay is caused by a necessary external server access. Depending on the connection quality and load, this results in different system response times. In contrast, the response of the *chat* setting is generated internally and presented immediately on the interface.

Even though the differences between the *embodied* and *chat* condition regarding the aggregated categories IPS, COM, SB, DI, UIS, ACC and ARG (see Tab. 7.5, for complete overview see Tab. B.4.1) were insignificant, we can perceive a consistent, category-overlapping tendency. Especially regarding the perceived naturalness (SB 4, DI 1), a significant influence of the *embodied* condition compared to the *chat* condition is observed. These findings imply that an *embodied* condition seems to influence the user's perception of naturalness during interaction, underpinning the claim in state-of-the-art literature that embodied agents can be used to design a more human-like, natural conversation.

The significant difference in the *expected help* the system should have provided (DI 6) implies that the *embodied* condition, on the one hand, tends to raise the expectation to the one of a human conversational partner, and on the other if these expectations are met could lead to a significantly stronger acceptance comparable to a human conversational partner.

The voluntary free-text remarks of participants on the study and the system underpin this. Fig. 7.3 shows the generated word clouds of the *free-text* answers.



(a) Chat-based agent

(b) Embodied agent

Figure 7.3: Word clouds highlighting the most common words in the free text answers.

Participants in the *chat* condition mainly commented on the performance and technical limitations and suggested potential ways to enhance it (e.g., "*the system might as well have been a list of pros and cons. . .*", "*speed up the results would make the experience better*" or "*it would be good if there are more buttons representing the options to interact with BEA or a drop-down menu of the previously typed words*"). In the *embodied* condition, participants focused their comments more on the overall impression of the system and suggestions for improvement that increase the naturalness and flexibility of the dialogue ("*I think the system needs more detailed arguments not just stating statistics*" or "*[. . .] The system moved very smoothly and was engaging. I found this study extremely interesting and very valid as we move toward more automated systems*").

The results in Tab. 7.6 and Tab. B.4.2 implicate that users seem to have a tendency to trust the *embodied* condition more than the *chat* condition, especially regarding the propensity to trust (PT) and trust in automation (TA 1). The fact that participants within the *embodied* condition are not significantly more familiar with similar systems (F 1) implies that an *embodied agent* interface can help users access arguments intuitively and positively influence the user's trust. This finding supports our decision to use an *embodied agent* and could help increase user trust by individualizing the agent. This study is subject to three limitations:

1. First, we did not compare different embodied agent settings personalized to the user, such as gender, age, variation of motion, mimics, or realism, which have a great impact on the user's social perception (Wessler et al., 2022; Zanbaka et al., 2006). We chose an easily implementable, commonly accessible, representative embodied agent rather than a highly individualized one. The comparison to a purely chat-based agent aimed to evaluate the influence of avatars on argumentative interactions in general. The focus was to determine whether the mere visualization of an avatar leads to a bias in opinion formation or influences the perception of the provided argumentative content and conversational engagement of users.
2. Second, as prior research studies relevant to our observations are limited and no comparable results exist, our quantitative analysis focuses on the participants' self-assessment answers and argument exploration behavior. In future work, the validity of our findings would be strengthened if they could be compared to a baseline and supported by qualitative analysis (e.g., free text responses of users).
3. Third, during interaction with the embodied agent, instances occurred where the time required for a response was perceived by the participants to be excessively delayed, which should be addressed in future research.

Still, we can conclude that the embodied agent exhibited significantly higher overall quality compared to its chat-based counterpart. Addressing the aforementioned limitations and adapting the embodied agent to the user's needs and preferences could further enhance this impression.

7.1.4.2 Impact of Gamification and Embodiment on AVQ

We also investigated the interaction effect of *gamification* strategies on the user's AVQ. To facilitate this, we used the AVQ as a performance score and presented users with their current AVQ score alongside two alternative scores: one reflecting their choice and another reflecting the potential score if they followed the agent's intervention. Displaying the performance score as a gamified element aimed to incentivize users to explore challenger arguments actively. By framing the score as a performance measure and associating it with the potential for improvement, we sought to motivate users to consider alternative viewpoints more often.

However, the effectiveness of the gamified condition was similar to that of the non-gamified one. This raises the question: Did users perceive the score as meaningful feedback on their exploration behavior? The intrinsic motivational value of the performance score seems unclear. Users might need a benchmark for what constitutes a *good* or *bad* score to have a realistic goal rather than merely aiming for the maximum possible score (100), for which there might be no incentive to reach. Thus, a different scoring system might be more beneficial, such as awarding points for each visited challenger argument (including those chosen without intervention). Additionally, comparing high scores achieved by other users could further motivate users.

In the same regard, users likely did not fully understand the performance score, which led to a lack of intrinsic motivation. Clarity regarding the mechanics of the performance score computation and the implications of user actions are essential for user motivation. Users should understand how their actions affect the score and why exploring challenger arguments is beneficial.

It is also possible that the intervention's pop-up window, which displayed the three scores, disrupted the dialogue flow within the interaction. This could have distracted the users from the ongoing interaction, reducing the effectiveness of the gamification strategy. Alternatively, one could integrate the gamified interventions directly into the dialogue flow instead of showing a pop-up window. This way, users might be more engaged and motivated to change their behavior during interaction to earn points and avoid losing them, thereby fostering a more reflective argument exploration process.

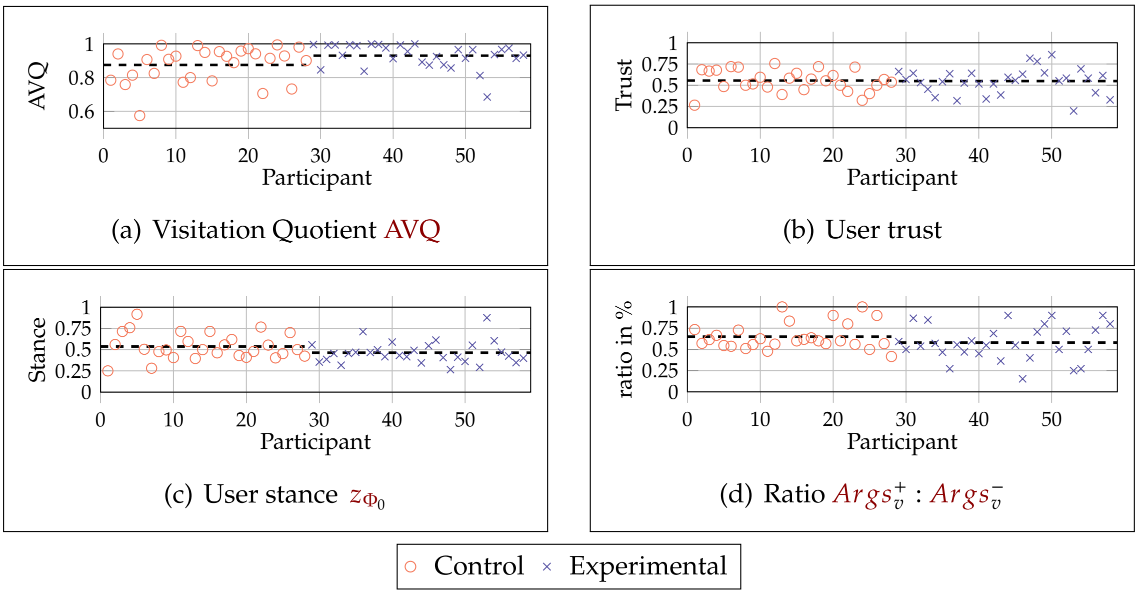
Besides *gamification*, we examined the interaction effect of agent *embodiment* on AVQ by showing an embodied agent instead of a chat-based agent. Although the embodied agent was rated higher than chat regarding enjoyment, engagement, and trust (as discussed in Sec. 7.1.4.1), it did not significantly motivate users to follow the interventions more often. While the embodied agent enhances the overall user experience, this does not necessarily translate into increased compliance with the intervention, i.e., users do not find the agent's presence compelling enough to change their behavior. One possible reason is that the embodied agent, despite being engaging and enjoyable, might not convey the importance of considering challenger arguments effectively. The effectiveness of embodiment depends on its personality traits and how well the agent communicates the rationale behind the interventions. If users do not perceive the agent as an authoritative or persuasive figure, such as a mentor, its ability to influence the user's choices does not increase compared to the *chat-based* agent. The adaptive intervention strategy (Sec. 6.2.2) could help overcome this limitation.

7.2 Study 2: Effects of Intervention

> The second study aimed to analyze the research questions **Q2.2** and **Q2.5**, and the derived hypotheses...
>
> - **H1** (main effect of *intervention* on AVQ).
> - **H4** (main effect of *intervention* on *argument visitation*).
> - **H5** (no main effect of *intervention* on *trust*).
>
> The section was previously published by the author in a similar form in peer-reviewed papers (Aicher et al., 2024; Weber et al., 2023a).

The second study was conducted with 60 participants from 1st July -09th July 2022 using the *chat-based* agent divided into two groups (control condition without intervention, experimental condition with intervention) to evaluate the effectiveness of the non-adaptive intervention strategy. The participant introduction and interaction procedures remained consistent with those outlined in Sec. 7.1.1.

Fig 7.4 plots the collected data. Appx. B.3.2.1 provides a complete overview.



(a) Visitation Quotient AVQ

(b) User trust

(c) User stance z_{Φ_0}

(d) Ratio $Args_v^+ : Args_v^-$

Figure 7.4: Collected data: Calculated metric AVQ, Self-reported user trust, user stance z_{Φ_0}, ratio of visited argument $Args_v^+ : Args_v^-$. Dotted lines denote the means.

Two participants had to be omitted either due to *corrupted data*, *unusual short response time*, *invalid response and interaction patterns* (e.g., all answers were identical or unusual fast system interactions).

7.2.1 Analyzing Hyp. H1, H4, and H5: AVQ, Argument Focus, and Trust

7.2.1.1 AVQ (H1):

Concerning the calculated metric AVQ (Fig. 7.4(a)), the homogeneity of variances was rejected utilizing the Levene's test ($F = 5.64$, $p = .021$) and the assumption of normal distribution was rejected using the Shapiro-Wilk test ($W = 0.895$, $p < .001$).

Thus, we applied the *Mann-Whitney-U test* (McKnight & Najab, 2010) to verify **H1** that there is a main effect of the intervention on AVQ, i.e., we checked if the score AVQ increased significantly in the experimental group. We found a moderate [22] significant effect ($U = 273$, $n_1 = 30$, $n_2 = 28$, $p \leq .01, r = .35$). In addition to that, we checked the total amount of interventions. There were 262 interventions in the experimental condition (8.73 per user), 201 (76%) of which were accepted by the user.

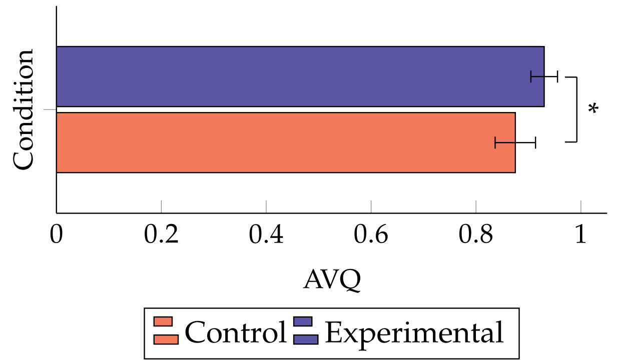


Figure 7.5: Means including 95% confidence interval denoted by bars of AVQ. (*) $p < .05$, (**) $p < .01$. See Appx. B.3.2.1 for data.

7.2.1.2 Engagement with Challenger Arguments (H4):

To test **H4**, we analyzed how many participants were more engaged with *challenger arguments* by counting the number of participants that heard more challenger arguments than arguments supporting their own stance, e.g., if the user stance was negative ($z_{\Phi_0} < 0.5$) and more pro than con arguments were heard ($Args_v^+ > Args_v^-$), it implicates a higher engagement with challenger arguments.

[22] Following J. Cohen (2013), Field (2013), and Tomczak and Tomczak (2014), we consider the effect size as **small** for $d, r < .3, \eta_p^2 \leq .06$, as **moderate** for $d, r < .5, \eta_p^2 \leq .14$ and as **large** for $d, r \geq .5, \eta_p^2 =\geq .14$.

We found that with intervention, nearly 80% of participants were more engaged with *challenger arguments*, while only 53% in the control condition did so, which is a total increase of 51% (see Tab. 7.8).

Table 7.8: Contingency table of engagement with challenger arguments per condition. Values in brackets show the number of arguments

	Engagement with ···		
Condition	Challenger arg.	Non-challenger arg.	Total
Experimental	*24 (374)*	*6 (261)*	*30 (635)*
Control	*15 (339)*	*13 (303)*	*28 (642)*
Total	*39 (713)*	*19 (564)*	*58 (1277)*

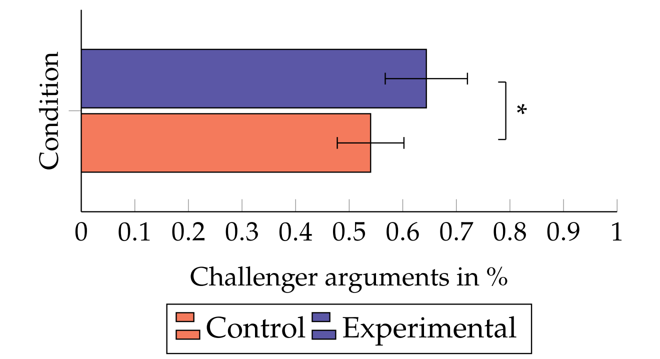


Figure 7.6: Means including 95% confidence interval denoted by bars of engagement with challenger arguments. (*) $p < .05$. See Appx. B.3.2.1 for data.

A chi-square test of independence (McHugh, 2013) was performed to examine the relation between *condition* and *engagement with challenger arguments* showing a significant relation, $(\mathcal{X}^2(1, N = 58) = 4.5924, p = .032)$.

Analyzing the main effect of intervention on the total percentage of heard *challenger arguments* (see Fig. 7.6) revealed a large significant main effect (*T-Test*, $t(56) = 2.0903, p = .02, d = .55$), showing that the users in the experimental group engaged significantly more with *challenger arguments* than in the control condition.

7.2.1.3 Trust (H5):

After the interaction, we asked the users about their trust in the system using the questionnaire of Körber (2019) consisting of six scales to measure trust.

- Reliability and competence scale (RC)
- Understanding and predictability scale (UP)
- Familiarity scale (F)

- Intention of developers scale (ID)
- Propensity to trust scale (PT)
- Trust in automation scale (TA)

Since certain scales were not applicable, we excluded RC and ID. The aggregated values of the remaining scales are depicted in Figure 7.4(b). Tab. 7.9 shows the aggregated values of the sub-scales separated by *metric* and *intervention*.

Table 7.9: Trust scales. Means (μ_{yes}, μ_{no}) and 95% confidence intervals denoted by bars of the questionnaire items regarding user trust (Körber, 2019) (1 = *strongly disagree*, 5 = *strongly agree*) grouped by the following categories: understanding/predictability (UP), familiarity (tba (F)), propensity to trust (PT), trust in automation (TA) and the respective aggregated values (AGG). Non-significant values with Bayes Factor (B_N) < $1/3^{rd}$ (Dienes, 2021) are check-marked [✓].

Cat.	Intervention	μ_{yes}/μ_{no}	p	d	B_N	
UP	*Yes*	***2.03***	*.418*	*0.055*	*0.12*	✓
	No	*2.00*				
F	*Yes*	***2.82***	***.006***	*0.678*	*4.73*	✗
	No	*2.20*				
PT	*Yes*	*3.31*	*.611*	*-0.074*	*0.12*	✓
	No	***3.35***				
TA	*Yes*	*3.10*	*.571*	*-0.048*	*0.23*	✓
	No	***3.14***				
AGG	*Yes*	*2.813*	*.764*	*-0.191*	*0.17*	✓
	No	***2.908***				

Because the conditions of normal distribution (Shapiro Wilk, $W = 0.989$, $p = .892$) and homogeneity of variances (Levene's test, $F = 0.25306$, $p = .617$) were met, we performed a Student's t-test on the trust score AGG. No significant difference was found between the baseline condition (without intervention) and the experimental condition ($t = -0.7254$, $p = .764$) in trust scores. A power analysis using the tool *G*Power 3* (Faul et al., 2007; 1 - $\beta - error$) resulted in a value of 0.84. We then computed the Bayes factor [23] (B) to assess strength of evidence (Dienes, 2014). The calculated Bayes factor $B_N(2.813, 0.48) = 0.17$ indicates moderate evidence that there is no difference in user trust because of being less than $1/3^{rd}$ (Dienes, 2021), thus, confirming **H5**. When looking into the sub-scales we also see that none of them is significant with moderate evidence except for `F` ($p = .006$, $d = 0.678$, $B_N(2.82, 0.48) = 4.27$) with moderate effect, which is unexpected as participants were randomly assigned to the groups during the study. There are various possible reasons for this:

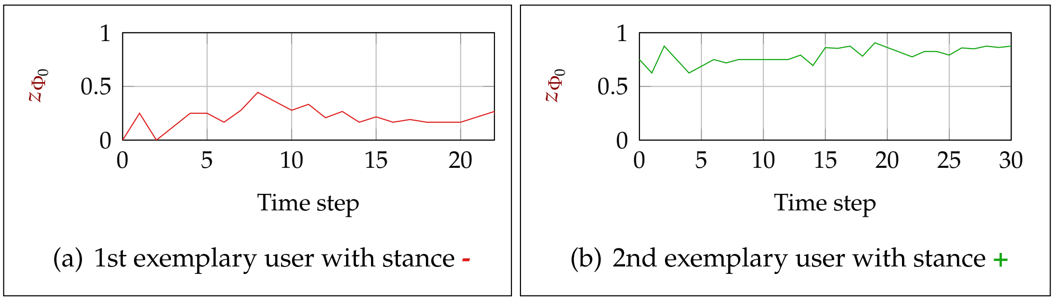
1. There may be unaccounted variables that differ between groups due to random variability (for instance leading to unexpected significant results.
2. When conducting hypothesis tests, there is always a chance of making a *Type I error*, where we incorrectly reject the null hypothesis.
3. The sample size may be too low and significance may not hold in a larger study.
4. There may be outliers in the data which cause the significant result.

Since the study was conducted online, reasons like *measurement bias* (Kopec & Esdaile, 1990) can be excluded. To better understand the unexpected significant result of sub-scale `F`, further investigation would be needed with additional control measures to shed light on the potential reasons for this unexpected finding. Thus, this result needs to be interpreted with caution. Nevertheless, the Bayes factors across all other sub-scales and the aggregated scales strongly indicate that the *intervention* has no negative impact on user trust.

7.2.1.4 Are Stances Oscillating?

As we used a dynamic stance estimator, the stance changed over time. It is thus necessary to avoid oscillating stance estimation. Only visited arguments are considered during stance estimation, and the system uses the initial subjective user opinion on the topic as the initial stance and incorporates the feedback given by the user to update the user's stance. We reviewed the stance data manually and verified no oscillating stance. In Fig. 7.7, we sketch two exemplary users and their estimated stances over time, showing that changes only happen gradually.

[23] https://harry-tattan-birch.shinyapps.io/bayes-factor-calculator, (Accessed on 04th August 2023).



(a) 1st exemplary user with stance -

(b) 2nd exemplary user with stance +

Figure 7.7: Two exemplary users with opposing stances and their estimated stance z_{Φ_0} over time depicting that the stance changes gradually over time depending on the user's provided feedback (*agree*, *disagree*).

7.2.2 Discussion of the Main Effects of Intervention

In the second study, we investigated the main effect of *intervention* on AVQ without considering *embodiment* or *gamification* using a *chat-based* agent. We found a significant and substantial main effect of *intervention* on AVQ. We further found a significant main effect of *intervention* on the overall user's engagement with *challenger argument*. This shows that the intervention mechanism along with the proposed metric leads to a significantly higher engagement with *challenger arguments*.

With respect to trust, it remains unclear why there was a significant main effect of *intervention* on the *familiarity scale* (F) (see Tab. 7.9), which can only be explained by some unaccounted variables causing random variability. Factors, such as *user characteristics* were not taken into account in this study.

Further, as the main effect of *intervention* was only investigated within the *chat-based* agent, there is an open question whether or not users actually read the arguments when they were presented to them. This however is crucial to assess whether or not users were more reflective during interaction.

Summarized, the second study is subject to five main limitations.

1. First, we did not investigate adaptive intervention strategies personalized to individual users. The focus was to determine whether the simple intervention strategy of suggesting opposing arguments already leads to a change in argument exploration. Future research should investigate the impact of personalized intervention strategies and how different agent's personality traits (e.g., dominance, friendliness/politeness) affect interventions' success and user perception.

2. Second, the user study focused on one topic (*Marriage is an outdated institution*) derived from a single source. We selected this topic because its dataset fulfills our criteria of being sufficiently large, balanced in terms of argument stance (pro/con), of high quality, and having depth in arguments. Thus, the reproducibility of our findings concerning other topics needs to be demonstrated in future research.
3. Third, while the user-agent interaction may seem constrained and artificial because users are unable to introduce counterarguments, this decision was deliberate. The aim of the intelligent agent was to neutrally confront users with pro/con arguments on a given topic, allowing them to explore without being directly engaged in a debate. As pointed out by Paul, 1990 due to the users' tendency to defend their own view, a system that confronts them with an opposing stance might not lead to an unbiased argument exploration but rather the opposite.
4. Fourth, although the *intervention* strategy exhibited a significant main effect, there is a lack of evidence confirming users' active reading of the arguments. We specifically aim for users to actively engage in processing the presented arguments, not merely navigate the system and follow interventions.
5. Last, we further did not investigate the main and interaction effects of UCs. Understanding how individual user traits may influence the effectiveness of the *intervention* strategy is essential for a comprehensive understanding of the intervention's efficacy to develop methods for personalization and enhancing the subjective experience in future work. Personality traits (e.g., Need for Cognition (NFC) and Locus of Control (LOC)) and cognitive abilities (e.g., Perceptual Speed (PS)) have been identified as significant factors affecting user behavior and performance (Conati et al., 2021; Toker & Conati, 2014; Toker et al., 2013; Ziemkiewicz et al., 2011). For instance, Ziemkiewicz et al. (2011) found that LOC influences user performance, while Toker et al. (2013) demonstrated that individuals with slower PS focus more on key elements in visualizations compared to those with faster PS.

The last two limitations are addressed in a final follow-up study (Sec. 7.3) in which we show among other interaction effects of *user characteristics* that the *intervention* strategy along with our proposed metric AVQ has significant main effects on *attention to arguments* (measured using eye-tracking data).

7.3 Follow-Up Study: Exploring UCs and Eye-Gaze

ℹ We conducted the follow-up study as an exploratory study without formulating specific hypotheses to answer research questions. . .

- **Q2.6**, i.e. the main effects of *intervention* on *eye gaze* (referred to as *attention to arguments*).
- **Q2.7**, i.e., interaction effects of UCs on AVQ, and *eye gaze*.

This section was previously published by the author in a similar form in a peer-reviewed paper (Weber et al., 2024).

To analyze the impact of BEA interventions on user's eye-gaze and attention to arguments (**Q2.6**), we collect eye-tracking data. Eye gaze is widely recognized as an indicator of attention (Beattie et al., 2017; Cheng & Yang, 2022; Cullipher et al., 2018). While the metric AVQ (see Sec. 7.2) assessed argument visitations and the user's exploration behavior, incorporating eye-tracking data in the second study enables us to explore how the agent's interventions influence the user's visual focus, and whether or not they pay different attention to arguments they agent suggested them using interventions.

We further examine the influence of UCs on interactions with the BEA system (**Q2.7**). As previously mentioned, addressing **Q2.7** will allow us to assess the influence of selected UCs on the interaction with BEA. Analyzing how UCs impact the intervention and user focus will provide insights into potential improvements and intervention strategies tailored to specific user groups, aiming to enhance the subjective experience and increase the intended goal of the agent.

7.3.1 Participants, Apparatus, and Procedure

The study was conducted in-person in our lab with 45 participants from the University of Augsburg (25 female, 20 male, aged 18-31) divided into two groups (an experimental condition with intervention and a control condition without intervention) using the same system as in Sec. 4.1 and 4.2.

After a short introduction and eye calibration, they were shown and explained in detail how the interaction with the system works, i.e., how they can ask for *pro* and *con* arguments, how the displayed argument graph is read, how the feedback (*agree*, *disagree*) buttons are used to express their opinion if they have one. The participants were not told anything about the underlying metric but only to select at least ten arguments to build a well-founded opinion on the topic *Marriage is an* *outdated institution*.

To ensure that they understood the interaction, the system prompted them to type a command to request a pro argument, which the system validated before proceeding. To assess the initial user stance z_{Φ_0}, they were asked to rate their opinion on the topic on a 5-point Likert scale.

During the study, we collected the following data anonymously [24].

1. **Independent Measures** UCs (see Sec. 7.3.2 for details).
2. **Objective Dependent Measures**:
   - Calculated metric AVQ.
   - Eye-Tracking data (see Sec. 7.3.3 for details).

7.3.2 User Characteristics (UCs)

The study participants took a collection of validated psychological tests to investigate the impact of various UCs (see Tab. 7.10 and Appx. B.3.3.3 for data) and address **Q2.7**. The collected UCs are utilized as co-variates during the analysis process. This allows us to assess the effects of UCs on interventions and attention to arguments (eye-gaze). We specifically investigated selected *personality traits*, *cognitive abilities*, *curiosity traits*, and *technical affinity*.

Table 7.10: Collected UCs along with the definitions and employed test scales.

UC	Definition	Test Scale
	Personality traits	
Need for Cognition (NFC)	Construct that measures the inclination of an individual to engage in challenging cognitive tasks (Bauer & Stiner, 2020; Cacioppo et al., 1984).	Need for Cognition Scale (Cacioppo et al., 1984)
Conscientiousness (CS)	Individual tendency to be self-disciplined, reliable, responsible, diligent, and structured (Barrick & Mount, 2012; Roberts et al., 2014).	Ten Item Personality Measure (Gosling et al., 2003)
Openness (OP)	Captures the individual differences in being open to new ideas, art, and values (McCrae & Sutin, 2009).	Ten Item Personality Measure (Gosling et al., 2003).

[24] In line with applicable privacy policy each user has voluntarily agreed to.

. . . continued

UC	Definition	Test Scale
Locus of Control (LOC)	Describes an individual's generalized belief system regarding the extent of control they have over their own life (Steca, 2021).	Rotters Internal-External Locus of Control Scale (Rotter, 1966)
	Cognitive abilities	
Perceptual Speed (PS)	Cognitive ability, which allows individuals to carry out simple tasks with visual information by measuring the speed for comparing figures and shapes (Ekstrom & Harman, 1976; Gnambs et al., 2021).	P-3 Identical Pictures Test (Ekstrom & Harman, 1976)
Visual Working Memory (VWM)	The capacity of information (e.g., colors and shapes) that can be temporarily retained and manipulated (Mance & Vogel, 2013; Vogel et al., 2001).	Colored Squares Sequential Comparison Task (Vogel et al., 2001)
Reading Proficiency (RP)	Measures the English vocabulary size and reading comprehension ability (Meara, 1992).	XLex Vocabulary Test (Meara, 1992)
	Curiosity traits	
Deprivation Sensitivity (DS)	The desire of individuals to close or reduce gaps in their knowledge, as they cause anxiety or tension (Kashdan et al., 2018).	Five-Dimensional Curiosity Scale (Kashdan et al., 2018)
Social Curiosity (SC)	Tendency of individuals to observe others, to find out what other people are thinking, feeling, and how they are behaving (Kashdan et al., 2018).	Five-Dimensional Curiosity Scale (Kashdan et al., 2018)
	Technical affinity	
Affinity for technology (ATI)	Refers to the extent to which people want to explore technical systems (Franke et al., 2019).	ATI Scale (Franke et al., 2019)

They were chosen from established psychological scales (e.g., Big-Five) based on their potential effects, derived from their definitions and findings from other studies as below. From traits that were not chosen from the scales, we did not find evidence in the literature for potential effects within our domain:

Personality traits:

- NFC (Bauer & Stiner, 2020; Cacioppo et al., 1984) has been shown to affect users' attention to explanations (Conati et al., 2021).
- OP (Gosling et al., 2003; McCrae & Sutin, 2009) has a recorded impact on users' intention to use a technical system (Millecamp et al., 2020).
- As CS (Barrick & Mount, 2012; Gosling et al., 2003; Roberts et al., 2014) impacts how people attend to their given tasks, it might influence interaction with the system.
- LOC (Rotter, 1966) is recognized to have an impact on performance measures in visualization tasks (Ottley et al., 2012; Ziemkiewicz et al., 2011, 2013).

Cognitive abilities:

- Both PS (Ekstrom & Harman, 1976; Gnambs et al., 2021) and VWM (Mance & Vogel, 2013; Vogel et al., 2001) have been found to impact time spent processing information (Carenini et al., 2014).
- Users' interaction with the system might be influenced by their RP, as it impacts how well they understand presented arguments (Meara, 1992).

Curiosity traits:

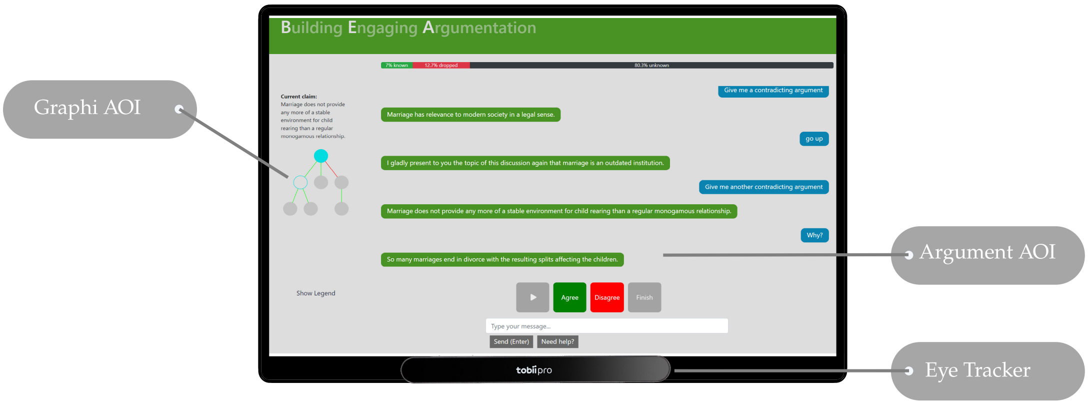
- DS (Kashdan et al., 2018) and SC (Kashdan et al., 2018) are both subdimensions of the construct curiosity. As curiosity fosters exploration and information seeking (Jirout, 2020), they could affect users' argumentation-seeking behavior.

Technical affinity:

- ATI (Franke et al., 2019) refers to how one approaches and deals with technology and might therefore impact how users interact with BEA.

7.3.3 Eye Tracking and Measures

We used the Tobii Pro Fusion eye tracker attached to the bottom of the monitor with a sampling rate of 120Hz. We utilized the Tobii Pro Lab software for eye gaze calibration and analysis. Two Areas of interest (AOIs) were defined (see Fig. 7.8): 1) The *Argument* AOI displaying the latest argument and 2) the *Graph* AOI showing the argument structure.



Current claim:
Marriage does not provide any more of a stable environment for child rearing than a regular monogamous relationship.

Give me a contradicting argument

Marriage has relevance to modern society in a legal sense.

go up

I gladly present to you the topic of this discussion again that marriage is an outdated institution.

Give me another contradicting argument

Marriage does not provide any more of a stable environment for child rearing than a regular monogamous relationship.

Why?

So many marriages end in divorce with the resulting splits affecting the children.

Type your message...

Figure 7.8: Setup: AOIs *Argument* and *Graph* and Eye Tracker attached to the bottom of the monitor to track the user's eye gaze movements.

We analyzed two fixation measures to track participants' visual focus and attention, similarly to Beattie et al. (2017):

- Average Number of Fixations (ANF): Signifying how frequently individuals process information and the importance they attribute to it (Cullipher et al., 2018).
- Average Duration of Fixations (ADF): Signifying how how long users pay attention to AOIs (Negi & Mitra, 2020).

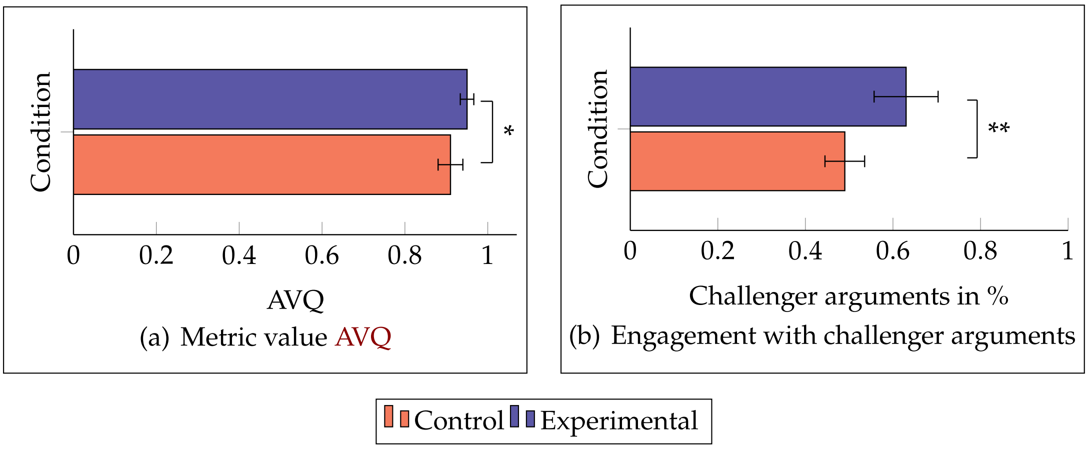
7.3.4 Statistical Analysis and Discussion

In the following, we report and discuss the statistical analysis results in detail. We selected the appropriate statistical model for each analysis after testing against relevant assumptions. We employed Shapiro-Wilk to test for normality. A parametric T-test was used if the data adhered to the normality assumption; otherwise, we opted for the non-parametric Mann-Whitney-U-Test. For the interaction effect analysis, we tested the data on the homogeneity of variances using Levene's Test and checked for extreme outliers; if those assumptions were met, we proceeded with an Analysis of Covariance (ANCOVA) test. If the assumptions of normality could not be met, we chose a General Linear Model (GLM) for the analysis. Furthermore, in each model, we analyzed each UC individually as a co-variate to prevent overfitting (Babyak, 2004). We excluded one participant from the analysis due to technical errors during the study.

7.3.4.1 Objective Dependent Measure: AVQ

We found a moderate [25] main effect of *condition* on AVQ with a significant increase (*Mann-Whitney-U-Test*, $U = 331, p = .018, r = .317$, **Fig. 7.9(a)**). We further found a significant relation between *condition* and *engagement with challenger arguments* ($\chi^2(1, N = 44) = 6.3043, p = .012$, **Tab. 7.11**).

Analyzing the main effect of intervention on the percentage of *challenger arguments* revealed a large significant main effect (*T-Test*, $t(42) = 3.1075, p = .003, d = .95$, **Fig. 7.9(b)**) confirming the results of our study discussed in Sec. 7.2.



(a) Metric value AVQ

(b) Engagement with challenger arguments

Figure 7.9: Means including 95% confidence interval denoted by bars of AVQ (**Fig. a**) and engagement with challenger arguments (**Fig. b**). (*) $p < .05$, (**) $p < .01$. See Appx. B.3.3.1 for data.

Table 7.11: Contingency table of engagement with challenger arguments per condition. Values in brackets show the number of arguments.

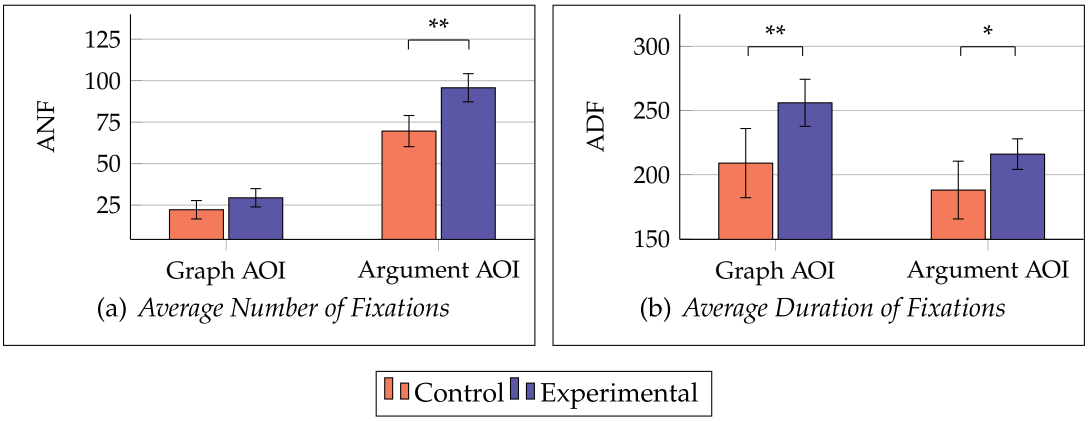
	Engagement with ···		
Condition	Challenger arg.	Non-challenger arg.	Total
Experimental	*20 (302)*	*3 (195)*	*23 (497)*
Control	*11 (248)*	*10 (251)*	*21 (499)*
Total	*31 (550)*	*13 (446)*	*44 (996)*

[25] Following J. Cohen (2013), Field (2013), and Tomczak and Tomczak (2014), we consider the effect size as **small** for $d, r < .3, \eta_p^2 \leq .06$, as **moderate** for $d, r < .5, \eta_p^2 \leq .14$ and as **large** for $d, r \geq .5, \eta_p^2 = \geq .14$.

7.3.4.2 Eye Gaze Analysis (Q2.6)

In the following, we will first investigate the effect of *intervention* on user attention and then analyze the impact of UCs. We excluded data from four participants due to technical issues leading to an absence of eye-tracking data.

Analyzing the eye-gaze behavior, we found the intervention significantly affects the user's attention to the graph and argument on the average number and duration of fixations. This shows that the intervention not only increased the calculated metric AVQ but also positively impacted the behavior of users in the experimental condition, as it shifted their attention to critical AOIs and prolonged their fixation on it.



(a) *Average Number of Fixations*

(b) *Average Duration of Fixations*

Figure 7.10: Means including 95% confidence interval denoted by bars of *Average Number of Fixations* (ANF, **Fig. a**) per minute and *Average Duration of Fixations* (ADF, **Fig b**) in milliseconds for *Graph* and *Argument* AOI. (*) $p < .05$, (**) $p < .01$. See Appx. B.3.3.2 for data.

Graph AOI. We observed a small trend of intervention on participants' ANF within the *Graph* AOI (*Mann-Whitney-U-Test*, $U = 140, p = .059, r = .25$, **Fig. 7.10(a)**) demonstrating an increased visual focus as participants directed their attention toward it more frequently (Culliphser et al., 2018).

There is further a moderate main effect of the intervention on the average duration of fixations (ADF) on the *Graph* AOI (Mann-Whitney-U-Test, $U = 295, p = .004, r = .417$, **Fig. 7.10(b)**), with the experimental group showing longer fixations than the control.

Argument AOI. There is a moderately significant main effect of the intervention on ANF (*Mann-Whitney-U-Test*, $U = 94, p = .002, r = .45$, **Fig. 7.10(a)**), whereby users in the experimental condition show a higher number of ANF compared to users in the control condition.

There is also a moderately significant main effect size of intervention on ADF (*T-Test*, $t(26.21) = 2.149, p = .021, d = .718,$ **Fig. 7.10(b)**), where users in the experimental condition have higher ADF compared to users in the control condition.

These results collectively demonstrate that the interventions increase users' attention to relevant parts of the arguments (graph and last argument presented), which, together with the increased metric AVQ, suggest improved processing overall.

7.3.4.3 Impact of UCs on attention (Q2.7)

The interaction effects reveal which users benefit from an intervention, as it directs their visual focus, in terms of ANF and ADF, towards the task's crucial AOIs, positively affecting their interaction with the system. Interaction effects were observed for the **personality traits** NFC and LOC, and **cognitive abilities** PS & VWM. We divided the UC data into binary classes using a median split for analysis.

1) NFC. We observed significant interaction effects and trends for NFC on ANF and ADF within the *Graph* and *Argument* AOI.

For the *Graph* AOI, we found a significant large interaction effect of NFC on ANF (*GLM*, $T(1,36) = -2.553, p < .015, \eta_p^2 == .15,$ **Fig. 7.11(a)**), as well as a trend for a moderate interaction effect of NFC on ADF (*GLM*, $T(1,36) = -2.016, p = .051, \eta_p^2 == .10,$ **Fig. 7.11(b)**). In the control condition, high NFC users have lower attention than their low-level counterparts, which is found to be moderately significant on ANF (*Mann-Whitney-U-Test*, $U = 17, p = .020, r = .489,$ **Fig. 7.11(c)**). Interventions primarily impact attention for high NFC users showing a large significant increase of ANF (*Mann-Whitney-U-Test*, $U = 11, p = .008, r = .583,$ **Fig. 7.11(c)**) and a moderate significant increase of ADF (*Mann-Whitney-U-Test*, $U = 17, p = .037, r = .444,$ **Fig. 7.11(d)**), whereas interventions have no impact on ANF for low NFC but a moderate significant main effect on ADF (*Mann-Whitney-U-Test*, $U = 35, p = .042, r = .368,$ **Fig. 7.11(d)**).

Parallel findings were observed for the *Argument* AOI. Here, we noted a trend for a moderate interaction effect between the intervention and NFC on ANF (*GLM*, $T(1,36) = -2.002, p = .053, \eta_p^2 == .10,$ **Fig. 7.12(a)**). Additionally, we found a moderate trend of NFC on ADF (*GLM*, $T(1,36) = 1.914, p < .064, \eta_p^2 = .0.09,$ **Fig. 7.12(b)**). Similarly to the *Graph* AOI, we also found a large significant main

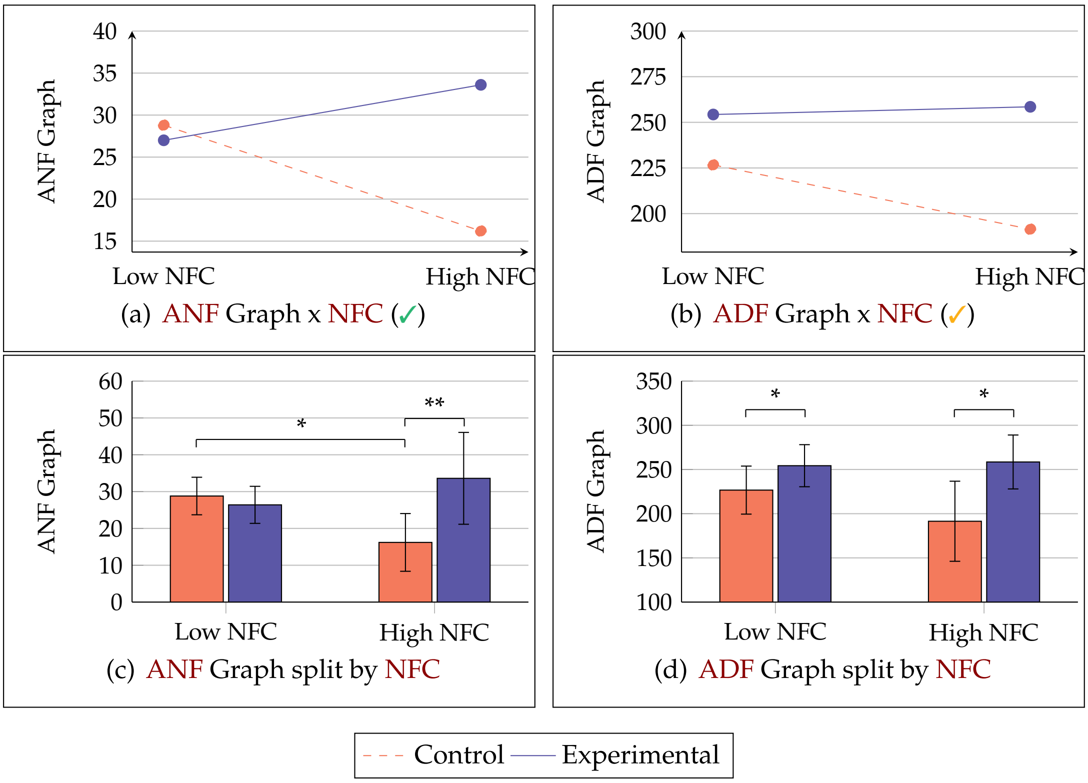
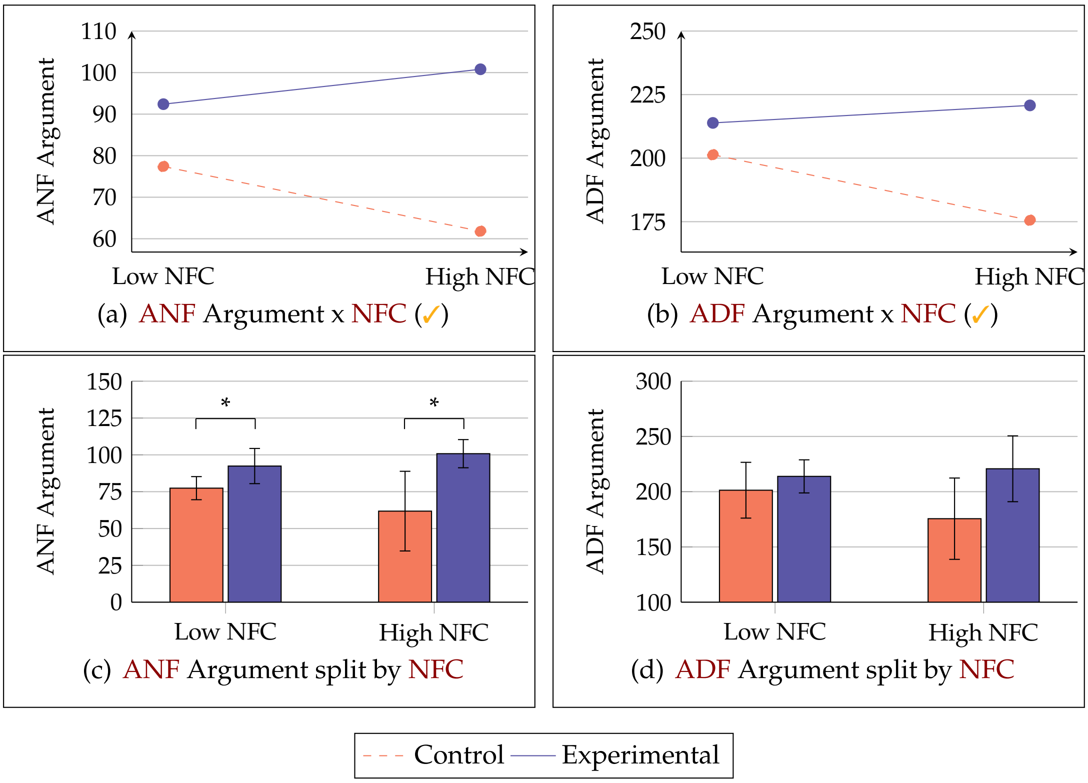


Figure 7.11: Graph AOI: Interaction effects of NFC on ANF/ADF (**Fig. a** and **b**) and main effects of condition × NFC on ANF/ADF (**Fig. c** and **d**). (✓) denotes a moderate trend, (✓) denotes a moderately significant effect. (*) p < .05, (**) p < .01.

effect of intervention on ANF (*Mann-Whitney-U-Test*, $U = 14, p = .018, r = .513$, **Fig. 7.12(c)**) as well as a moderate trend on ADF (*Mann-Whitney-U-Test*, $U = 25.5, p = .057, r = .397$, **Fig. 7.12(d)**) for high NFC users. For low NFC users, there was no main effect of the intervention on ADF, but a moderately significant effect on ANF (*Mann-Whitney-U-Test*, $U = 36, p = .048, r = .355$, **Fig. 7.12(c)**).

The results in the control condition seem contrary to the anticipated behavior of high NFC individuals, who are generally characterized by engaging in cognitively challenging tasks and reflection (Bauer & Stiner, 2020) and investing more effort into processing information and paying more attention to the ongoing task, notably when dealing with arguments (Q. Liu & Nesbit, 2024). Therefore, high NFC users are expected to show heightened attention and processing for important AOIs compared to low NFC users. However, Garner (2003) and Coppens et al. (2019) also report low performance of high NFC users in their respective research, suggesting that the association between high NFC and performance-related measures may be more prominent in more cognitive challenging tasks (Q. Liu & Nesbit, 2024). To



(a) ANF Argument x NFC (✓)

(b) ADF Argument x NFC (✓)

(c) ANF Argument split by NFC

(d) ADF Argument split by NFC

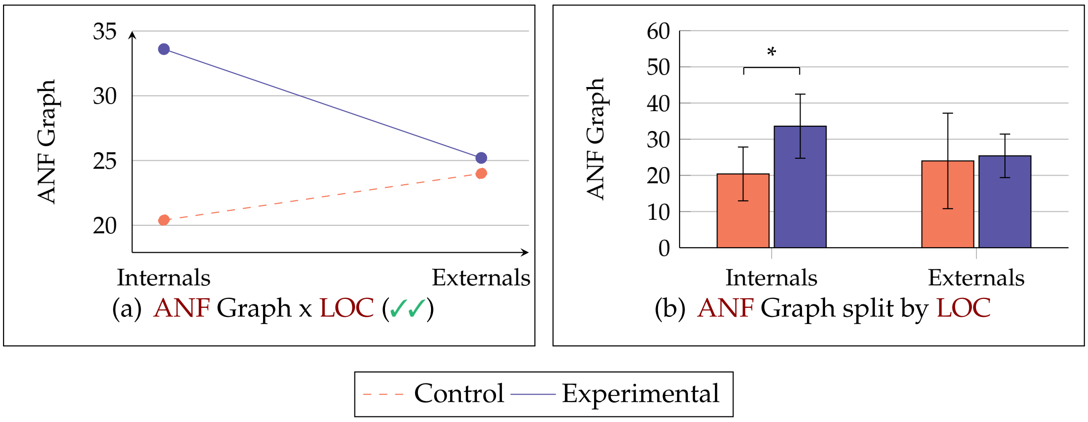
Figure 7.12: Argument AOI: Interaction effects of NFC on ANF/ADF (**Fig. a** and **b**) and main effects of condition × NFC on ANF/ADF (**Fig. c** and **d**). (✓) denotes a moderate trend. (*) p < .05, (**) p < .01.

fully explain the lower fixation measures of high NFC, further research is needed,

Nevertheless, the intervention in the experimental condition, which suggests the user taking a look at challenger arguments, proved successful in re-directing their attention to the argument and graph area and increasing their average duration on each AOI, potentially enhancing argument comprehension (Culliper et al., 2018), which demonstrates that the attention of high NFC users can be positively affected and be re-directed to critical aspects of the system.

2) LOC. We found a large significant interaction effect between LOC and condition on ANF for the *Graph* AOI (*GLM*, $T(1,36) = 3.159, p < .003, \eta_p^2 == .22$, **Fig. 7.13(a)**).

Individuals with an internal LOC (= Internals) feel they have substantial control over their life, while people with an external LOC (=Externals) attribute life events to factors beyond their control (Steca, 2021). This belief affects academic achievement, favoring Internals (Findley & Cooper, 1983). Interestingly, in



(a) ANF Graph x LOC (✓✓)

(b) ANF Graph split by LOC

Figure 7.13: Interaction effects between ANF and LOC. (✓✓) denotes a large significant interaction effect. (*) $p < .05$.

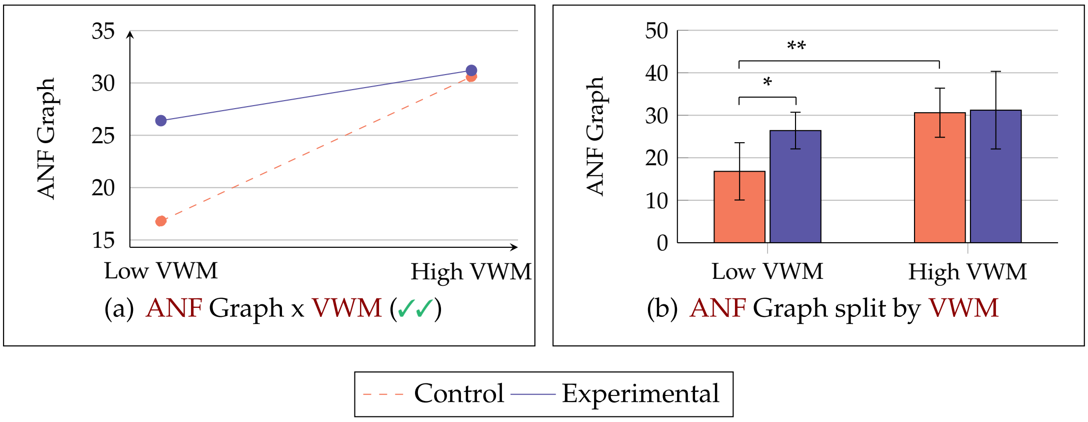
visualization studies, Externals tend to perform better in terms of speed (Ottley et al., 2012) and accuracy (Ziemkiewicz et al., 2013). Ziemkiewicz et al. (2013) explain this using an idea from distributed cognition: *Externalization*. In Externalization, slower cognitive processes are substituted with faster perceptual processes to enhance the efficiency of cognitive tasks such as problem-solving (Z. Liu et al., 2008). In contrast to Internals, Externals may depend more on external information, allowing them to adapt more quickly and make sense of new visualizations faster (Ziemkiewicz et al., 2013). In our findings, in the control condition, Internals have a reduced ANF for the *Graph* AOI, in contrast to Externals, suggesting that they rely less on external information and perceive the graph as less crucial.

The intervention significantly increases the ANF for internal LOC users (*Mann-Whitney-U-Test*, $U = 22, p = .019, r = .467$, **Fig. 7.13(b)**) with a moderate effect, heightening their attention. In the case of the externals, the intervention did not lead to an increase or decrease in ANF; therefore, another form of intervention might be necessary here for Externals.

3) Visual abilities PS and VWM. The user characteristics *perceptual speed* (PS) and *visual working memory* (VWM) are reported and discussed together as they both can be categorized as visual abilities.

For VWM, we found a large significant interaction effect on ANF in *Graph* AOI (*GLM*, $T(1, 36) = 2.544, p < .015, \eta_p^2 == .15$, **Fig. 7.14(a)**).

Further, a moderate significant interaction effect of PS on ANF (*GLM*, $T(1, 36) = 2.281, p = .029, \eta_p^2 == .13$, **Fig. 7.15(a)**), and a moderate trend of PS on ADF (*GLM*, $T(1, 36) = -1.985, p = .055, \eta_p^2 = .10$, **Fig. 7.15(b)**) was found for *Argument* AOI.



(a) ANF Graph x VWM (✓✓)

(b) ANF Graph split by VWM

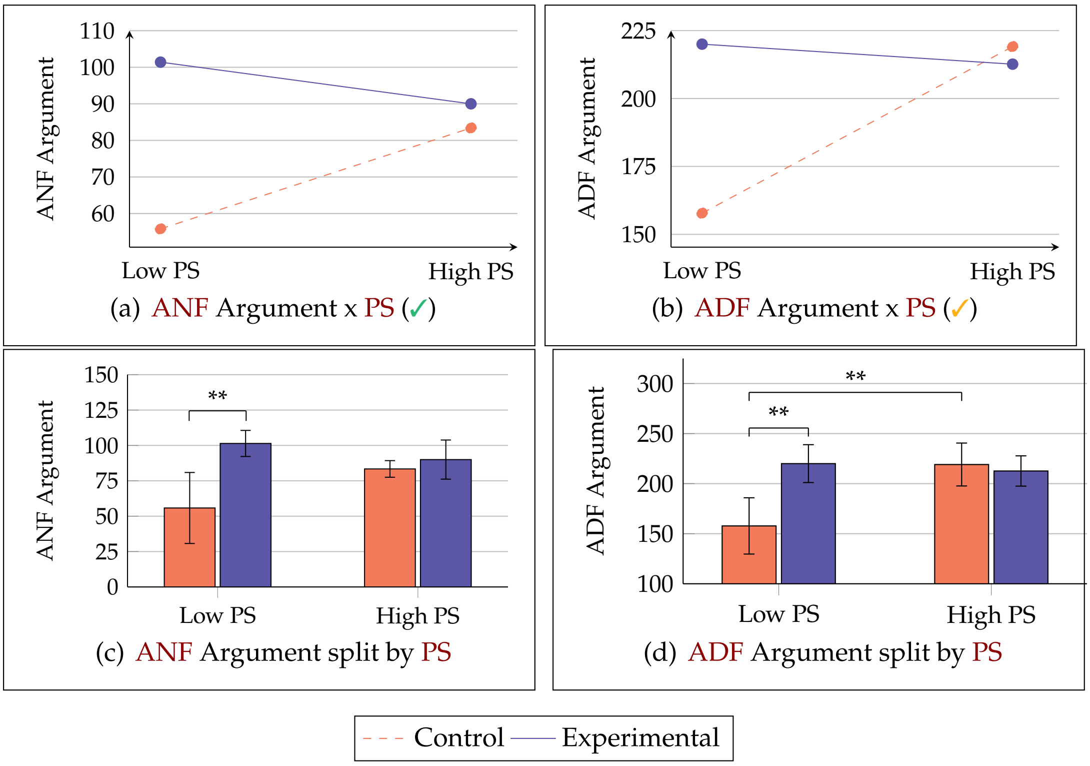
Figure 7.14: Interaction effects between VWM and ANF. (✓✓) denotes a large significant interaction effect. (*) $p < .05$, (**) $p < .01$.

These interaction effects indicate that users of the control condition with low PS/VWM levels pay much less attention to the relevant AOIs than their high-level counterparts, which is large significant for *Graph* AOI for VWM on ANF (*Mann-Whitney-U-Test*, $U = 65, p = .008, r = .566$, **Fig. 7.14(b)**) and large significant for *Argument* AOI for PS on ADF (*Mann-Whitney-U-Test*, $U = 36.5, p = .003, r = .666$, **Fig. 7.15(d)**).

This contrasts the findings of Toker and Conati (2014), who reported that individuals with lower PS and VWM levels allocated more time to vital AOIs and demonstrated greater transitions between information sources. In Toker and Conati (2014), participants were given specific tasks, such as comparing and performing aggregations. In our study, participants had a broader task assignment, which allowed them to choose what to focus on, resulting in reduced attention and cognitive processing of the *Graph* and *Argument* AOI. This suggests that when presented with a broader task description, individuals with lower visual abilities may not focus on crucial AOIs, with the underlying mechanisms needing to be clarified in future work.

The interventions significantly increase this attention, bringing it to the same level as the high-level users, which is large significant for *Graph* AOI for VWM on ANF (*Mann-Whitney-U-Test*, $U = 20, p = .013, r = .501$, **Fig. 7.14(b)**) and large significant for *Argument* AOI for PS on ANF (*Mann-Whitney-U-Test*, $U = 12, p = .002, r = .637$, **Fig. 7.15(c)**) and on ADF (*Mann-Whitney-U-Test*, $U = 8.5, p = .001, r = .697$, **Fig. 7.15(d)**).

For high visual abilities users, the intervention had no impact, as they had already directed their attention to the critical information.



(a) ANF Argument x PS (✓)

(b) ADF Argument x PS (✓)

(c) ANF Argument split by PS

(d) ADF Argument split by PS

Figure 7.15: Interaction effects between PS and ADF/ANF. (✓) denotes a moderate trend, (✓) denotes a moderate significant interaction effect. (**) $p < .01$.

7.4 Key Points & Summary

Within this chapter...

- ...we evaluated the intervention strategies, system design, and effects on exploration behavior.
- ...we demonstrated a main effect of *intervention* on user's attention to *challenger arguments*.
- ...we demonstrated the impact of the co-variate *embodiment* (chat vs. embodied) on system perception, trust, and user engagement.
- .... we demonstrated a main effect of *intervention* on *eye-gaze* behavior along with other interaction effects between *intervention* and User Characteristics.

This chapter addressed research questions **Q2.2-Q2.7**:

Q2.2 *"Does the intervention mechanism impact the user's engagement with challenger arguments, i.e., leads to an increase of AVQ?"*

Q2.3 *"Do the gamification mechanism and agent embodiment affect intervention success positively?"*

Q2.4 *"Does agent embodiment affect system perception, trust, and the user's CE positively?"*

Q2.5 *"Do interventions affect user trust negatively?"*

Q2.6 *"Do interventions impact users' eye gaze behavior (attention to arguments)?"*

Q2.7 *"Is there an interaction effect of User Characteristics (UCs) on the exploration behavior, and agent interactions?"*

In this chapter, we presented the results of three user studies to evaluate different aspects of our intervention strategies (see Tab. 7.12 for the summary).

In the first study, we investigated how the embodied agent is perceived compared to the chat-based agent with 84 participants. Additionally, we examined the effect of the intervention mechanism, the gamification strategy, and the embodied agent on AVQ in a preliminary setup with fewer participants.

In the second study, we investigated the main effect of the intervention mechanism on the user's AVQ and engagement with challenger arguments with 60 participants. The results showed a significant main effect of the intervention on the user's engagement with challenger arguments, indicating that our interventions effectively encouraged users to consider a less biased set of arguments.

The third study incorporated eye-tracking measures to evaluate user attention more deeply. We demonstrated that users who were nudged by the interventions not only focused more frequently on the presented arguments (ANF) but also maintained their focus for a longer duration (ADF) compared to those in the control condition. This study highlighted the effectiveness of our interventions in promoting a more critical and sustained exploration of arguments.

We also examined interaction effects with various User Characteristics (UCs), specifically NFC (Need for Cognition), PS (Perceptual Speed), VWM (Visual Working Memory), and LOC (Locus of Control). Our findings indicated that users with low levels of PS and VWM, and those with high NFC, derived substantial benefits from the interventions compared to their counterparts. Notably, in terms of ANF, users with both low and high NFC showed a significantly heightened level of attention to the arguments compared to the control condition.

Table 7.12: Key results of investigated main (+), no main (−), and interaction (∗) effects derived from the research questions. The metric score AVQ, propensity to trust (PT), trust in automation (TA), system behavior (SB), dialogue (DI), argumentation (ARG), and overall quality (QLT), average number of fixations (ANF), average duration of fixations (ADF), need for cognition (NFC), perceptual speed (PS), visual working memory (VWM), locus of control LOC. (✗) denotes no effect, (✓) denotes a partial effect, (✓) denotes a significant effect.

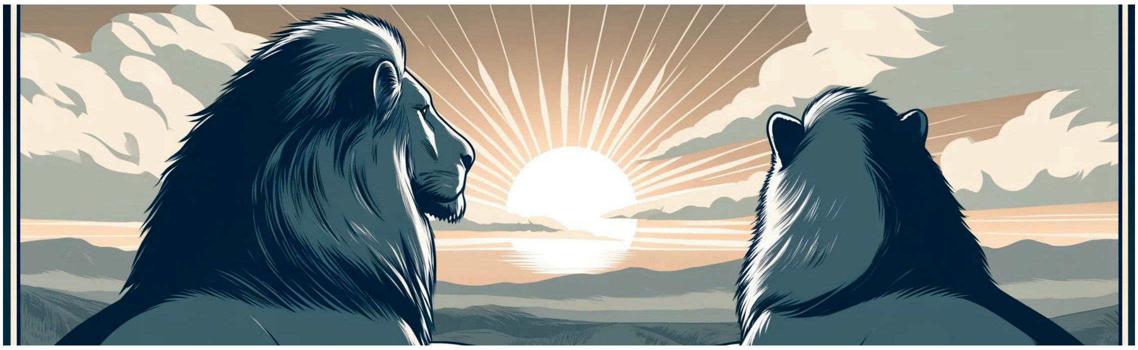
RQ	Hyp.	Effects Type	Items	Conf.?	Ch.
Q2.2	H1	+ *Intervention on AVQ*	*AVQ*	✓	7.2
	H4	+ *Intervention on argument visitation*		✓	
Q2.3	H2	∗ *Embodiment × interv. on AVQ*		✗	7.1.2
		∗ *Gamification × interv. on AVQ*		✗	
Q2.4	H3	+ *Embodiment on system perception*	*SB/DI/ARG*	✓	7.1.3
			QLT	✓	
		+ *Embodiment on trust*	*PT/TA*	✓	
		+ *Embodiment on CE*	*worthwhile*	✓	
Q2.5	H5	− *Intervention on user trust*		✓	7.2
Q2.6	E	+ *Intervention on eye gaze*	*Arg. ANF*	✓	7.3
			ADF	✓	
Q2.7		∗ *UCs × intervention on AVQ*		✗	
		∗ *UCs × intervention on* eye gaze	*NFC*	✓	
			PS/VWM	✓	
			LOC	✓	
			other	✗	

7.5 Relevant Publications

- Weber, K., Aicher, A., Minker, W., Ultes, S., & André, E. (2023a). Fostering User Engagement in the Critical Reflection of Arguments. *Proceedings of the 13th International Workshop on Spoken Dialogue Systems (IWSDS)*, 1–16
- Weber, K., Hogh, N., Conati, C., & André, E. (2024). A Gaze into Argumentative Chatbots: Exploring the Influence of Challenger Arguments on Reflection and Attention. *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents (IVA)*, 1–10
- Aicher, A., Weber, K., Minker, W., André, E., & Ultes, S. (2023). The Influence of Avatar Interfaces on Argumentative Dialogues. *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents (IVA)*, 1–8
- Aicher, A., Weber, K., André, E., Minker, W., & Stefan, U. (2024). BEA: Building Engaging Argumentation. *Proceedings of the 1st International Conference on Robust Argumentation Machines (RATIO)*, 1–17

Chapter 8

Conclusions and Future Work



"The important thing is not to stop questioning. Curiosity has its reason for existence." (Albert Einstein)

The ability to reflect and engage in reflective practices has become increasingly important. Considering the limited coverage of (political) events by traditional news media (*Corona* for instance) and the widespread dissemination of information through social media platforms, it is even more critical to understand reflection and to equip ourselves with the necessary tools to reflect appropriately, particularly when it comes to controversial topics. This thesis focused on methods and methodologies to become aware of behavior-based reflection bias and overcome the content-based reflection bias through argumentative dialogue systems.

8.1 Contributions

In the following, we summarize the three contributions along with the respective research questions defined in Sec. 1.3:

1. Methodological (Ch. 3)
2. Technical (Ch. 4, 5, and 6)
3. Empirical (Ch. 7)

8.1.1 Methodological Contribution: Raising Awareness

With the application of explainable artificial intelligence (XAI) techniques to investigate the impact of subliminal cues of individual users (see Sec. 3.2, 3.3), we proposed a novel methodology (see Sec. 3.1). By employing explainable AI, our research aimed to shed light on the behavior-based reflection bias of visual stimuli and raise awareness of them by comparing their impact on individual users to investigate *why* a speaker is perceived as persuasive. This approach goes beyond traditional studies that merely vary stimuli to investigate *what stimuli* have an effect. Varying stimuli mainly contributes to an understanding of visible, allegedly supraliminal cues (e.g., gesture vs. no gesture), but does not help raise awareness of the subliminal influence of these cues.

> Q1.1 *"Can we effectively uncover behavior-based reflection bias in political speeches and provide satisfactory explanations?"*

To uncover behavior-based reflection bias, we proposed a methodology involving training neural networks based on visual input from annotated video datasets. In comparison to state-of-the-art approaches varying stimuli to compare their persuasiveness, our methodology explores what users (annotators) focus on and explains *why* an image was rated as *persuasive* or *not persuasive* using XAI techniques by generating explanations highlighting specific regions of the input frame contributing to network predictions.

Despite the challenge of highly subjective data, we demonstrated the feasibility of the methodology by training a neural network on aggregated subjective annotations to predict perceived persuasiveness (see Sec.3.2) based on the visual input frames only. The network training result showed high accuracy (see Tab. 3.1), supporting the approach's feasibility.

We analyzed the ability to provide satisfactory explanations regarding what part of the input frame the network learned to focus on by testing several XAI methods at a *micro* and *macro* level, revealing the network's focus on contours and posture and gesture in particular. In the hypothetical scenario where no behavior-based bias of social cues existed, we expect a network not to be able to learn anything from such videos, which means a random focus and prediction contradicting a subliminal bias, thus demonstrating the overall feasibility of the methodology.

Q1.2 *"Can XAI contribute to highlighting and understanding subjective differences in persuasive cues in political speeches?"*

To answer this research question, we opted for a multi-network approach, training separate neural networks on each annotator's data (see Sec. 3.3), applied XAI once again, and investigated the differences between the annotators (see Sec. 3.3).

Our analysis revealed that the networks prioritize hands and hand gestures as the most significant indicators of perceived persuasiveness, aligning with existing literature (Maricchiolo et al., 2009; Newman et al., 2016; J. Peters & Hoetjes, 2017). This underscores the meaningful role of hand movements in shaping perceived persuasiveness and validates our approach because if these gestures held no genuine influence, the network's focus would have been random, without any discernible pattern.

For rater three, we identified a stronger focus on body movement and gestures, specifically highlighted by the network's distinct focus on the hands, even for the neutral class, contrary to the other raters, where we identified a stronger focus on the eye gaze direction for class neutral.

This comparison demonstrates that XAI contributes to highlighting and understanding subjective differences in persuasive cues within political speeches.

Overall, our presented approach and the analysis represent another step toward a more profound comprehension and heightened awareness of the subtle impact of non-verbal cues. In contrast to existing studies that merely vary stimuli to investigate *what stimuli* impact perceived persuasiveness, our explanation-based method provides a more intuitive way to comprehend the tangible influence of these subtle cues. It shows *why* speakers are perceived as persuasive by an NN. It therefore empowers individuals to witness these highlighted effects directly, making them easier to comprehend. This, in turn, fosters a greater awareness of subliminal persuasive cues and can reduce the tendency for peripheral processing of messages. As highlighted in J. Peters and Hoetjes (2017), individuals with low elaboration are susceptible to subliminal persuasion, in contrast to those with high elaboration. This susceptibility extends to the perceived speaker's performance and the speech's factual accuracy. Hence, when individuals become aware of subliminal processes and learn to mitigate peripheral processing, they can indirectly enhance their critical thinking skills.

8.1.2 Technical Contribution: Fostering Reflection

To mitigate the reflection bias in argument exploration, we developed an intelligent agent (Ch. 4) that monitors the user's exploration focus using a metric, namely Argument Visitation Quotient (AVQ, see Ch. 5), and guides the user's focus using interventions (Ch. 6).

> Q2.1 *"How can we formulate a computational metric for RE that is operationally feasible and can be programmatically implemented allowing the agent to guide the user's argument visitation focus?"*

To allow for a dynamic stance estimation during interaction and thus a non-static intervention behavior, we derived a user stance prediction model to determine the stance of users during their interaction with the system (see Sec. 5.2). The model utilizes direct user feedback as input to predict the current user stance during the argument exploration. Under the assumption that users tend to agree with a particular stance more when they agree with arguments supporting the *Major Claim*, it utilizes the user's feedback to calculate the stance on a fine-grained level with high accuracy. We incorporated this model into the computational metric AVQ (see Sec. 5.1) for RE. It aims to monitor the users' focus on arguments during their interaction with the agent in relation to their own stance.

We developed an argumentative intelligent agent (chat-based and embodied, see Ch. 4) and utilized the metric to enable the agent to monitor the user's argument exploration focus and foster a less biased argument exploration if users stick to arguments supporting their own opinion by intervening in case of a monitored low score (see Ch. 6), and proposing so-called *challenger arguments* that challenge the users' position. We equipped the system with adaptive and non-adaptive intervention strategies (see Ch. 6), allowing the agent to guide the user's exploration behavior and learn strategies to increase the intervention's success.

While most argumentative systems cope with enhancing speaking skills and learning the process of effective argumentation, our work is, to the best of our knowledge, the first that enabled an intelligent agent to foster a reflective argument exploration (see Sec. 2.3.2).

These technical contributions fill a gap in the existing literature and provide a tool for users to delve into controversial topics in a structured and reflective manner. The system's combination of argument organization, user interaction, and the metric AVQ for RE sets it apart from other argumentative dialogue systems, including *ChatGPT*, which do not foster critical argument exploration by default. They also tend to reinforce a user's confirmation bias (Sharma et al., 2024).

Integrating our system as a frontend with *ChatGPT* as the argument retrieval knowledge base could reduce the problem of confirmation bias in large language models like *ChatGPT* when used for information retrieval. Our system can detect when users only engage with supporting arguments and propose alternative perspectives. This setup ensures a more reflective dialogue, leveraging *ChatGPT's* extensive knowledge base while guiding users towards a reflective and less biased exploration of controversial topics. The flexibility and adaptability of our system's architecture, which uses argument trees consisting of support and attack arguments as a knowledge base, allows it to build and adapt argument structures at runtime dynamically. This inherent versatility allows the incorporation of various controversial topics. It enables the effective use of *ChatGPT* as a back-end, amplifying the system's impact and potential across different domains and topics.

8.1.3 Empirical Contributions

This thesis includes several empirical contributions to human-agent interaction. Across multiple studies conducted within this thesis, we have delved deeply into understanding the impact of the metrics, methodologies, and systems on user perception, motivation, and reflection.

The empirical contribution is thereby three-folded:

1. Argument Visitation Quotient (AVQ, see Sec. 7.1.2, Sec. 7.2 and Sec. 7.3)
2. Conversational Engagement (CE, see Sec. 7.1.3)
3. Attention to arguments (see Sec. 7.3)

We conducted three studies, with the first two conducted online (Sec. 7.1, Sec. 7.2), and the third one conducted at our lab (Sec 7.3) to answer the empirical research questions **Q2.2** - **Q2.7**.

Q2.2 *"Does the intervention mechanism impact the user's engagement with challenger arguments, i.e., leads to an increase of AVQ?"*

Q2.5 *"Do interventions affect user trust negatively?"*

Q2.6 *"Do interventions impact users' eye gaze behavior (attention to arguments)?"*

We showed a moderate significant main effect of the intervention on AVQ ($p \leq .01$, Sec. 7.2), i.e., the AVQ increases when the system uses interventions. We further demonstrated that significantly *more users* were more engaged with challenger arguments ($p = .032$) with a large significant main effect of intervention on *the amount of visited challenger arguments* ($p = .02, d = .55$), proving that the system is able to significantly shift the *user's focus* to *challenger arguments*.

In another study (Sec. 7.3), we confirmed these results and additionally investigated the profound impact of such simple intervention mechanisms on attention to arguments based on eye-tracking analysis. Users did not only focus more frequently on presented arguments (ANF) but also did so for a longer duration (ADF) compared to those who were not nudged to engage in a more critical exploration by interventions within the control condition.

Also, we found no negative impact of interventions on trust, which is a critical aspect of increasing the likeliness that people interact with the system more often. This is because when actions are perceived as controlling or coercive, it can negatively impact and reduce trust over time due to trust being a dynamic variable (Rhim et al., 2023).

Q2.3 *"Do the gamification mechanism and agent embodiment affect intervention success positively?"*

To enhance the effectiveness of the interventions, we explored the use of gamification strategies (*gamification* condition) in the first study. Research suggests that integrating game mechanics into non-gaming contexts can enhance user motivation and engagement by making interactions more enjoyable and rewarding (Deterding et al., 2011). Additionally, we investigated the interaction effect of *embodiment* on the intervention's success. However, we found no significant interaction effects between *gamification*, *embodiment*, and *intervention*.

In detail, we discussed the reasons for this in Sec. 7.1.4.2. In short, the lack of the users' understanding of the performance score, the absence of a clear benchmark for a good score, and the disruptive nature of the pop-up intervention likely contributed to the limited effectiveness of the gamification strategy.

Similarly, an embodied agent did not significantly increase compliance with the intervention compared to a chat-based agent despite enhancing the overall user experience (see below). This may be due to 1) the embodied agent's failure to convey authority or persuasiveness and 2) a potential disruption caused by the pop-up window again, which could have detracted from the perceived coherence of the intervention, the interaction with the agent, and the dialogue flow. The presented adaptive intervention strategy (Sec. 6.2.2) could help overcome this limitation by tailoring the interventions, giving the agent more personality.

Q2.4 *“Does agent embodiment affect system perception, trust, and the user’s CE positively?”*

Our experiments show a strong tendency towards the *embodied agent* over the *chat agent*.

Regarding *system perception*, we found a significant increase in perceived naturalness (`DI 3`, $p = .032$) and felt engagement (`ARG 5`, $p = .039$). Additionally, the impression of the *embodied agent* was rated significantly higher than that of the *chat agent* (`QLT 1`, $p = .047$).

With respect to trust, the *embodied agent* was rated as significantly more trustworthy (`TA 1`, $p = .039$).

Regarding Conversational Engagement (CE), we found the *embodied agent* to be perceived as more worthwhile (`RW 1`, $p = .022$).

In Sec. 4.3, we said that it was unclear whether avatars have an impact on the course of argumentative debates (Blount et al., 2015), which raised the question of whether the usage of an embodied agent as a counterpart in argumentative discussions is perceived as motivating and engaging.

Our findings further indicate that our proposed system enhances user acceptance when employing an *embodied agent*, mitigating the perception of technical difficulties compared to a standard *chat agent*.

This aligns with existing literature findings (Miao et al., 2022; Qiu et al., 2021), extending them to the context of argumentation.

Q2.7 *“Is there an interaction effect of User Characteristics (UCs) on the exploration behavior, and agent interactions?”*

We found interaction effects with several User Characteristics (UCs), that are NFC (Need for Cognition), PS (Perceptual Speed), VWM (Visual Working Memory), and LOC (Locus of Control). These interaction effects indicated that users with low levels of PS, VWM, and with high NFC derived substantial benefits from the interventions compared to their counterparts with lower levels of these UCs. Notably, concerning ANF, users with low and high NFC demonstrated a significantly heightened level of attention to the arguments compared to the control condition.

These findings overall amplify the impact of the proposed methods, metrics, and system and demonstrate that UCs significantly influence how individuals interact with and perceive the system and whether or not they benefit from the interventions. The difference in how interventions influence the behavior and interaction experience based on UCs indicates that they should be considered in future work when tailoring interventions based on users' personalities.

Overall, we can conclude that our herein presented approach, the intervention mechanisms, and the employed metric have a significant impact on argument exploration and attention to *challenger arguments*, thus fostering critical thinking.

8.2 Future Work and Outlook

In closing, we outline potential directions for further research and applications across diverse platforms where individuals engage with information.

First, regarding intervention strategies, we could explore whether integrating them directly into the dialogue flow, rather than using pop-up windows, enhances the intervention success, especially for the gamification strategy. Embedding the interventions within the conversation could increase the naturalness and reduce user's distraction, thereby increasing user engagement, enjoyment, and motivation to follow the interventions more often. In addition, investigating the effects of adaptive personalized intervention strategies tailored to individual users' personality traits, such as NFC, could enhance the success of these interventions, especially in cases where users ignore them.

Second, a lot of future work can be done regarding the metric AVQ. While our metric is intended to be a simple approximation of RE that measures the user's argument visitation focus, it does not account for several important psychological factors. For instance, incorporating *motivational factors* such as NFC or elaboration could be beneficial, as these factors play a crucial role in reflective thinking. Furthermore, integrating the user's *bias awareness* could enhance the metric as well, which concerns the user's awareness of their own biases and understanding how these biases influence their argument exploration behavior. Additionally, considering the extent to which users engage with and understand the provided arguments could further improve the metric.

Further, as outlined in the discussion, our application could incorporate large language models like ChatGPT as a back-end to facilitate the exploration of various controversial topics and investigate the intervention's effect on topics with different levels of divisiveness. Similarly, personal assistants and chatbots powered by large language models could provide more unbiased and reflective responses, improving user decision-making. Alternatively, our approach could be integrated into search engines like Google to not only retrieve information but also to present it in a manner that encourages critical thinking if users search for one-sided information too often. Future studies could evaluate the impact of these enhancements on daily information queries, personal decision-making, and their effect on people's "*self-imposed filter bubbles*" (Ekström et al., 2022).

As previously discussed, social media platforms often reinforce users' biases by filtering information based on past search queries and interactions. This issue could be addressed by extending our approach to social media platforms. Such platforms could detect when users predominantly engage with content that aligns with their existing opinions and expose them to challenging content (opposing their existing opinion) to promote a more critical perspective. In the current age, where social media is increasingly used for political election campaigns, such a mechanism could help users look beyond their self-imposed filter bubbles. Future research could investigate how such an approach impacts user engagement, the quality of online discussions, and whether it reduces polarization on social media.

Similarly, news websites could adopt our approach to present diverse viewpoints on controversial topics, helping readers develop a more critical and unbiased understanding of issues. Since some media sources are often biased themselves, our approach could be used to aggregate content from multiple sources to present a less biased perspective. Future research could explore the impact on reader satisfaction, trust in media, and the development of more informed opinions.

On a broader scale, our approach could be used to develop a healthcare decision-support system with which patients and doctors could obtain an unbiased and reflective overview of the pros and cons of several treatment options and their potential outcomes. This could increase user satisfaction and overall health outcomes.

This thesis lays the groundwork for fostering critical thinking within an argument exploration task. From improving information retrieval, personal assistants, and healthcare decision support to mitigating bias on social media and news platforms, our presented approach has the potential to significantly enhance how individuals engage with information and develop critical thinking skills. Integrating our framework with search engines can further broaden its impact by promoting balanced perspectives and mitigating confirmation bias. By addressing these future research directions, we can develop better applications for information engagement, consequently contributing to a more reflective and informed society.

Bibliography

Abro, W. A., Aicher, A., Rach, N., Ultes, S., Minker, W., & Qi, G. (2022). Natural Language Understanding for Argumentative Dialogue Systems in the Opinion Building Domain. *Knowledge-Based Systems*, *242*, 108318:1–21. https://doi.org/10.1016/j.knosys.2022.108318 (cit. on pp. 96, 99).

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052 (cit. on p. 31).

Aesop. (2020). The Boy Who Cried Wolf (The Shepherd's Boy and the Wolf). In J. Zipes (Ed.), *The Complete Fables of Aesop* (pp. 112–113). Penguin Classics. (Cit. on p. 59).

Aicher, A., Minker, W., & Ultes, S. (2021a). Determination of Reflective User Engagement in Argumentative Dialogue Systems. *Proceedings of the 2021 Workshop on Computational Models of Natural Argument (CMNA)*, 1–8. https://ceur-ws.org/Vol-2937/paper1.pdf (cit. on p. 109).

Aicher, A., Rach, N., Minker, W., & Ultes, S. (2021b). Opinion Building Based on the Argumentative Dialogue System BEA. *Proceedings of the 10th International Workshop on Spoken Dialogue Systems (IWSDS): Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, 307–318. https://doi.org/10.1007/978-981-15-9323-9_27 (cit. on pp. 64, 112, 125).

Aicher, A., Weber, K., André, E., Minker, W., & Stefan, U. (2024). BEA: Building Engaging Argumentation. *Proceedings of the 1st International Conference on Robust Argumentation Machines (RATIO)*, 1–17 (cit. on pp. 95, 106, 108, 127, 151, 159, 181, 259).

Aicher, A., Weber, K., Minker, W., André, E., & Ultes, S. (2023). The Influence of Avatar Interfaces on Argumentative Dialogues. *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents (IVA)*, 1–8. https://doi.org/10.1145/3570945.3607343 (cit. on pp. 95, 106, 151, 181, 259).

Alam, M. U., Baldvinsson, J. R., & Wang, Y. (2022). Exploring LRP and Grad-CAM Visualization to Interpret Multi-Label-Multi-Class Pathology Prediction

Using Chest Radiography. *Proceedings of the 35th IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, 258–263. https://doi.org/10.1109/CBMS55023.2022.00052 (cit. on pp. 31, 78).

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a Convolutional Neural Network. *Proceedings of the 2017 International Conference on Engineering and Technology (ICET)*, 1–6. https://doi.org/10.1109/ICEngTechnol.2017.8308186 (cit. on p. 22).

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., Samek, W., Müller, K. R., Dähne, S., & Kindermans, P. J. (2019). iNNvestigate Neural Networks. *Journal of Machine Learning Research*, *20*, 1–8. https://www.jmlr.org/papers/volume20/18-540/18-540.pdf (cit. on p. 77).

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, *31*(2), 211–36. https://doi.org/10.1257/jep.31.2.211 (cit. on pp. 8, 52).

Amgoud, L., & Ben-Naim, J. (2018). Weighted Bipolar Argumentation Graphs: Axioms and Semantics. *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 5194–5198. https://doi.org/10.24963/ijcai.2018/720 (cit. on pp. 112, 113).

André, E. (2011). Experimental Methodology in Emotion-Oriented Computing. *IEEE Pervasive Computing*, *10*(3), 54–57. https://doi.org/10.1109/MPRV.2011.50 (cit. on p. 57).

André, E., Bayer, S., Benke, I., Benlian, A., Cummins, N., Gimpel, H., Hinz, O., Kersting, K., Maedche, A., Muehlhaeuser, M., et al. (2019). Humane Anthropomorphic Agents: The Quest for the Outcome Measure. *Proceedings of the Pre-ICIS Workshop "Values and Ethics in the Digital Age"*, *12*(4), 1–16 (cit. on p. 258).

APA Dictionary of Psychology. (2023). APA Dictionary of Psychology's Definition of 'Need For Cognition'. Retrieved 2023-12-05, from https://dictionary.apa.org/need-for-cognition (cit. on p. 6).

Arapakis, I., Lalmas, M., & Valkanas, G. (2014). Understanding Within-Content Engagement Through Pattern Analysis of Mouse Gestures. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM)*, 1439–1448. https://doi.org/10.1145/2661829.2661909 (cit. on pp. 11, 66).

Argyle, M., & Cook, M. (1976). *Gaze and Mutual Gaze*. Cambridge University Press. (Cit. on p. 59).

Babyak, M. A. (2004). What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. *Psychosomatic Medicine*, *66*(3), 411–421. https://doi.org/10.1097/01.psy.0000127692.23278.a9 (cit. on p. 170).

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation (O. D. Suarez, Ed.). *PLoS ONE*, *10*(7), 1–46. https://doi.org/10.1371/journal.pone.0130140 (cit. on pp. 33, 35).

Barrick, M. R., & Mount, M. K. (2012). Select on Conscientiousness and Emotional Stability. *Handbook of Principles of Organizational Behavior: Indispensable Knowledge for Evidence-Based Management*, 19–39. https://doi.org/10.1002/9781119206422.ch2 (cit. on pp. 167, 169).

Bauer, B., & Stiner, E. (2020). Need for Cognition. *Encyclopedia of Personality and Individual Differences*, 3122–3125. https://doi.org/10.1007/978-3-319-24612-3_1093 (cit. on pp. 52, 167, 169, 174).

Baumer, E. P., Khovanskaya, V., Matthews, M., Reynolds, L., Schwanda Sosik, V., & Gay, G. (2014). Reviewing Reflection: On the Use of Reflection in Interactive System Design. *Proceedings of the 2014 Conference on Designing Interactive Systems (DIS)*, 93–102. https://doi.org/10.1145/2598510.2598598 (cit. on p. 61).

Baur, T., Heimerl, A., Lingenfelser, F., Wagner, J., Valstar, M. F., Schuller, B., & André, E. (2020). Explainable Cooperative Machine Learning with Nova. *Joint German/Austrian Conference on Artificial Intelligence - Künstliche Intelligenz (KI)*, 1–22. https://doi.org/10.1007/s13218-020-00632-3 (cit. on p. 70).

Beattie, G., Marselle, M., McGuire, L., & Litchfield, D. (2017). Staying Over-Optimistic About the Future: Uncovering Attentional Biases to Climate Change Messages. *Semiotica*, *2017*(218), 21–64. https://doi.org/10.1515/sem-2016-0074 (cit. on pp. 166, 170).

Bilmes, J. (2020). Underfitting and Overfitting in Machine Learning [Accessed 01-03-2024]. *UW ECE Course Notes*, *5*, 1–6. https://people.ece.uw.edu/bilmes/classes/ee511/ee511_spring_2020/overfitting_underfitting.pdf (cit. on p. 27).

Bjorck, N., Gomes, C. P., Selman, B., & Weinberger, K. Q. (2018). Understanding Batch Normalization. *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*, *31*, 1–12. https://proceedings.neurips.cc/paper_files/paper/2018/file/36072923bfc3cf47745d704feb489480-Paper.pdf (cit. on p. 72).

Black, E., & Hunter, A. (2009). An Inquiry Dialogue System. *Proceedings of the 2009 International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, *19*, 173–209. https://doi.org/10.1007/s10458-008-9074-5 (cit. on p. 64).

Blasco, P. G., Moreto, G., Blasco, M. G., Levites, M. R., & Janaudis, M. A. (2015). Education Through Movies: Improving Teaching Skills and Fostering Reflection Among Students and Teachers. *Journal for Learning through the Arts*, *11*(1), 1–17. https://doi.org/10.21977/D911122357 (cit. on p. 62).

Blount, T., Millard, D. E., & Weal, M. J. (2015). On the Role of Avatars in Argumentation. *Proceedings of the 2015 Workshop on Narrative & Hypertext (NHT)*, 17–19. https://doi.org/10.1145/2804565.2804569 (cit. on pp. 104, 189).

Brightside. (2023). 15 Incredible Tricks Advertisers Use to Make Food Look Delicious. Retrieved 2023-12-04, from https://brightside.me/wonder-curiosities/15-incredible-tricks-advertisers-use-to-make-food-look-delicious-296860/ (cit. on p. 2).

Brown, P., & Levinson, S. C. (2009). Politeness: Some Universals in Language Usage [Chapter 1, Reprint]. In *Sociolinguistics: Critical Concepts [volume III: Interactional Sociolinguistics]* (pp. 311–323). Routledge. (Cit. on pp. 134, 136).

Burgoon, J. K., Birk, T., & Pfau, M. (1990). Nonverbal Behaviors, Persuasion, and Credibility. *Human Communication Research*, *17*(1), 140–169. https://doi.org/10.1111/j.1468-2958.1990.tb00229.x (cit. on p. 53).

Busoniu, L., Babuska, R., Schutter, B. D., & Ernst, D. (2010). *Reinforcement Learning and Dynamic Programming Using Function Approximators* (1st). CRC Press, Inc., USA. (Cit. on pp. 44, 45).

Buss, D. M. (1987). Selection, Evocation, and Manipulation. *Journal of Personality and Social Psychology*, *53*(6), 1214–1221. https://doi.org/10.1037/0022-3514.53.6.1214 (cit. on pp. 3, 4).

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The Efficient Assessment of Need for Cognition. *Journal of Personality Assessment*, *48*(3), 306–307. https://doi.org/10.1207/s15327752jpa4803_13 (cit. on pp. 52, 167, 169).

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2019). OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *Proceedings of the 2021 IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *43*(1), 172–186. https://doi.org/10.1109/TPAMI.2019.2929257 (cit. on p. 81).

Carenini, G., Conati, C., Hoque, E., Steichen, B., Toker, D., & Enns, J. (2014). Highlighting Interventions and User Differences: Informing Adaptive Information Visualization Support. *Proceedings of the 2014 SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 1835–1844. https://doi.org/10.1145/2556288.2557141 (cit. on p. 169).

Castellano, G., Kessous, L., & Caridakis, G. (2008). Emotion Recognition through Multiple Modalities: Face, Body Gesture, Speech. In C. Peter & R. Beale (Eds.), *Affect and Emotion in Human-Computer Interaction: From Theory to Applications* (pp. 92–103). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-85099-1_8 (cit. on p. 58).

Chaiken, S. (1989). Heuristic and Systematic Information Processing Within and Beyond the Persuasion Context. *Unintended Thought*, 212–252 (cit. on p. 53).

Chalaguine, L. A., & Hunter, A. (2020). A Persuasive Chatbot Using a Crowd-Sourced Argument Graph and Concerns. *Proceedings of the 2020 International Conference on Computational Models of Argument (COMMA)*, *326*, 9–20. https://doi.org/10.3233/FAIA200487 (cit. on p. 64).

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953 (cit. on p. 84).

Chen, F. S., Minson, J. A., Schöne, M., & Heinrichs, M. (2013). In the Eye of the Beholder: Eye Contact Increases Resistance to Persuasion. *Psychological science*, *24*(11), 2254–2261. https://doi.org/10.1177/0956797613491968 (cit. on p. 59).

Cheng, C.-H., & Yang, F.-Y. (2022). Analyzing Visual Attention During Tap Learning and the Effect of Epistemic Beliefs on the Understanding of Argument Components. *International Journal of Science Education*, *44*(8), 1336–1355. https://doi.org/10.1080/09500693.2022.2076950 (cit. on p. 166).

Cicero, M. T. (2001). *On the Ideal Orator (de Oratore)* (J. May J.M. - Wisse, Trans.) [Translated by May, J.M. and Wisse, J.]. Oxford University Press. (Cit. on p. 57).

Cocarascu, O., & Toni, F. (2016). Detecting Deceptive Reviews Using Argumentation. *Proceedings of the 1st International Workshop on AI for Privacy and Security*, 1–8. https://doi.org/10.1145/2970030.2970031 (cit. on p. 117).

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Routledge. https://doi.org/10.4324/9780203771587 (cit. on pp. 150, 160, 171).

Cohen, P. R., & Perrault, C. R. (1979). Elements of a Plan-Based Theory of Speech Acts. *Cognitive Science*, *3*(3), 177–212. https://doi.org/10.1016/S0364-0213(79)80006-3 (cit. on p. 36).

Collins, A. (1986). A Sample Dialogue Based on a Theory of Inquiry Teaching. *Center for the Study of Reading Technical Report*, 367:1–38 (cit. on p. 63).

Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward Personalized XAI: A Case Study in Intelligent Tutoring Systems. *Artificial Intelligence*, *298*, 1–23. https://doi.org/10.1016/j.artint.2021.103503 (cit. on pp. 13, 165, 169).

Coppens, L. C., Hoogerheide, V., Snippe, E. M., Flunger, B., & van Gog, T. (2019). Effects of Problem–Example and Example–Problem Pairs on Gifted and Nongifted Primary School Students' Learning. *Instructional Science*, *47*(3), 279–297. https://doi.org/10.1007/s11251-019-09484-3 (cit. on p. 174).

Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555 (cit. on p. 28).

Cullipher, S., Hansen, S. J., & VandenPlas, J. R. (2018). Eye Tracking As a Research Tool: An Introduction. In *Eye Tracking for the Chemistry Education Researcher* (pp. 1–9). ACS Publications. https://doi.org/10.1021/bk-2018-1292.ch001 (cit. on pp. 166, 170, 172, 175).

D'Alfonso, A. (2016). The UK 'Rebate' on the EU Budget: An Explanation of the Abatement and Other Correction Mechanisms. Retrieved 2023-12-04, from https://www.europarl.europa.eu/RegData/etudes/BRIE/2016/577973/EPRS_BRI(2016)577973_EN.pdf (cit. on p. 3).

Danciu, V. (2014). Manipulative Marketing: Persuasion and Manipulation of the Consumer Through Advertising. *Theoretical and Applied Economics*, *2*(591), 19–34. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=5cbe76dbb0b4707d32c8439ce6d4c184a591251f (cit. on p. 2).

Dann, C. (2012). *Algorithms for Fast Gradient Temporal Difference Learning* (tech. rep.). TU Darmstadt. (Cit. on p. 49).

Darmawansah, D., Lin, C.-J., & Hwang, G.-J. (2022). Empowering the Collective Reflection-Based Argumentation Mapping Strategy to Enhance Students' Argumentative Speaking. *Computers & Education*, *184*, 104516. https://doi.org/10.1016/j.compedu.2022.104516 (cit. on p. 63).

Dean, D., & Kuhn, D. (2003). Metacognition and Critical Thinking. *Educational Resources Information Center*, 1–11. Retrieved 2023-09-08, from https://files.eric.ed.gov/fulltext/ED477930.pdf (cit. on p. 56).

de Melo, C. M., Carnevale, P., & Gratch, J. (2012). The Effect of Virtual Agents' Emotion Displays and Appraisals on People's Decision Making in Negotiation. *Proceedings of the 2012 International Conference on Intelligent Virtual Agents (IVA)*, 53–66. https://doi.org/10.1007/978-3-642-33197-8_6 (cit. on p. 58).

Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., & Cieliebak, M. (2021). Survey on Evaluation Methods for Dialogue Systems. *Artificial Intelligence Review*, *54*, 755–810. https://doi.org/10.1007/s10462-020-09866-x (cit. on p. 36).

DeSteno, D., Petty, R. E., Rucker, D. D., Wegener, D. T., & Braverman, J. (2004). Discrete Emotions and Persuasion: The Role of Emotion-Induced Expectancies. *Journal of Personality and Social Psychology*, *86*(1), 43–56. https://doi.org/10.1037/0022-3514.86.1.43 (cit. on p. 57).

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From Game Design Elements to Gamefulness: Defining Gamification. *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments (MindTrek)*, 9–15. https://doi.org/10.1145/2181037.2181040 (cit. on pp. 132, 188).

Dewey, J. (1910). *How We Think*. Heath & Co. https://yourknow.com/uploads/books/5e018bd9bd4cd.pdf (cit. on p. 65).

Dewey, J. (1933). *How We Think*. Buffalo, N Y: Prometheus Books. (Cit. on p. 65).

Dienes, Z. (2014). Using Bayes to Get the Most out of Non-Significant Results. *Frontiers in Psychology: Quantitative Psychology and Measurement*, *5*, 781:1–17. https://doi.org/10.3389/fpsyg.2014.00781 (cit. on p. 163).

Dienes, Z. (2021). How to Use and Report Bayesian Hypothesis Tests. *Psychology of Consciousness: Theory, Research, and Practice*, *8*(1), 1–39. https://doi.org/10.31234/osf.io/bua5n (cit. on pp. 162, 163).

Dijk, J. v., Roest, J. v. d., Lugt, R. v. d., & Overbeeke, K. C. (2011). NOOT: A Tool for Sharing Moments of Reflection During Creative Meetings. *Proceedings of the 8th ACM Conference on Creativity and Cognition (C&C)*, 157–164. https://doi.org/10.1145/2069618.2069646 (cit. on p. 61).

Dole, J. A., & Sinatra, G. M. (1998). Reconceptalizing Change in the Cognitive Construction of Knowledge. *Educational Psychologist*, *33*(2-3), 109–128. https://doi.org/10.1080/00461520.1998.9653294 (cit. on pp. 6, 55).

Donadello, I., Dragoni, M., & Eccher, C. (2019). Persuasive Explanation of Reasoning Inferences on Dietary Data. *Joint Proceedings of the 6th International Workshop on Dataset Profiling and Search & the 1st Workshop on Semantic Explainability co-located with the 18th International Semantic Web Conference (ISWC 2019)*, *2465*, 46–61. https://cris.fbk.eu/handle/11582/319876 (cit. on p. 29).

Dupret, G., & Lalmas, M. (2013). Absence Time and User Engagement: Evaluating Ranking Functions. *Proceedings of the 6th ACM International Conference on*

Web Search and Data Mining (WSDM), 173–182. https://doi.org/10.1145/2433396.2433418 (cit. on pp. 11, 66).

Ekman, P. (2004). Emotional and Conversational Nonverbal Signals. *Language, Knowledge, and Representation: Proceedings of the 6th International Colloquium on Cognitive Science (ICCS-99)*, 39–50. https://doi.org/10.1007/978-1-4020-2783-3_3 (cit. on p. 58).

Ekman, P., & Friesen, W. V. (1971). Constants Across Cultures in the Face and Emotion. *Journal of Personality and Social Psychology*, *17*(2), 124–129. https://doi.org/10.1037/h0030377 (cit. on p. 57).

Ekstrom, R. B., & Harman, H. H. (1976). Manual for Kit of Factor-Referenced Cognitive Tests. *Princeton, New Jersey: Educational Testing Service* (cit. on pp. 168, 169).

Ekström, A. G., Niehorster, D. C., & Olsson, E. J. (2022). Self-Imposed Filter Bubbles: Selective Attention and Exposure in Online Search. *Computers in Human Behavior Reports*, *7*, 100226, 1–10. https://doi.org/10.1016/j.chbr.2022.100226 (cit. on pp. 5, 65, 191).

Escalante, H. J., Guyon, I., Escalera, S., Jacques, J. C. S., Madadi, M., Baró, X., Ayache, S., Viegas, E., Güçlütürk, Y., Güçlü, U., van Gerven, M. A. J., & van Lier, R. (2017). Design of an Explainable Machine Learning Challenge for Video Interviews. *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN)*, 3688–3695. https://doi.org/10.1109/IJCNN.2017.7966320 (cit. on p. 29).

Farr, F., & Riordan, E. (2012). Students' Engagement in Reflective Tasks: An Investigation of Interactive and Non-Interactive Discourse Corpora. *Classroom Discourse*, *3*(2), 129–146. https://doi.org/10.1080/19463014.2012.716622 (cit. on pp. 7, 11, 56, 66).

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146 (cit. on pp. 150, 162).

Fazzinga, B., Galassi, A., & Torroni, P. (2021). An Argumentative Dialogue System for COVID-19 Vaccine Information. *Proceedings of the 2021 International Conference on Logic and Argumentation (CLAR)*, 477–485. https://doi.org/10.1007/978-3-030-89391-0_27 (cit. on p. 64).

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. Sage Publications. (Cit. on pp. 150, 160, 171).

Findley, M. J., & Cooper, H. M. (1983). Locus of Control and Academic Achievement: A Literature Review. *Journal of Personality and Social Psychology*, *44*(2), 419–427. https://doi.org/10.1037/0022-3514.44.2.419 (cit. on p. 175).

Fischer, K., Langedijk, R. M., Nissen, L. D., Ramirez, E. R., & Palinko, O. (2020). Gaze-Speech Coordination Influences the Persuasiveness of Human-Robot Dialog in the Wild. *Proceedings of the 2020 International Conference on Social Robotics (ICSR)*, 157–169. https://doi.org/10.1007/978-3-030-62056-1_14 (cit. on pp. 57, 60).

Fortmann-Roe, S. (2012). Understanding the Bias-Variance Tradeoff [Accessed 01-03-2024]. https://scott.fortmann-roe.com/docs/BiasVariance.html (cit. on p. 27).

Franke, T., Attig, C., & Wessel, D. (2019). A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ati) scale. *International Journal of Human–Computer Interaction*, *35*(6), 456–467. https://doi.org/10.1080/10447318.2018.1456150 (cit. on pp. 168, 169).

Garner, J. K. (2003). *The Persuasion Model of Conceptual Change and Its Application to Misconceptions in Evolution* [Doctoral dissertation]. The Pennsylvania State University ProQuest Dissertations & Theses. https://www.proquest.com/docview/287990048 (cit. on p. 174).

Gass, R. H., & Seiter, J. S. (2018). What Constitutes Persuasion. In *Persuasion: Social Influence and Compliance Gaining* (6th ed., pp. 30–46). Routledge. https://doi.org/10.4324/9781315209302 (cit. on p. 3).

Gelter, H. (2003). Why Is Reflective Thinking Uncommon. *Reflective Practice*, *4*(3), 337–344. https://doi.org/10.1080/1462394032000112237 (cit. on p. 111).

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, *4*(1), 1–58. https://doi.org/10.1162/neco.1992.4.1.1 (cit. on pp. 9, 27).

Gennari, R., Matera, M., Melonio, A., Rizvi, M., & Roumelioti, E. (2021). Reflection and Awareness in the Design Process: Children Ideating, Programming and Prototyping Smart Objects. *Multimedia Tools and Applications*, *80*(26-27), 34909–34932. https://doi.org/10.1007/s11042-020-09927-x (cit. on p. 61).

George, D., & Mallery, P. (2002). *SPSS for Windows Step by Step: A Simple Study Guide and Reference, 11.0 update, 10/e* (Vol. 4). Routledge. (Cit. on p. 29).

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning.

Proceedings of the IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 80–89. https://doi.org/10.1109/DSAA.2018.00018 (cit. on p. 30).

Glorot, X., & Bengio, Y. (2010). Understanding the Difficulty of Training Deep Feedforward Neural Networks. *Proceedings of the 5th International Conference on Data Science and Advanced Analytics (DSAA)*, *9*, 249–256. https://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf (cit. on p. 72).

Gnambs, T., Scharl, A., & Rohm, T. (2021). Comparing Perceptual Speed Between Educational Contexts: The Case of Students with Special Educational Needs. *Psychological Test Adaptation and Development*, 1–9. https://doi.org/10.1027/2698-1866/a000013 (cit. on pp. 168, 169).

Goda, Y., Yamada, M., Matsukawa, H., Hata, K., & Yasunami, S. (2014). Conversation with a Chatbot Before an Online EFL Group Discussion and the Effects on Critical Thinking. *The Journal of Information and Systems in Education*, *13*(1), 1–7. https://doi.org/10.12937/ejsise.13.1 (cit. on p. 64).

Goodfellow, I., Bengio, Y., & Courville, A. (2017). *Deep Learning* (Vol. 1). The MIT Press Cambridge, MA, USA. (Cit. on pp. 16, 17, 23–25, 71).

Gordon, G., Spaulding, S., Westlund, J. K., Lee, J. J., Plummer, L., Martinez, M., Das, M., & Breazeal, C. (2016). Affective Personalization of a Social Robot Tutor for Children's Second Language Skills. *Proceedings of the 30th Conference on Artificial Intelligence (AAAI)*, *30*(1), 1–7. https://doi.org/10.1609/aaai.v30i1.9914 (cit. on p. 44).

Gosling, S. D., Rentfrow, P. J., & Swann Jr, W. B. (2003). A Very Brief Measure of the Big-Five Personality Domains. *Journal of Research in Personality*, *37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1 (cit. on pp. 167, 169).

Govaerts, S., Verbert, K., Duval, E., & Pardo, A. (2012). The Student Activity Meter for Awareness and Self-reflection. *Extended Abstracts on Human Factors in Computing Systems (CHI EA)*, 869–884. https://doi.org/10.1145/2212776.2212860 (cit. on pp. 11, 62, 66).

Great Britain & Treasury. (2018). *European Union Finances 2017: Statement on the 2017 EU Budget and Measures to Counter Fraud and Financial Mismanagement*. APS Group on behalf of the Controller of Her Majesty's Stationery Office. (Cit. on p. 3).

Greydanus, S., Koul, A., Dodge, J., & Fern, A. (2018). Visualizing and Understanding Atari Agents. *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 1787–1796. https://proceedings.mlr.press/v80/greydanus18a.html (cit. on p. 35).

Grigoriadou, M., Tsaganou, G., & Cavoura, T. (2005). Historical Text Comprehension Reflective Tutorial Dialogue System. *Journal of Educational Technology & Society*, *8*(4), 31–41. http://hdl.handle.net/11615/28224 (cit. on p. 63).

Griskevicius, V., Shiota, M. N., & Neufeld, S. L. (2010). Influence of Different Positive Emotions on Persuasion Processing: A Functional Evolutionary Approach. *Emotion*, *10*(2), 190–206. https://doi.org/10.1037/a0018421 (cit. on p. 2).

Guo, K., Zhong, Y., Li, D., & Chu, S. K. W. (2023). Effects of Chatbot-Assisted In-Class Debates on Students' Argumentation Skills and Task Motivation. *Computers & Education*, *203*, 1–19. https://doi.org/10.1016/j.compedu.2023.104862 (cit. on p. 63).

Hailpern, J., Hinterbichler, E., Leppert, C., Cook, D., & Bailey, B. P. (2007). TEAM STORM: Demonstrating an Interaction Model for Working with Multiple Ideas During Creative Group Work. *Proceedings of the 6th ACM Conference on Creativity & Cognition (C&C)*, 193–202. https://doi.org/10.1145/1254960.1254987 (cit. on p. 61).

Ham, J., Bokhorst, R., Cuijpers, R., van der Pol, D., & Cabibihan, J.-J. (2011). Making Robots Persuasive: The Influence of Combining Persuasive Strategies (Gazing and Gestures) by a Storytelling Robot on its Persuasive Power. *Proceedings of the 3rd International Conference on Social Robotics (ICSR)*, 71–83. https://doi.org/10.1007/978-3-642-25504-5_8 (cit. on pp. 57, 59).

Hammer, S., Lugrin, B., Bogomolov, S., Janowski, K., & André, E. (2016). Investigating Politeness Strategies and Their Persuasiveness for a Robotic Elderly Assistant. In A. Meschtscherjakov, B. De Ruyter, V. Fuchsberger, M. Murer, & M. Tscheligi (Eds.), *Persuasive Technology* (pp. 315–326). Springer International Publishing. https://doi.org/10.1007/978-3-319-31510-2_27 (cit. on pp. 11, 134).

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254 (cit. on p. 13).

Handelman, S. (2009). *Thought Manipulation: The Use and Abuse of Psychological Trickery*. Praeger Publishers. (Cit. on p. 4).

Harmon-Jones, E. (2000). Cognitive Dissonance and Experienced Negative Affect: Evidence that Dissonance Increases Experienced Negative Affect even in the Absence of Aversive Consequences. *Personality and Social Psychology Bulletin*, *26*(12), 1490–1501. https://doi.org/10.1177/01461672002612004 (cit. on p. 64).

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling Validated Versus Being Correct: A Meta-Analysis of Selective Exposure to Information. *Psychological Bulletin*, *135*(4), 555–588. https://doi.org/10.1037/a0015701 (cit. on p. 64).

Heimerl, A., Mertes, S., Schneeberger, T., Baur, T., Liu, A., Becker, L., Rohleder, N., Gebhard, P., & André, E. (2022). Generating Personalized Behavioral Feedback for a Virtual Job Interview Training System Through Adversarial Learning. *Proceedings of the 23rd International Conference on Artificial Intelligence in Education (AIED)*, 679–684. https://doi.org/10.1007/978-3-031-11644-5_67 (cit. on p. 30).

Howell, D. C. (2012). *Statistical Methods for Psychology*. WADSWORTH INC FULFILLMENT. (Cit. on p. 28).

Huang, C.-M., & Mutlu, B. (2013). Modeling and Evaluating Narrative Gestures for Humanlike Robots. *Proceedings of Robotics: Science and Systems Conference IX (RSS)*, *2*, 1–8. https://roboticsproceedings.org/rss09/p26.pdf (cit. on p. 59).

Huang, L., & Yeh, Y. (2017). Meaningful Gamification for Journalism Students to Enhance Their Critical Thinking Skills. *International Journal of Game-Based Learning (IJGBL)*, *7*(2), 47–62. https://doi.org/10.4018/IJGBL.2017040104 (cit. on p. 63).

Huber, T., Limmer, B., & André, E. (2022). Benchmarking Perturbation-Based Saliency Maps for Explaining Atari Agents. *Frontiers in Artificial Intelligence*, *5*, 903875:1–15. https://doi.org/10.3389/frai.2022.903875 (cit. on p. 78).

Huber, T., Schiller, D., & André, E. (2019). Enhancing Explainability of Deep Reinforcement Learning Through Selective Layer-wise Relevance Propagation. *Joint German/Austrian Conference on Artificial Intelligence - Künstliche Intelligenz (KI)*, 188–202. https://doi.org/10.1007/978-3-030-30179-8_16 (cit. on p. 35).

idebate. (2022). This House Believes that Marriage is an Outdated Institution [Accessed 11-12-2023]. https://idebate.net/this-house-believes-that-marriage-is-an-outdated-institution~b872/ (cit. on p. 96).

Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 37:448–456. https://proceedings.mlr.press/v37/ioffe15.html (cit. on p. 72).

Iordanou, K., & Constantinou, C. P. (2015). Supporting Use of Evidence in Argumentation Through Practice in Argumentation and Reflection in the

Context of SOCRATES Learning Environment. *Science Education*, *99*(2), 282–311. https://doi.org/10.1002/sce.21152 (cit. on p. 63).

Isbister, K., Nakanishi, H., Ishida, T., & Nass, C. (2000). Helper Agent: Designing an Assistant for Human-Human Interaction in a Virtual Meeting Space. *Proceedings of the 2000 Conference on Human Factors in Computing Systems (CHI)*, 57–64. https://doi.org/10.1145/332040.332407 (cit. on p. 65).

Jirout, J. J. (2020). Supporting Early Scientific Thinking Through Curiosity. *Frontiers in Psychology*, 1–7. https://doi.org/10.3389/fpsyg.2020.01717 (cit. on p. 169).

Johnson, W. L., Mayer, R. E., André, E., & Rehm, M. (2005). Cross-Cultural Evaluation of Politeness in Tactics for Pedagogical Agents. *Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED)*, 298–305 (cit. on p. 136).

Juel, W. K., Haarslev, F., Ramírez, E. R., Marchetti, E., Fischer, K., Shaikh, D., Manoonpong, P., Hauch, C., Bodenhagen, L., & Krüger, N. (2020). SMOOTH Robot: Design for a Novel Modular Welfare Robot. *Journal of Intelligent & Robotic Systems*, *98*(1), 19–37. https://doi.org/10.1007/s10846-019-01104-z (cit. on p. 60).

Kahneman, D. (2012a). Of 2 Minds: How Fast and Slow thinking Shape Perception and Choice. *Scientific American*, *15*. https://www.scientificamerican.com/article/kahneman-excerpt-thinking-fast-and-slow/ (cit. on p. 6).

Kahneman, D. (2012b). *Thinking, Fast and Slow* (1st ed.). Penguin. (Cit. on pp. 5, 55).

Kaptein, M., Lacroix, J., & Saini, P. (2010). Individual Differences in Persuadability in the Health Promotion Domain. *Proceedings of the 5th International Conference on Persuasive Technology*, 94–105. https://doi.org/10.1007/978-3-642-13226-1_11 (cit. on p. 71).

Karpfinger, C. (2017). *Höhere Mathematik in Rezepten: Begriffe, Sätze und zahlreiche Beispiele in kurzen Lerneinheiten* (3rd). Springer Spektrum. (Cit. on pp. 18, 20, 46).

Kashdan, T. B., Stiksma, M. C., Disabato, D. J., McKnight, P. E., Bekier, J., Kaji, J., & Lazarus, R. (2018). The Five-Dimensional Curiosity Scale: Capturing the Bandwidth of Curiosity and Identifying Four Unique Subgroups of Curious People. *Journal of Research in Personality*, *73*, 130–149. https://doi.org/10.1016/j.jrp.2017.11.011 (cit. on pp. 168, 169).

Kember, D., Leung, D. Y., Jones, A., Loke, A. Y., McKay, J., Sinclair, K., Tse, H., Webb, C., Yuet Wong, F. K., Wong, M., et al. (2000). Development of a Questionnaire to Measure the Level of Reflective Thinking. *Assessment &*

Evaluation in Higher Education, *25*(4), 381–395. https://doi.org/10.1080/713611442 (cit. on pp. 11, 66).

Kharrufa, A., Leat, D., & Olivier, P. (2010). Digital Mysteries: Designing for Learning at the Tabletop. *Proceedings of the 2010 ACM International Conference on Interactive Tabletops and Surfaces (ITS)*, 197–206. https://doi.org/10.1145/1936652.1936689 (cit. on pp. 11, 62, 66).

King, P. M., & Kitchener, K. S. (1994). *Developing Reflective Judgment: Understanding and Promoting Intellectual Growth and Critical Thinking in Adolescents and Adults*. Jossey-Bass Inc.,U.S. (Cit. on p. 55).

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 1–15. https://doi.org/10.48550/arXiv.1412.6980 (cit. on pp. 25, 72).

Kipp, M., & Gebhard, P. (2008). IGaze: Studying Reactive Gaze Behavior in Semi-immersive Human-avatar Interactions. *International Workshop on Intelligent Virtual Agents (IVA)*, 191–199. https://doi.org/10.1007/978-3-540-85483-8_19 (cit. on p. 57).

Kong, S. C., & Song, Y. (2015). An Experience of Personalized Learning Hub Initiative Embedding BYOD For Reflective Engagement in Higher Education. *Computers & Education*, *88*, 227–240. https://doi.org/10.1016/j.compedu.2015.06.003 (cit. on pp. 7, 56).

Konidaris, G., Osentoski, S., & Thomas, P. (2011). Value Function Approximation in Reinforcement Learning Using the Fourier Basis. *Proceedings of the 25th Conference on Artificial Intelligence (AAAI)*, 380–385. http://lis.csail.mit.edu/pubs/konidaris-aaai11a.pdf (cit. on pp. 45–48).

Kopec, J. A., & Esdaile, J. M. (1990). Bias in Case-control Studies - A Review. *Journal of Epidemiology and Community Health*, *44*(3), 179–189. https://doi.org/10.1136/jech.44.3.179 (cit. on p. 163).

Körber, M. (2019). Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. *Proceedings of the 20th Congress of the International Ergonomics Association (IEA)*, 13–30. https://doi.org/10.1007/978-3-319-96074-6_2 (cit. on pp. xxii–xxiv, 148, 151, 153, 154, 161, 162, 240, 252).

Kotikalapudi, R., & contributors. (2017). Keras-vis. https://github.com/raghakot/keras-vis (cit. on p. 74).

Krämer, N. C., Simons, N., & Kopp, S. (2007). The Effects of an Embodied Conversational Agent's Nonverbal Behavior on User's Evaluation and Behavioral

Mimicry. *Proceedings of the 7th ACM International Conference on Intelligent Virtual Agents (IVA)*, 238–251. https://doi.org/10.1007/978-3-540-74997-4_22 (cit. on p. 104).

Kusajima, S., & Sumi, Y. (2018). Activating Group Discussion by Topic Providing Bots. *IEICE Transactions on Information and Systems*, *101-D*(4), 856–864. https://doi.org/10.1587/transinf.2016IIP0034 (cit. on p. 65).

Langhammer, S. (2018). *A Debating Ontology for Argumentative Dialogue Systems* [Bachelor's Thesis]. Universität Ulm. https://oparu.uni-ulm.de/xmlui/handle/123456789/6805 (cit. on pp. 96–98, 109).

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. *Nature Communications*, *10*(1), 1096:1–8. https://doi.org/10.1038/s41467-019-08987-4 (cit. on pp. 30, 35, 69).

Le, D. T., Nguyen, C.-T., & Nguyen, K. A. (2018). Dave the Debater: A Retrieval-Based and Generative Argumentative Dialogue Agent. *Proceedings of the 5th Workshop on Argument Mining*, 121–130. https://doi.org/10.18653/v1/W18-5215 (cit. on p. 64).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539 (cit. on p. 31).

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. (1989). Handwritten Digit Recognition with a Back-Propagation Network. *Proceedings of the 1989 Conference on Neural Information Processing Systems (NeurIPS)*, *2*, 1–9 (cit. on p. 22).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791 (cit. on p. 22).

Lee, M. L., & Dey, A. K. (2011). Reflecting on Pills and Phone Use: Supporting Awareness of Functional Abilities for Older Adults. *Proceedings of the 2011 SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2095–2104. https://doi.org/10.1145/1978942.1979247 (cit. on pp. 11, 66).

Leijen, Ä., Lam, I., Wildschut, L., Simons, P. R.-J., & Admiraal, W. (2009). Streaming Video to Enhance Students' Reflection in Dance Education. *Computers & Education*, *52*(1), 169–176. https://doi.org/10.1016/j.compedu.2008.07.010 (cit. on pp. 11, 66).

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, *18*(17), 1–5. https://www.jmlr.org/papers/volume18/16-365/16-365.pdf (cit. on p. 84).

Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., & Bhogal, R. S. (1997). The Persona Effect: Affective Impact of Animated Pedagogical Agents. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 359–366. https://doi.org/10.1145/258549.258797 (cit. on p. 103).

Lin, Y.-T., Doong, H.-S., & Eisingerich, A. B. (2021). Avatar Design of Virtual Salespeople: Mitigation of Recommendation Conflicts. *Journal of Service Research*, *24*(1), 141–159. https://doi.org/10.1177/1094670520964872 (cit. on p. 103).

Liu, Q., & Nesbit, J. C. (2024). The Relation Between Need for Cognition and Academic Achievement: A Meta-Analysis. *Review of Educational Research*, (2), 155–192. https://doi.org/10.3102/00346543231160474 (cit. on p. 174).

Liu, Z., Nersessian, N., & Stasko, J. (2008). Distributed Cognition As a Theoretical Framework for Information Visualization. *IEEE Transactions on Visualization and Computer Graphics*, *14*(6), 1173–1180. https://doi.org/10.1109/TVCG.2008.121 (cit. on p. 176).

Lloyd, N., Chowdhry, A., & Lewis, P. R. (2020). Trustworthy Embodied Virtual Agents. In N. Lee (Ed.), *Encyclopedia of Computer Graphics and Games* (pp. 1–6). Springer International Publishing. https://doi.org/10.1007/978-3-319-08234-9_524-1 (cit. on p. 103).

Lyons, N. (2006). Reflective Engagement As Professional Development in the Lives of University Teachers. *Teachers and Teaching: Theory and Practice*, *12*(2), 151–168. https://williammarylyons.com/assets/docs/NonaReflectiveEngagement.277151038.pdf (cit. on pp. 7, 56).

Mance, I., & Vogel, E. K. (2013). Visual Working Memory. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(2), 179–190. https://doi.org/10.1002/wcs.1219 (cit. on pp. 168, 169).

Maricchiolo, F., Gnisci, A., Bonaiuto, M., & Ficca, G. (2009). Effects of Different Types of Hand Gestures in Persuasive Speech on Receivers' Evaluations. *Language and Cognitive Processes*, *24*(2), 239–266. https://doi.org/10.1080/01690960802159929 (cit. on pp. 57, 58, 185).

Mason, M. (2007). Critical Thinking and Learning. *Educational Philosophy and Theory*, *39*(4), 339–349. https://doi.org/10.1111/j.1469-5812.2007.00343.x (cit. on pp. 56, 111).

Mataillet, D. (2019). Fifteen Adjectives to Describe Fine Wines. Retrieved 2023-12-04, from https://france-amerique.com/en/fifteen-adjectives-to-describe-fine-wines/ (cit. on p. 2).

McCrae, R. R., & Sutin, A. R. (2009). Openness to Experience. *Handbook of Individual Differences in Social Behavior*, 257–373. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=f763a87db2e473584b6577076e6ff43a2348bfee (cit. on pp. 167, 169).

McHugh, M. L. (2013). The Chi-Square Test of Independence. *Biochemia Medica*, *23*(2), 143–149. https://doi.org/10.11613/BM.2013.018 (cit. on p. 161).

McKnight, P. E., & Najab, J. (2010). Mann-Whitney U Test. In *The Corsini Encyclopedia of Psychology* (pp. 1–1). WILEY Online Library. https://doi.org/10.1002/9780470479216.corpsy0524 (cit. on pp. 153, 160).

McNeill, D. (1992). Hand and Mind: What Gestures Reveal about Thought. *University of Chicago Press* (cit. on p. 58).

Meara, P. M. (1992). *EFL Vocabulary Tests*. Citeseer. https://eric.ed.gov/?id=ED362046 (cit. on pp. 168, 169).

Miao, F., Kozlenkova, I. V., Wang, H., Xie, T., & Palmatier, R. W. (2022). An Emerging Theory of Avatar Marketing. *Journal of Marketing*, *86*(1), 67–90 (cit. on pp. 96, 103, 104, 189).

Millecamp, M., Htun, N.-N., Conati, C., & Verbert, K. (2020). What's in a User? Towards Personalising Transparency for Music Recommender Interfaces. *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP)*, 173–182. https://doi.org/10.1145/3340631.3394844 (cit. on p. 169).

Mischel, W., & Ebbesen, E. B. (1970). Attention in Delay of Gratification. *Journal of Personality and Social Psychology*, *16*(2), 329–337. https://doi.org/10.1037/h0029815 (cit. on p. 41).

Mishra, K., Samad, A. M., Totala, P., & Ekbal, A. (2022). PEPDS: A Polite and Empathetic Persuasive Dialogue System for Charity Donation. *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, 424–440. https://aclanthology.org/2022.coling-1.34/ (cit. on p. 64).

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable Ai Systems. *ACM Transactions on Interactive Intelligent Systems (TIIS)*, *11*(3-4), 1–45. https://doi.org/10.1145/3387166 (cit. on p. 78).

Möller, S. (2003). Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems. *ITU-T Recommendation* (cit. on pp. 148, 151, 152, 249).

Molnar, C. (2019). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Lulu.com. https://christophm.github.io/interpretable-ml-book (cit. on p. 30).

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K.-R. (2019). Layer-wise Relevance Propagation: An Overview. In *Explainable Ai: Interpreting, Explaining and Visualizing Deep Learning* (pp. 193–209). Springer. https://doi.org/10.1007/978-3-030-28954-6_10 (cit. on pp. 34, 35).

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition. *Pattern Recognition*, *65*, 211–222. https://doi.org/10.1016/j.patcog.2016.11.008 (cit. on p. 35).

Negi, S., & Mitra, R. (2020). Fixation Duration and the Learning Process: An Eye Tracking Study with Subtitled Videos. *Journal of Eye Movement Research*, *13*(6), 1–15. https://doi.org/10.16910/jemr.13.6.1 (cit. on p. 170).

Newman, R., Furnham, A., Weis, L., Gee, M., Cardos, R., Lay, A., & McClelland, A. (2016). Non-verbal Presence: How Changing Your Behaviour Can Increase Your Ratings for Persuasion, Leadership and Confidence. *Psychology*, *7*(4), 488–499. https://doi.org/10.4236/psych.2016.74050 (cit. on pp. 76, 78, 79, 93, 185).

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., & Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS)*, *29*, 3395–3403. https://dl.acm.org/doi/abs/10.5555/3157382.3157477 (cit. on p. 30).

Nojavanasghari, B., Gopinath, D., Koushik, J., Baltrušaitis, T., & Morency, L.-P. (2016). Deep Multimodal Fusion for Persuasiveness Prediction. *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI)*, 284–288. https://doi.org/10.1145/2993148.2993176 (cit. on p. 71).

Novikova, J., Dušek, O., & Rieser, V. (2017). The E2E Dataset: New Challenges For End-to-End Generation. *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 201–206. https://doi.org/10.18653/v1/W17-5525 (cit. on p. 36).

O'Brien, H. L., Cairns, P., & Hall, M. (2018). A Practical Approach to Measuring User Engagement with the Refined User Engagement Scale (UES) and New UES Short Form. *International Journal of Human-Computer Studies*, *112*, 28–39. https://doi.org/10.1016/j.ijhcs.2018.01.004 (cit. on pp. xxi–xxiv, 148, 151, 154, 254).

Ortony, A., Clore, G. L., & Collins, A. (2022). *The Cognitive Structure of Emotions*. Cambridge University Press. (Cit. on p. 57).

Ottley, A., Crouser, R., Ziemkiewicz, C., & Chang, R. (2012). Priming Locus of Control to Affect Performance. *Computer Science: Faculty Publications*, 1–3. https://scholarworks.smith.edu/csc_facpubs/143 (cit. on pp. 169, 176).

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. *Proceedings of the 2022 Conference on Neural Information Processing Systems (NeurIPS)*, 1–68. https://doi.org/10.48550/arXiv.2203.02155 (cit. on p. 64).

Pariser, E. (2012). *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin Publishing Group. (Cit. on p. 5).

Paul, R. W. (1981). Teaching Critical Thinking in The "Strong" Sense: A Focus on Self-deception, World Views, and a Dialectical Mode of Analysis. *Informal Logic - Reasoning and Argumentation in Theory and Practice*, *4*(2), 1–6. https://doi.org/10.22329/il.v4i2.2766 (cit. on p. 56).

Paul, R. W. (1990). Critical and Reflective Thinking: A Philosophical Perspective. In *Dimensions of Thinking and Cognitive Instruction* (pp. 445–494). Taylor & Francis Inc. (Cit. on pp. 56, 110, 111, 130, 165).

Peters, J., & Hoetjes, M. (2017). The Effect of Gesture on Persuasive Speech. *Proceedings of the 2017 Conference of the International Speech Communication Association (INTERSPEECH)*, 659–663. http://hdl.handle.net/2066/181138 (cit. on pp. 6, 7, 57, 59, 76, 78, 79, 93, 185).

Peters, U. (2022). What Is the Function of Confirmation Bias? *Erkenntnis*, *87*(3), 1351–1376. https://doi.org/10.1007/s10670-020-00252-1 (cit. on pp. 5, 130).

Petty, R. E., & Cacioppo, J. T. (1986). The Elaboration Likelihood Model of Persuasion. In *Communication and Persuasion. Springer Series in Social Psychology* (pp. 1–24). Springer, New York, NY. https://doi.org/10.1007/978-1-4612-4964-1_1 (cit. on pp. 6, 52).

Petukhova, V., Mayer, T., Malchanau, A., & Bunt, H. (2017). Virtual Debate Coach Design: Assessing Multimodal Argumentation Performance. *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI)*, 41–50. https://doi.org/10.1145/3136755.3136775 (cit. on p. 64).

Poggi, I., & Vincze, L. (2009). Gesture, Gaze and Persuasive Strategies in Political Discourse. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04793-0_5 (cit. on p. 57).

Polyak, B. T. (1964). Some Methods of Speeding Up the Convergence of Iteration Methods. *USSR Computational Mathematics and Mathematical Physics*, *4*(5), 1–17. https://doi.org/10.1016/0041-5553(64)90137-5 (cit. on p. 24).

Ponnuswami, A. K., Pattabiraman, K., Wu, Q., Gilad-Bachrach, R., & Kanungo, T. (2011). On Composition of a Federated Web Search Result Page: Using Online Users to Provide Pairwise Preference for Heterogeneous Verticals. *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM)*, 715–724. https://doi.org/10.1145/1935826.1935922 (cit. on p. 11).

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., & Androutsopoulos, I. (2015). Semeval-2015 Task 12: Aspect Based Sentiment Analysis. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval)*, 486–495. https://aclanthology.org/S15-2082.pdf (cit. on pp. 117, 118, 225, 233).

Prakken, H. (2000). On Dialogue Systems with Speech Acts, Arguments, and Counterarguments. In M. Ojeda-Aciego, I. P. de Guzmán, G. Brewka, & L. Moniz Pereira (Eds.), *Logics in Artificial Intelligence* (pp. 224–238). Springer. https://doi.org/10.1007/3-540-40006-0_16 (cit. on p. 36).

Prakken, H. (2005). Coherence and Flexibility in Dialogue Games for Argumentation. *Journal of Logics and Computation*, *15*(6), 1009–1040. https://doi.org/10.1093/logcom/exi046 (cit. on p. 36).

Qiu, S., Bozzon, A., Birk, M. V., & Gadiraju, U. (2021). Using Worker Avatars to Improve Microtask Crowdsourcing. *5*(CSCW2), 1–28. https://doi.org/10.1145/3476063 (cit. on pp. 96, 103, 189).

Rach, N. (2022). *Towards Flexible Argumentation with Conversational Agents* [Doctoral dissertation, Universität Ulm]. https://oparu.uni-ulm.de/xmlui/handle/123456789/44152 (cit. on p. 116).

Rach, N., Minker, W., & Ultes, S. (2018a). Markov Games for Persuasive Dialogue. *Proceedings of the 2018 International Conference on Computational Models of Argument (COMMA)*, *302*, 213–220. https://doi.org/10.3233/978-1-61499-906-5-213 (cit. on pp. 64, 124).

Rach, N., Weber, K., Aicher, A., Lingenfelser, F., André, E., & Minker, W. (2019). Emotion Recognition Based Preference Modelling in Argumentative Dialogue Systems. *Proceedings of the 2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 838–843. https://doi.org/10.1109/PERCOMW.2019.8730650 (cit. on p. 257).

Rach, N., Weber, K., Pragst, L., André, E., Minker, W., & Ultes, S. (2018b). EVA: A Multimodal Argumentative Dialogue System. *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI)*, 551–552. https://doi.org/10.1145/3242969.3266292 (cit. on p. 257).

Rach, N., Weber, K., Yang, Y., Ultes, S., André, E., & Minker, W. (2021). EVA 2.0: Emotional and Rational Multimodal Argumentation between Virtual Agents. *it - Information Technology*, *63*(1), 17–30. https://doi.org/10.1515/itit-2020-0050 (cit. on pp. 36, 37, 64, 116, 124, 125, 129, 135–137, 144, 258).

Rakshit, G., Bowden, K. K., Reed, L., Misra, A., & Walker, M. A. (2019). Debbie, the Debate Bot of the Future. *Proceedings of the 8th International Workshop on Spoken Dialog Systems (IWSDS): Advanced Social Interaction with Agents*, 45–52. https://doi.org/10.1007/978-3-319-92108-2_5 (cit. on p. 64).

Ras, G., Xie, N., Van Gerven, M., & Doran, D. (2022). Explainable Deep Learning: A Field Guide for the Uninitiated. *Journal of Artificial Intelligence Research*, *73*, 329–396. https://doi.org/10.1613/jair.1.13200 (cit. on pp. 31–33).

Rebolledo-Mendez, G., Burden, D., & de Freitas, S. (2008). A Model of Motivation for Virtual-Worlds Avatars. *Lecture Notes in Computer Science*, *5208*, 535–536. https://doi.org/10.1007/978-3-540-85483-8_76 (cit. on p. 103).

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084*, 1–11. https://doi.org/10.48550/arXiv.1908.10084 (cit. on p. 99).

Rezaei Khavas, Z. (2021). A Review on Trust in Human-Robot Interaction. *arXiv e-prints*, 1–10. https://doi.org/10.48550/arXiv.2105.10045 (cit. on p. 13).

Rhim, J., Kwak, S. S., Lim, A., & Millar, J. (2023). The Dynamic Nature of Trust: Trust in Human-Robot Interaction Revisited. *arXiv preprint arXiv:2303.04841*, 1–6. https://doi.org/10.48550/arXiv.2303.04841 (cit. on pp. 13, 188).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 1135–1144. https://doi.org/10.1145/2939672.2939778 (cit. on p. 35).

Ritschel, H., Baur, T., & André, E. (2017). Adapting a Robot's Linguistic Style Based on Socially-Aware Reinforcement Learning. *Proceedings of the 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 378–384. https://doi.org/10.1109/ROMAN.2017.8172330 (cit. on pp. 44, 137, 142).

Ritschel, H., Kiderle, T., Weber, K., & André, E. (2020a). Multimodal Joke Presentation for Social Robots Based on Natural-language Generation and Nonverbal Behaviors. *Proceedings of the 2nd Workshop on Natural Language Generation for Human–robot Interaction (NLG4HRI)*, 1–3. https://hbuschme.github.io/nlg-hri-workshop-2020/assets/papers/NLG4HRI_paper_3.pdf (cit. on p. 258).

Ritschel, H., Kiderle, T., Weber, K., Lingenfelser, F., Baur, T., & André, E. (2020b). Multimodal Joke Generation and Paralinguistic Personalization for a Socially-aware Robot. *Proceedings of the 22nd International International Conference on Practical Applications of Agents, Multi-Agent Systems, and Trustworthiness. (PAAMS)*, 278–290. https://doi.org/10.1007/978-3-030-49778-1_22 (cit. on p. 258).

Ritschel, H., Seiderer, A., Janowski, K., Wagner, S., & André, E. (2019). Adaptive Linguistic Style for an Assistive Robotic Health Companion Based on Explicit Human Feedback. *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA)*, 247–255. https://doi.org/10.1145/3316782.3316791 (cit. on pp. 134, 135).

Roberts, B. W., Lejuez, C., Krueger, R. F., Richards, J. M., & Hill, P. L. (2014). What Is Conscientiousness and How Can It Be Assessed? *Developmental Psychology*, *50*(5), 1315–1330. https://doi.org/10.1037/a0031109 (cit. on pp. 167, 169).

Rodgers, C. (2002). Defining Reflection: Another Look at John Dewey and Reflective Thinking. *Teachers College Record*, *104*(4), 842–866. https://doi.org/10.1111/1467-9620.00181 (cit. on p. 56).

Rodman, G. J. (2010). Facilitating the Teaching-Learning Process Through the Reflective Engagement of Pre-service Teachers. *Australian Journal of Teacher Education*, *35*(2), 20–34. https://doi.org/10.14221/ajte.2010v35n2.2 (cit. on pp. 7, 56).

Rogers, M., & Smith, K. H. (1993). Public Perceptions of Subliminal Advertising: Why Practitioners Shouldn't Ignore This Issue. *Journal of Advertising Research*, *33*(2), 10–18. https://psycnet.apa.org/record/1993-39581-001 (cit. on p. 8).

Rosenfeld, A., & Kraus, S. (2016). Strategical Argumentative Agent for Human Persuasion. *Proceedings of the 22nd European Conference on Artificial Intelligence (ECAI)*, *16*, 320–329. https://doi.org/10.3233/978-1-61499-672-9-320 (cit. on pp. 64, 124).

Rotter, J. B. (1966). Generalized Expectancies for Internal Versus External Control of Reinforcement. *Psychological Monographs: General and Applied*, *80*(1), 1–28. https://doi.org/10.1037/h0092976 (cit. on pp. 168, 169).

Roussou, M., Perry, S., Katifori, A., Vassos, S., Tzouganatou, A., & McKinney, S. (2019). Transformation through Provocation? *Proceedings of the 2019 Conference on Human Factors in Computing Systems (CHI)*, 1–13. https://doi.org/10.1145/3290605.3300857 (cit. on p. 64).

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Representations by Back-Propagating Errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0 (cit. on pp. 16, 19).

Russell, J. A. (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714 (cit. on p. 57).

Santos, J. L., Verbert, K., Govaerts, S., & Duval, E. (2013). Addressing Learner Issues with Stepup! An Evaluation. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge (LAK)*, 14–22. https://doi.org/10.1145/2460296.2460301 (cit. on pp. 11, 62, 66).

Schön, D. A. (1983). *The Reflective Practicioner: How Professionals Think in Action*. Basic Books, New York. (Cit. on p. 65).

Schraagen, J. M., Elsasser, P., Fricke, H., Hof, M., & Ragalmuto, F. (2020). Trusting the X in XAI: Effects of Different Types of Explanations by a Self-Driving Car on Trust, Explanation Satisfaction and Mental Models. *Proceedings of the 2020 Human Factors and Ergonomics Society Annual Meeting*, *64*(1), 339–343. https://doi.org/10.1177/1071181320641077 (cit. on p. 29).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626. https://doi.org/10.1109/ICCV.2017.74 (cit. on pp. 31, 32, 74).

Sharma, N., Liao, Q. V., & Xiao, Z. (2024). Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. *Proceedings of the 2024 Conference on Human Factors in Computing Systems (CHI)*, 1–17. https://doi.org/10.1145/3613904.3642459 (cit. on pp. 65, 186).

Sidner, C. L., Kidd, C. D., Lee, C., & Lesh, N. (2004). Where to Look: A Study of Human-robot Engagement. *Proceedings of the 9th International Conference on Intelligent User Interfaces*, 78–84. https://doi.org/10.1145/964442.964458 (cit. on p. 10).

Siegel, M., Breazeal, C., & Norton, M. I. (2009). Persuasive Robotics: The Influence of Robot Gender on Human Behavior. *Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2563–2568. https://doi.org/10.1109/IROS.2009.5354116 (cit. on p. 150).

Silpasuwanchai, C., Ma, X., Shigemasu, H., & Ren, X. (2016). Developing a Comprehensive Engagement Framework of Gamification for Reflective Learning. *Proceedings of the 2016 ACM Conference on Designing Interactive Systems (DIS)*, 459–472. https://doi.org/10.1145/2901790.2901836 (cit. on pp. 62, 132).

Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). Hand Keypoint Detection in Single Images using Multiview Bootstrapping. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4645–4653. https://doi.org/10.1109/CVPR.2017.494 (cit. on p. 81).

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *Proceedings of the International Conference on Learning Representations (ICLR)*, 1–8. https://doi.org/10.48550/arXiv.1312.6034 (cit. on p. 31).

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*, 1–14. https://doi.org/10.48550/arXiv.1409.1556 (cit. on p. 29).

Sixt, L., Granz, M., & Landgraf, T. (2020). When Explanations Lie: Why Modified BP Attribution Fails. *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 9046–9057. https://proceedings.mlr.press/v119/sixt20a.html (cit. on p. 31).

Slonim, N., Bilu, Y., Alzate, C., Bar-Haim, R., Bogin, B., Bonin, F., Choshen, L., Cohen-Karlik, E., Dankin, L., Edelstein, L., et al. (2021). An Autonomous Debating System. *nature*, *591*(7850), 379–384. https://doi.org/10.1038/s41586-021-03215-w (cit. on p. 64).

Smith, B. N., Xu, A., & Bailey, B. P. (2010). Improving Interaction Models for Generating and Managing Alternative Ideas During Early Design Work. *Proceedings of Graphics Interface (GI)*, 121–128. https://dl.acm.org/doi/abs/10.5555/1839214.1839236 (cit. on p. 61).

Sorlin, S. (2016). Manipulative Moves: Between Persuasion and Coercion. In S. Sorlin (Ed.), *Language and Manipulation in House of Cards: A Pragma-Stylistic Perspective* (pp. 107–141). Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-55848-0_4 (cit. on p. 3).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958. https://dl.acm.org/doi/abs/10.5555/2627435.2670313 (cit. on p. 82).

Stab, C., & Gurevych, I. (2014). Annotating Argument Components and Relations in Persuasive Essays. *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, 1501–1510. https://aclanthology.org/C14-1142.pdf (cit. on p. 38).

Steca, P. (2021). Locus of Control. In *Encyclopedia of Quality of Life and Well-being Research* (pp. 1–4). Springer. https://doi.org/10.1007/978-3-319-69909-7_1688-2 (cit. on pp. 168, 175).

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. F. (2020). Learning to Summarize with Human Feedback. *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, *33*, 3008–3021. https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf (cit. on p. 64).

Streeck, J. (2008). Gesture in Political Communication: A Case Study of the Democratic Presidential Candidates During the 2004 Primary Campaign. *Research on Language and Social Interaction*, *41*(2), 154–186. https://doi.org/10.1080/08351810802028662 (cit. on p. 53).

Susser, D., Roessler, B., & Nissenbaum, H. (2019). Technology, Autonomy, and Manipulation. *Internet Policy Review*, *8*(2), 1–22. https://doi.org/10.14763/2019.2.1410 (cit. on p. 4).

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. The MIT Press Cambridge, MA, USA. (Cit. on pp. 40–44, 139).

Tinnes, L. (2022). *Understanding Subliminal Persuasive Body Language in Political Speeches via Explainable Artificial Intelligence* [Bachelor's Thesis. Under joint sup. of Weber, K. & Huber, T.]. Universität of Augsburg. (Cit. on p. 68).

Toker, D., & Conati, C. (2014). Eye Tracking to Understand User Differences in Visualization Processing with Highlighting Interventions. *Proceedings of the 22nd International Conference on User Modeling, Adaptation, and Personalization (UMAP)*, 219–230. https://doi.org/10.1007/978-3-319-08786-3_19 (cit. on pp. 13, 165, 177).

Toker, D., Conati, C., Steichen, B., & Carenini, G. (2013). Individual User Characteristics and Information Visualization: Connecting the Dots Through Eye Tracking. *Proceedings of the 2013 SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 295–304. https://doi.org/10.1145/2470654.2470696 (cit. on pp. 13, 165).

Tomar, P. S., Mathur, K., & Suman, U. (2024). Fusing Facial and Speech Cues for Enhanced Multimodal Emotion Recognition. *International Journal of Information Technology*, *16*(3), 1397–1405. https://doi.org/10.1007/s41870-023-01697-7 (cit. on p. 58).

Tomczak, M., & Tomczak, E. (2014). The Need to Report Effect Size Estimates Revisited. An Overview of Some Recommended Measures of Effect Size. *Trends in Sport Sciences*, *21*(1), 19–24 (cit. on pp. 150, 160, 171).

van Kleef, G. (2014). Emotions As Agents of Social Influence. In *The Oxford Handbook of Social Influence* (pp. 237–256). Oxford University Press. https:

//doi.org/10.1093/oxfordhb/9780199859870.013.19 (cit. on pp. 3, 5, 53, 54, 58).

van Kleef, G., van den Berg, H., & Heerdink, M. W. (2015). The Persuasive Power of Emotions: Effects of Emotional Expressions on Attitude Formation and Change. *Journal of Applied Psychology*, *100*(4), 1124–1142. https://doi.org/10.1037/apl0000003 (cit. on pp. 3, 5).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, *30*, 6000–6010. https://dl.acm.org/doi/abs/10.5555/3295222.3295349 (cit. on p. 99).

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of Features, Conjunctions, and Objects in Visual Working Memory. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(1), 92–144. https://doi.org/10.1037/0096-1523.27.1.92 (cit. on pp. 168, 169).

Walton, D. (2005). *Fundamentals of Critical Argumentation*. Cambridge University Press. (Cit. on pp. 6, 36).

Walton, D. (2010). Types of Dialogue and Burdens of Proof. *Proceedings of the 2010 International Conference on Computational Models of Argument (COMMA)*, 13–24. https://dl.acm.org/doi/abs/10.5555/1860828.1860832 (cit. on p. 36).

Wang, Y., Lucas, G., Khooshabeh, P., De Melo, C., & Gratch, J. (2015). Effects of Emotional Expressions on Persuasion. *Social Influence*, *10*(4), 236–249. https://doi.org/10.1080/15534510.2015.1081856 (cit. on p. 57).

Ward, A., & Litman, D. J. (2011). Adding Abstractive Reflection to a Tutorial Dialog System. *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, 1–6. https://cdn.aaai.org/ocs/2575/2575-11242-1-PB.pdf (cit. on p. 62).

Watkins, C. J. (1989). *Learning from Delayed Rewards* [Doctoral dissertation]. King's College, Cambridge United Kingdom. (Cit. on p. 43).

Watkins, C. J., & Dayan, P. (1992). Q-Learning. *Machine Learning*, *8*, 279–292. https://doi.org/10.1007/BF00992698 (cit. on p. 43).

Weber, K. (2017). *Adaption eines Sozialen Roboters auf Basis von Bestärkendem Lernen mit Linearer Funktionsapproximation und Sozialen Signalen* [Master's Thesis]. University of Augsburg. (Cit. on pp. 40, 48, 49, 141, 257).

Weber, K., Aicher, A., Minker, W., Ultes, S., & André, E. (2023a). Fostering User Engagement in the Critical Reflection of Arguments. *Proceedings of the 13th International Workshop on Spoken Dialogue Systems (IWSDS)*, 1–16.

https://doi.org/10.48550/arXiv.2308.09061 (cit. on pp. 95, 106, 108, 127, 129, 144, 159, 181, 258).

Weber, K., Hogh, N., Conati, C., & André, E. (2024). A Gaze into Argumentative Chatbots: Exploring the Influence of Challenger Arguments on Reflection and Attention. *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents (IVA)*, 1–10. https://doi.org/10.1145/3652988.3673935 (cit. on pp. 95, 106, 108, 127, 129, 144, 166, 181, 259).

Weber, K., Janowski, K., Rach, N., Weitz, K., Minker, W., Ultes, S., & André, E. (2020a). Predicting Persuasive Effectiveness for Multimodal Behavior Adaptation using Bipolar Weighted Argument Graphs. *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 1476–1484. https://dl.acm.org/doi/abs/10.5555/3398761.3398931 (cit. on pp. 36, 112, 116, 127, 129, 135, 144, 258).

Weber, K., Rach, N., Minker, W., & André, E. (2020b). How to Win Arguments: Empowering Virtual Agents to Improve Their Persuasiveness. *Datenbank-Spektrum*, *20*, 161–169. https://doi.org/10.1007/s13222-020-00345-9 (cit. on pp. 36, 64, 112, 124, 125, 127, 129, 135, 144, 258).

Weber, K., Ritschel, H., Aslan, I., Lingenfelser, F., & André, E. (2018a). How to Shape the Humor of a Robot - Social Behavior Adaptation Based on Reinforcement Learning. *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI)*, 154–162. https://doi.org/10.1145/3242969.3242976 (cit. on pp. 40, 129, 135–137, 144, 257).

Weber, K., Ritschel, H., Lingenfelser, F., & André, E. (2018b). Real-Time Adaptation of a Robotic Joke Teller Based on Human Social Signals. *Proceedings of the 17th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2259–2261. https://dl.acm.org/doi/abs/10.5555/3237383.3238141 (cit. on p. 257).

Weber, K., Tinnes, L., Huber, T., & André, E. (2023b). Exploring the Effect of Visual-Based Subliminal Persuasion in Public Speeches Using Explainable AI techniques. *Proceedings of the 25th International Conference on Human-Computer Interaction (HCII)*, 381–397. https://doi.org/10.1007/978-3-031-35891-3_23 (cit. on pp. 15, 29, 68, 94, 259).

Weber, K., Tinnes, L., Huber, T., Heimerl, A., Pohlen, E., Reinecker, M.-L., & Andé, E. (2020c). Towards Demystifying Subliminal Persuasiveness: Using XAI-Techniques to Highlight Persuasive Markers of Public Speeches. *Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS)*, 113–128. https://doi.org/10.1007/978-3-030-51924-7_7 (cit. on pp. 15, 29, 68, 94, 258).

Weitz, K., Hassan, T., Schmid, U., & Garbas, J.-U. (2019). Deep-Learned Faces of Pain and Emotions: Elucidating the Differences of Facial Expressions with the Help of Explainable AI methods. *tm-Technisches Messen*, *86*(7-8), 404–412. https://doi.org/10.1515/teme-2019-0024 (cit. on p. 29).

Wessler, J., Schneeberger, T., Christidis, L., & Gebhard, P. (2022). Virtual Backlash: Nonverbal Expression of Dominance Leads to Less Liking of Dominant Female Versus Male Agents. *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents (IVA)*, 1–8. https://dl.acm.org/doi/abs/10.1145/3514197.3549682 (cit. on pp. 150, 157).

Wiewiora, E. (2010). Reward Shaping. In G. I. Sammut Claude and Webb (Ed.), *Encyclopedia of Machine Learning* (pp. 863–865). Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_731 (cit. on p. 137).

Wood, A. W. (2014). Coercion, Manipulation, Exploitation. *Manipulation: Theory and practice*, 17–50 (cit. on p. 4).

Yi, X., Hong, L., Zhong, E., Liu, N. N., & Rajan, S. (2014). Beyond Clicks: Dwell Time for Personalization. *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys)*, 113–120. https://doi.org/10.1145/2645710.2645724 (cit. on pp. 11, 66).

Yip, J., Wong, S.-H., Yick, K.-L., Chan, K., & Wong, K.-H. (2019). Improving Quality of Teaching and Learning in Classes by Using Augmented Reality Video. *Computers & Education*, *128*, 88–101. https://doi.org/10.1016/j.compedu.2018.09.014 (cit. on p. 62).

Yokoyama, H., & Daibo, I. (2012). Effects of Gaze and Speech Rate on Receivers' Evaluations of Persuasive Speech. *Psychological Reports*, *110*(2), 663–676. https://doi.org/10.2466/07.11.21.28.PR0.110.2.663-676 (cit. on p. 53).

Zanbaka, C., Goolkasian, P., & Hodges, L. (2006). Can a Virtual Cat Persuade You? The Role of Gender and Realism in Speaker Persuasiveness. *Proceedings of the 2006 SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 1153–1162. https://doi.org/10.1145/1124772.1124945 (cit. on p. 157).

Zanot, E. J., Pincus, J. D., & Lamp, E. J. (1983). Public Perceptions of Subliminal Advertising. *Journal of Advertising*, *12*(1), 39–45. https://doi.org/10.1080/00913367.1983.10672829 (cit. on p. 8).

Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. *Proceedings of the 2014 European Conference on Computer Vision (ECCV)*, 818–833. https://doi.org/10.1007/978-3-319-10590-1_53 (cit. on p. 35).

Zhang, Y., & Chen, X. (2020). Explainable Recommendation: A Survey and New Perspectives. *Foundations and Trends in Information Retrieval*, *14*(1), 1–101. https://doi.org/10.1561/1500000066 (cit. on p. 29).

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning Deep Features for Discriminative Localization. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2921–2929. https://doi.org/10.1109/CVPR.2016.319 (cit. on pp. 31, 32).

Ziemkiewicz, C., Crouser, R. J., Yauilla, A. R., Su, S. L., Ribarsky, W., & Chang, R. (2011). How Locus of Control Influences Compatibility with Visualization Style. *Proceedings of the 2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 81–90. https://doi.org/10.1109/VAST.2011.6102445 (cit. on pp. 13, 165, 169).

Ziemkiewicz, C., Ottley, A., Crouser, R. J., Yauilla, A. R., Su, S. L., Ribarsky, W., & Chang, R. (2013). How Visualization Layout Relates to Locus of Control and Other Personality Factors. *Proceedings of the 2013 IEEE Conference on Transactions on Visualization and Computer Graphics (TVCG)*, *19*(7), 1109–1121. https://doi.org/10.1109/TVCG.2012.180 (cit. on pp. 169, 176).

Appendix A

Speech Links

- **Description.** This appendix is a repository of links to the videos that were annotated and employed in training the neural networks to investigate individual differences in subliminal persuasion. Some of these videos are no longer available (✗) as of the completion date of this thesis.
- **Data.** List of links to videos.
- **Source.** The data was manually collected from https://www.youtube.com.
- **Cross-Reference.** Section 3.3 provides additional context.

Speaker	Video URL	Accessed	Avail.
Abbot	https://youtu.be/uRZdTKxeAvE	*Aug. 15, 2020*	✓
Amthor	https://youtu.be/YeIZ7IcrNf4	*Aug. 15, 2020*	✓
Christmann	https://youtu.be/TfKzfuFZAqM	*Aug. 19, 2020*	✗
Brugger	https://youtu.be/FbpY5a4x3y4	*Aug. 15, 2020*	✗
Strasser	https://youtu.be/FD9ABAxVPSg	*Aug. 10, 2020*	✓
Högel	https://youtu.be/CehjnXths1M	*Aug. 19, 2020*	✓
Fauci	https://youtu.be/Vs7H-uNWifo	*Aug. 15, 2020*	✓
Gastel	https://youtu.be/xsYk_g2mdAM	*Aug. 19, 2020*	✗
Grundmann	https://youtu.be/ykqhOnQEfC4	*Aug. 15, 2020*	✓
Gysi	https://youtu.be/fXtjbwRIrdg	*Aug. 10, 2020*	✓
Hahn	https://youtu.be/G7Z1-y_I6JY	*Aug. 10, 2020*	✓
Andreae	https://youtu.be/MFkZsjodGDI	*Aug. 19, 2020*	✗

. . . continued

Speaker	Video URL	Accessed	Avail.
Künast	https://youtu.be/2h9V3YsERv8	*Aug. 10, 2020*	✗
Kurz	https://youtu.be/6U1VrLp_Hrk	*Aug. 19, 2020*	✓
Lindner	https://youtu.be/BkUxh91C9EU	*Aug. 10, 2020*	✗
Obama, M.	https://youtu.be/wWfvK-2JUqc	*Aug. 19, 2020*	✓
Neu	https://youtu.be/6DzAzrhc9RA	*Aug. 10, 2020*	✓
Özdemir	https://youtu.be/Y9y_t6ukki0	*Aug. 10, 2020*	✓
Ott	https://youtu.be/itgfOh_B46I	*Aug.15, 2020*	✓
Peterka	https://youtu.be/eHZpf1BdWzs	*Aug. 10, 2020*	✓
Ruppert	https://youtu.be/dRRyLWn8DAs	*Aug. 15, 2020*	✓
Saathoff	https://youtu.be/TnqoypdcetI	*Aug. 10, 2020*	✓
Sanders	https://youtu.be/xzfG7zApLT0	*Aug. 19, 2020*	✓
Sarrazin	https://youtu.be/4Opn3Cw3L9M	*Aug. 19, 2020*	✗
Sattelberger	https://youtu.be/Q-bnwpMIDAM	*Aug. 15, 2020*	✓
Schnieder	https://youtu.be/P12bsPda3vc	*Aug. 15, 2020*	✓
Theuer	https://youtu.be/rwzScawYuFk	*Aug. 19, 2020*	✓
Kartes	https://youtu.be/m2iCDQghhk0	*Aug. 15, 2020*	✓
Völlers	https://youtu.be/4HC8db2-94I	*Aug. 19, 2020*	✓
v.d. Leyen	https://youtu.be/rXEqI3pqF6g	*Aug. 10, 2020*	✓

Appendix

B

Data sets

B.1 Argument Strength Values of Hotel Reviews

B.1.1 Raw Data

- **Description.** This data set captures the perceived strength of the 43 arguments of the *SemEval-2015 Task 12* Test Data set (Pontiki et al., 2015) assessed by 105 participants using a 5-point Likert scale. This appendix provides an overview of the raw data for each participant and argument.
- **Data.** The `argument's id` (column 1), `rated strength` for arguments 1 - 22 (column 2 - 23, first half), `rated strength` for arguments 23 - 43 (column 2 - 23, second half).
- **Source.** This data was collected through an online crowd-sourcing survey.
- **Cross-Reference.** Section 5.2.3 provides additional context and analysis.

	Argument																					
ID	*1*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	*22*
1	*3*	*3*	*1*	*2*	*3*	*2*	*2*	*3*	*1*	*3*	*4*	*3*	*1*	*3*	*3*	*4*	*2*	*1*	*3*	*3*	*2*	*3*
2	*4*	*4*	*4*	*3*	*3*	*4*	*3*	*4*	*2*	*4*	*4*	*4*	*2*	*1*	*2*	*4*	*3*	*4*	*2*	*2*	*2*	*2*
3	*2*	*1*	*2*	*0*	*2*	*2*	*2*	*3*	*0*	*3*	*0*	*3*	*1*	*1*	*1*	*4*	*1*	*2*	*3*	*1*	*3*	*3*
4	*4*	*4*	*3*	*3*	*2*	*3*	*2*	*2*	*3*	*3*	*3*	*3*	*2*	*3*	*3*	*3*	*3*	*2*	*2*	*1*	*4*	*4*

. . . continued

	Argument																					
ID	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	22
5	0	0	3	4	4	4	3	3	1	3	1	2	1	1	1	1	3	2	3	2	2	2
6	3	4	4	4	4	3	4	4	0	4	2	4	1	3	3	4	4	3	4	3	4	4
7	0	1	3	3	3	3	3	4	2	3	2	0	0	1	1	0	3	3	3	3	0	1
8	4	4	4	4	4	4	4	4	0	4	3	4	3	4	2	4	4	4	4	4	4	4
9	2	1	3	3	3	3	3	3	3	3	1	2	1	3	2	2	4	2	3	2	4	3
10	3	3	2	3	3	2	4	3	2	3	0	4	2	4	3	2	3	2	3	2	2	2
11	2	2	2	2	2	2	2	3	3	3	2	3	3	3	3	3	3	3	3	3	2	3
12	3	3	2	2	3	2	2	3	1	2	3	4	3	0	2	3	2	2	3	2	2	2
13	4	4	2	3	3	4	3	4	0	2	4	4	3	4	4	4	2	3	4	3	3	3
14	1	1	1	1	2	1	1	2	3	1	2	2	1	2	2	1	1	1	3	1	2	2
15	0	0	4	4	4	4	4	4	0	4	4	4	0	0	0	0	4	4	4	4	0	0
16	3	3	3	3	1	2	2	3	1	3	3	3	4	1	3	3	1	1	2	2	4	2
17	1	1	2	3	3	2	2	3	0	3	0	4	4	4	1	1	4	3	0	3	1	3
18	0	0	3	3	2	3	3	3	2	3	2	0	1	2	1	0	2	4	2	2	2	1
19	0	0	4	4	4	4	4	4	2	4	4	0	0	0	1	1	4	4	4	3	1	1
20	4	3	0	0	0	0	0	0	2	0	2	2	3	3	2	3	1	1	0	1	2	3
21	4	3	4	4	4	4	4	4	0	4	3	3	4	4	2	2	4	3	4	4	2	4
22	4	3	4	3	3	3	3	3	1	2	2	3	3	4	2	4	2	4	2	1	2	3
23	4	4	2	4	4	4	4	3	0	1	4	3	2	0	2	4	3	2	4	1	1	1
24	1	2	3	4	3	3	4	3	2	3	3	2	2	0	2	3	2	3	3	2	2	1
25	0	0	4	4	4	4	4	4	0	4	0	0	0	0	0	0	4	4	4	4	0	0
26	0	0	4	4	4	4	4	4	1	4	3	0	0	0	1	0	2	3	4	3	0	0
27	1	1	2	1	1	1	1	3	1	2	2	3	2	1	2	1	3	1	3	3	2	3
28	2	4	4	4	4	4	4	4	1	4	3	2	0	1	2	2	2	3	4	0	0	0
29	2	1	2	2	2	2	2	2	1	2	2	3	2	1	3	2	2	2	2	3	3	1
30	4	2	3	4	1	4	4	3	0	4	3	4	2	4	2	4	4	4	3	1	1	4
31	2	3	2	4	2	3	2	3	2	3	1	0	2	3	2	3	3	3	4	2	3	2
32	2	4	4	3	2	4	2	4	3	4	3	4	3	3	3	4	4	4	4	2	2	4
33	4	3	2	3	3	2	3	3	1	3	3	4	0	0	3	4	3	2	4	2	0	2
34	2	1	4	4	4	4	4	4	3	3	3	2	1	1	3	2	3	3	2	2	3	2
35	0	0	3	2	4	3	3	3	2	3	1	1	1	0	2	1	3	3	3	2	1	1
36	0	0	4	4	4	4	3	4	3	4	0	0	0	2	2	0	3	4	4	3	1	0
37	0	0	4	0	0	0	4	4	4	4	4	4	0	4	4	4	4	4	4	0	0	4

. . . continued

	Argument																					
ID	*1*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	*22*
38	*3*	*4*	*0*	*0*	*0*	*1*	*1*	*0*	*2*	*0*	*4*	*4*	*4*	*1*	*2*	*3*	*0*	*0*	*0*	*0*	*2*	*4*
39	*0*	*0*	*4*	*4*	*4*	*4*	*4*	*4*	*0*	*4*	*0*	*0*	*0*	*0*	*0*	*0*	*4*	*4*	*4*	*4*	*0*	*0*
40	*2*	*3*	*2*	*2*	*2*	*2*	*3*	*3*	*0*	*3*	*2*	*3*	*0*	*0*	*3*	*3*	*3*	*2*	*3*	*3*	*1*	*1*
41	*4*	*4*	*3*	*4*	*2*	*4*	*3*	*4*	*1*	*4*	*2*	*4*	*4*	*4*	*2*	*3*	*2*	*4*	*4*	*0*	*2*	*2*
42	*3*	*2*	*2*	*1*	*3*	*2*	*3*	*4*	*1*	*3*	*2*	*1*	*3*	*0*	*3*	*3*	*1*	*3*	*1*	*2*	*2*	*2*
43	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*2*	*4*	*4*	*3*	*3*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*
44	*0*	*1*	*1*	*1*	*1*	*0*	*0*	*3*	*2*	*3*	*3*	*3*	*1*	*4*	*2*	*2*	*1*	*2*	*1*	*2*	*3*	*2*
45	*4*	*4*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*1*	*4*	*4*	*0*	*4*	*4*	*0*	*0*	*0*	*0*	*0*	*0*
46	*1*	*0*	*1*	*1*	*1*	*0*	*1*	*1*	*3*	*2*	*3*	*3*	*3*	*4*	*1*	*1*	*1*	*2*	*2*	*2*	*3*	*0*
47	*0*	*0*	*4*	*4*	*4*	*4*	*4*	*4*	*0*	*4*	*0*	*0*	*0*	*4*	*0*	*0*	*4*	*4*	*4*	*4*	*4*	*4*
48	*1*	*1*	*3*	*3*	*2*	*3*	*2*	*2*	*2*	*3*	*3*	*2*	*2*	*0*	*3*	*2*	*3*	*3*	*3*	*2*	*0*	*2*
49	*3*	*1*	*1*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*1*	*2*	*2*	*3*	*2*	*2*	*2*	*1*	*2*	*2*	*2*	*2*
50	*0*	*1*	*1*	*0*	*0*	*1*	*1*	*1*	*2*	*2*	*2*	*2*	*1*	*4*	*1*	*1*	*2*	*3*	*2*	*2*	*4*	*0*
51	*3*	*3*	*3*	*2*	*2*	*3*	*2*	*3*	*2*	*2*	*3*	*2*	*2*	*3*	*2*	*3*	*2*	*3*	*2*	*2*	*2*	*2*
52	*3*	*2*	*2*	*4*	*3*	*3*	*3*	*3*	*2*	*2*	*2*	*4*	*2*	*3*	*3*	*3*	*3*	*2*	*3*	*1*	*4*	*4*
53	*3*	*3*	*4*	*3*	*4*	*4*	*3*	*4*	*3*	*4*	*2*	*3*	*3*	*1*	*3*	*3*	*3*	*3*	*4*	*2*	*2*	*2*
54	*2*	*1*	*2*	*3*	*3*	*2*	*4*	*4*	*0*	*4*	*2*	*1*	*1*	*1*	*2*	*3*	*3*	*4*	*3*	*1*	*2*	*1*
55	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*2*	*3*	*0*	*2*	*4*	*2*	*4*	*0*	*1*	*1*	*1*	*2*	*2*	*3*	*0*
56	*4*	*4*	*3*	*3*	*3*	*3*	*2*	*4*	*2*	*2*	*3*	*4*	*3*	*3*	*3*	*4*	*2*	*2*	*3*	*3*	*2*	*3*
57	*3*	*3*	*1*	*1*	*1*	*0*	*1*	*1*	*2*	*1*	*3*	*4*	*3*	*4*	*1*	*2*	*1*	*1*	*0*	*2*	*3*	*3*
58	*4*	*4*	*1*	*1*	*1*	*1*	*1*	*1*	*2*	*1*	*3*	*3*	*3*	*2*	*2*	*2*	*1*	*2*	*2*	*1*	*1*	*1*
59	*1*	*1*	*2*	*3*	*3*	*1*	*4*	*4*	*4*	*3*	*1*	*4*	*4*	*0*	*2*	*4*	*3*	*3*	*4*	*4*	*0*	*3*
60	*0*	*0*	*3*	*4*	*4*	*4*	*4*	*4*	*1*	*4*	*0*	*0*	*0*	*1*	*1*	*0*	*3*	*4*	*0*	*4*	*3*	*0*
61	*2*	*3*	*3*	*2*	*3*	*2*	*3*	*4*	*1*	*3*	*2*	*3*	*1*	*4*	*2*	*3*	*3*	*3*	*2*	*3*	*3*	*3*
62	*0*	*0*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*0*	*0*	*0*	*4*	*0*	*0*	*4*	*4*	*1*	*0*	*0*	*0*
63	*1*	*4*	*1*	*3*	*2*	*3*	*1*	*3*	*1*	*3*	*1*	*0*	*3*	*4*	*0*	*1*	*2*	*3*	*2*	*2*	*3*	*3*
64	*3*	*3*	*1*	*1*	*1*	*1*	*0*	*1*	*1*	*1*	*3*	*3*	*3*	*4*	*0*	*3*	*1*	*0*	*1*	*2*	*3*	*3*
65	*1*	*1*	*3*	*1*	*3*	*1*	*2*	*4*	*3*	*3*	*3*	*4*	*3*	*1*	*4*	*2*	*3*	*2*	*2*	*3*	*2*	*3*
66	*1*	*1*	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*1*	*1*	*1*	*1*	*1*	*1*	*3*	*3*	*0*	*2*	*2*	*1*
67	*3*	*2*	*3*	*1*	*1*	*1*	*1*	*3*	*2*	*0*	*2*	*3*	*2*	*4*	*1*	*3*	*3*	*2*	*4*	*1*	*0*	*0*
68	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*2*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*2*	*4*	*4*
69	*1*	*1*	*4*	*3*	*2*	*3*	*3*	*3*	*3*	*2*	*2*	*0*	*1*	*0*	*2*	*1*	*3*	*3*	*4*	*2*	*2*	*1*
70	*1*	*1*	*3*	*3*	*3*	*3*	*4*	*3*	*1*	*3*	*1*	*1*	*1*	*0*	*2*	*1*	*3*	*3*	*3*	*2*	*1*	*1*

. . . continued

	Argument																					
ID	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	22
71	4	3	3	4	3	2	4	4	0	4	4	4	2	0	4	4	4	4	4	0	0	0
72	4	4	3	4	4	3	3	4	0	3	0	4	4	0	2	3	1	3	4	0	0	0
73	4	4	2	3	3	3	4	4	3	3	2	4	3	3	4	4	4	3	3	2	2	3
74	3	4	2	2	3	3	3	3	2	2	4	4	2	2	3	4	3	3	3	3	3	3
75	3	4	3	4	3	0	3	4	2	3	1	4	1	3	1	3	2	3	3	0	3	2
76	1	1	4	1	3	2	3	4	2	3	2	3	1	0	1	2	4	2	4	3	2	2
77	0	0	3	4	4	3	4	4	2	3	0	0	0	0	0	0	3	2	4	2	1	0
78	4	4	0	0	0	0	0	0	0	0	3	4	3	0	3	4	0	0	0	1	3	3
79	1	2	1	1	2	2	2	3	0	2	3	4	0	0	1	1	3	2	2	3	1	2
80	4	4	4	4	4	4	3	4	4	3	3	4	4	4	4	4	4	2	4	4	2	3
81	2	4	3	3	3	2	3	3	3	2	1	4	3	3	2	2	4	4	2	2	4	3
82	0	0	4	4	4	4	4	4	2	4	0	0	0	0	0	0	4	4	4	3	2	2
83	2	1	4	3	3	3	4	4	3	4	2	2	0	4	2	3	4	4	4	2	2	2
84	4	4	3	4	3	4	3	4	3	3	2	3	3	4	2	2	3	3	3	3	2	3
85	4	3	3	3	4	3	4	4	2	4	3	4	3	3	4	3	2	3	4	1	2	2
86	4	4	4	4	3	4	4	4	1	4	3	4	3	4	1	4	4	4	4	4	1	0
87	4	4	3	4	4	3	4	3	0	4	4	1	1	0	0	4	4	4	4	4	4	4
88	3	3	4	3	3	2	2	3	2	2	2	3	4	3	2	3	2	2	4	2	3	4
89	4	4	0	0	0	0	0	0	2	0	2	4	3	4	2	4	1	0	1	0	3	3
90	2	3	1	1	0	0	1	0	2	0	2	4	2	4	2	3	1	0	2	1	4	4
91	2	2	2	3	2	2	2	3	1	3	2	3	3	3	3	2	2	2	2	2	3	3
92	3	2	2	2	3	2	2	2	3	2	2	2	2	4	2	2	2	2	2	2	3	3
93	0	0	4	4	4	4	4	4	0	4	4	0	0	0	0	0	4	4	4	4	0	0
94	3	3	4	4	2	3	2	3	1	2	3	3	2	4	1	2	3	3	4	2	4	4
95	4	3	4	3	1	2	3	4	3	3	3	3	3	4	3	4	3	2	2	4	4	4
96	3	2	2	2	2	2	1	3	2	2	3	4	2	4	3	4	3	2	3	2	3	4
97	0	0	4	3	4	4	4	4	1	4	4	1	0	2	0	2	2	4	3	2	0	1
98	4	4	4	3	3	3	4	4	1	4	3	4	2	1	1	3	3	4	4	3	2	3
99	3	2	1	0	3	0	3	3	0	2	3	4	2	3	1	3	3	1	3	1	4	4
100	3	2	4	1	3	2	4	3	2	4	2	4	3	4	2	2	2	2	4	1	4	0
101	2	3	2	2	4	2	4	3	1	4	3	4	3	0	2	3	3	2	4	3	2	2
102	4	4	0	0	1	2	1	0	1	0	1	2	0	0	3	4	2	0	2	2	4	4
103	4	3	3	3	4	4	4	2	2	2	1	2	2	4	3	4	3	2	4	2	3	3

. . . continued

	Argument																				
ID	*1*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	*22*
104	*4*	*3*	*3*	*3*	*2*	*3*	*3*	*4*	*3*	*3*	*4*	*4*	*2*	*4*	*4*	*4*	*3*	*2*	*4*	*4*	*4*	*4*
105	*4*	*4*	*4*	*1*	*3*	*3*	*3*	*4*	*2*	*3*	*3*	*2*	*1*	*3*	*2*	*4*	*4*	*3*	*4*	*2*	*1*	*3*

	Argument																				
ID	*23*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	*43*
1	*3*	*3*	*4*	*2*	*4*	*2*	*4*	*3*	*4*	*2*	*2*	*2*	*3*	*3*	*3*	*3*	*2*	*2*	*3*	*4*	*3*
2	*2*	*2*	*4*	*3*	*4*	*4*	*2*	*2*	*2*	*0*	*2*	*3*	*3*	*3*	*3*	*3*	*3*	*4*	*4*	*4*	*3*
3	*3*	*4*	*3*	*1*	*2*	*3*	*0*	*0*	*0*	*1*	*2*	*1*	*1*	*2*	*2*	*3*	*0*	*3*	*2*	*3*	*2*
4	*4*	*3*	*3*	*3*	*3*	*3*	*2*	*0*	*0*	*1*	*2*	*3*	*3*	*2*	*3*	*3*	*4*	*2*	*2*	*2*	*2*
5	*2*	*0*	*2*	*3*	*1*	*3*	*1*	*2*	*1*	*1*	*2*	*3*	*3*	*3*	*3*	*4*	*3*	*4*	*4*	*2*	*1*
6	*4*	*2*	*3*	*1*	*3*	*4*	*1*	*3*	*1*	*0*	*3*	*3*	*4*	*4*	*4*	*3*	*4*	*4*	*4*	*4*	*4*
7	*1*	*2*	*3*	*3*	*1*	*1*	*1*	*1*	*1*	*1*	*3*	*2*	*4*	*3*	*3*	*3*	*3*	*4*	*3*	*4*	*2*
8	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*
9	*3*	*2*	*3*	*4*	*2*	*3*	*3*	*3*	*3*	*3*	*3*	*2*	*3*	*3*	*3*	*3*	*3*	*2*	*3*	*3*	*2*
10	*3*	*3*	*1*	*3*	*4*	*4*	*2*	*0*	*0*	*1*	*2*	*2*	*3*	*4*	*3*	*4*	*3*	*3*	*4*	*2*	*4*
11	*3*	*4*	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*2*	*3*	*3*	*2*	*2*	*2*	*3*	*3*	*3*	*2*
12	*2*	*3*	*2*	*2*	*3*	*2*	*3*	*2*	*2*	*0*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*4*
13	*3*	*4*	*4*	*0*	*4*	*4*	*0*	*1*	*1*	*0*	*1*	*4*	*3*	*1*	*4*	*4*	*3*	*3*	*1*	*4*	*4*
14	*2*	*2*	*2*	*2*	*2*	*2*	*3*	*2*	*2*	*2*	*2*	*1*	*1*	*2*	*1*	*2*	*1*	*1*	*1*	*2*	*1*
15	*4*	*0*	*4*	*4*	*0*	*4*	*0*	*0*	*0*	*0*	*4*	*4*	*0*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*
16	*3*	*3*	*1*	*4*	*3*	*3*	*3*	*2*	*2*	*0*	*1*	*2*	*2*	*2*	*3*	*3*	*2*	*0*	*2*	*3*	*2*
17	*3*	*1*	*4*	*0*	*3*	*3*	*1*	*2*	*2*	*2*	*3*	*2*	*2*	*3*	*3*	*4*	*4*	*3*	*0*	*3*	*3*
18	*1*	*0*	*0*	*2*	*1*	*4*	*0*	*0*	*0*	*2*	*2*	*3*	*3*	*3*	*4*	*2*	*3*	*4*	*3*	*3*	*1*
19	*0*	*0*	*1*	*4*	*0*	*4*	*0*	*0*	*0*	*0*	*0*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*0*
20	*3*	*4*	*2*	*1*	*3*	*0*	*4*	*2*	*2*	*0*	*1*	*1*	*1*	*0*	*0*	*0*	*0*	*1*	*0*	*0*	*2*
21	*4*	*2*	*2*	*0*	*4*	*4*	*1*	*1*	*1*	*0*	*3*	*3*	*4*	*3*	*4*	*4*	*4*	*4*	*4*	*4*	*2*
22	*3*	*2*	*1*	*3*	*4*	*3*	*4*	*1*	*1*	*2*	*2*	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*2*
23	*1*	*3*	*3*	*1*	*4*	*3*	*3*	*3*	*3*	*4*	*3*	*2*	*1*	*2*	*2*	*3*	*3*	*4*	*4*	*2*	*2*
24	*2*	*1*	*3*	*2*	*4*	*3*	*1*	*1*	*1*	*0*	*3*	*2*	*4*	*3*	*3*	*4*	*4*	*3*	*3*	*4*	*3*
25	*0*	*0*	*4*	*4*	*0*	*4*	*0*	*0*	*0*	*0*	*0*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*0*
26	*0*	*0*	*0*	*3*	*0*	*3*	*4*	*0*	*0*	*2*	*4*	*4*	*4*	*4*	*3*	*4*	*4*	*3*	*3*	*4*	*0*
27	*3*	*2*	*3*	*3*	*2*	*2*	*3*	*2*	*2*	*2*	*3*	*1*	*3*	*2*	*1*	*1*	*1*	*2*	*3*	*3*	*2*

. . . continued

	Argument																				
ID	23	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	43
28	2	2	0	0	4	4	1	0	0	0	2	4	2	4	4	4	1	2	2	4	3
29	2	2	2	1	3	2	2	3	3	1	3	2	3	2	2	2	2	2	2	3	2
30	4	4	0	2	3	4	3	2	2	0	4	1	2	2	3	4	4	4	4	2	1
31	3	3	3	2	3	3	3	2	2	2	2	2	2	2	2	2	4	2	2	4	1
32	4	2	3	3	4	3	2	3	4	1	3	2	4	3	4	3	4	4	4	4	4
33	1	1	3	3	3	3	2	3	3	3	2	3	3	3	4	3	3	2	4	2	4
34	3	2	3	3	2	3	3	2	3	3	2	4	1	2	4	4	4	3	4	4	2
35	1	1	1	3	0	4	0	2	1	1	3	2	2	3	3	4	4	3	3	3	1
36	1	1	0	2	0	4	0	0	0	2	1	3	2	3	4	4	3	4	4	3	1
37	0	4	4	0	4	4	4	0	0	0	4	0	4	4	4	4	0	4	4	4	4
38	4	1	2	0	3	0	4	4	4	2	4	2	1	0	0	0	0	0	0	4	4
39	4	0	0	4	0	4	4	0	0	0	4	4	4	4	4	4	4	4	4	4	0
40	1	2	3	2	3	2	3	1	1	2	3	2	2	3	2	2	2	2	2	3	1
41	2	2	1	2	4	4	2	1	1	0	3	2	3	4	4	4	4	4	4	2	1
42	2	4	0	0	2	3	3	0	0	1	2	2	4	4	3	4	1	3	4	3	3
43	4	4	4	4	4	4	4	4	4	2	4	4	4	4	4	4	4	4	4	4	4
44	3	3	3	2	2	4	3	4	2	3	3	1	3	1	1	1	1	1	4	2	2
45	0	4	4	0	4	0	0	4	4	0	0	0	0	0	0	0	0	0	0	0	4
46	3	4	3	3	3	3	4	4	3	3	3	1	3	2	2	2	2	2	3	3	1
47	4	4	0	4	0	4	0	0	0	0	4	4	4	4	4	4	4	4	4	4	0
48	1	2	2	2	2	2	2	2	2	0	2	4	3	3	3	3	2	2	2	3	2
49	2	1	2	2	1	2	2	2	2	3	2	2	2	2	3	2	3	1	2	2	2
50	3	4	4	3	1	1	2	3	4	3	3	1	2	1	2	1	2	1	2	3	1
51	3	2	3	2	3	3	2	3	3	2	3	2	2	2	3	2	2	3	3	2	1
52	3	2	2	3	3	2	2	3	3	1	1	3	4	2	3	2	3	3	4	3	1
53	2	2	3	2	3	3	3	2	3	2	3	3	3	3	3	4	3	3	3	4	2
54	1	0	1	4	3	4	3	0	0	0	1	1	3	4	3	4	4	3	4	4	4
55	4	3	4	4	1	1	4	2	3	3	2	0	1	0	0	0	1	1	1	1	1
56	3	4	2	3	4	3	3	4	4	2	3	2	2	2	2	4	4	3	3	3	3
57	3	2	2	1	3	1	3	3	2	1	3	1	1	2	1	1	1	1	1	0	4
58	2	3	2	2	2	1	2	2	2	2	2	2	1	1	1	1	1	2	2	1	2
59	0	3	4	4	2	3	4	4	4	4	4	2	1	4	3	3	2	4	2	3	1
60	1	0	4	4	0	4	0	0	0	0	4	4	4	3	4	4	4	4	4	4	0
61	3	3	3	2	3	3	2	1	1	2	3	2	2	3	3	3	2	3	3	3	2

. . . continued

	Argument																				
ID	*23*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	*43*
62	*0*	*0*	*4*	*4*	*0*	*0*	*0*	*0*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*0*
63	*2*	*1*	*4*	*3*	*4*	*3*	*2*	*3*	*3*	*2*	*2*	*2*	*1*	*2*	*3*	*3*	*4*	*2*	*1*	*3*	*2*
64	*3*	*2*	*2*	*1*	*3*	*1*	*3*	*3*	*3*	*1*	*3*	*1*	*0*	*2*	*0*	*1*	*1*	*0*	*1*	*1*	*3*
65	*3*	*3*	*4*	*3*	*3*	*3*	*3*	*2*	*2*	*3*	*4*	*1*	*2*	*4*	*2*	*3*	*2*	*3*	*3*	*4*	*1*
66	*2*	*1*	*1*	*3*	*1*	*3*	*1*	*0*	*0*	*1*	*3*	*2*	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*1*
67	*0*	*1*	*0*	*2*	*4*	*2*	*4*	*2*	*2*	*0*	*2*	*0*	*0*	*2*	*1*	*1*	*1*	*2*	*1*	*2*	*1*
68	*2*	*2*	*2*	*2*	*4*	*4*	*4*	*4*	*4*	*2*	*4*	*2*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*
69	*1*	*0*	*1*	*3*	*0*	*4*	*2*	*2*	*2*	*2*	*4*	*2*	*2*	*4*	*3*	*3*	*3*	*3*	*3*	*3*	*1*
70	*1*	*1*	*1*	*3*	*1*	*3*	*1*	*1*	*1*	*2*	*2*	*4*	*3*	*4*	*4*	*4*	*4*	*3*	*3*	*3*	*2*
71	*0*	*4*	*3*	*2*	*4*	*4*	*2*	*2*	*2*	*0*	*0*	*2*	*2*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*
72	*0*	*3*	*2*	*0*	*4*	*4*	*4*	*2*	*2*	*0*	*3*	*0*	*3*	*4*	*3*	*4*	*4*	*3*	*4*	*4*	*3*
73	*4*	*4*	*3*	*3*	*3*	*4*	*3*	*3*	*3*	*2*	*2*	*2*	*3*	*4*	*4*	*3*	*3*	*3*	*4*	*3*	*3*
74	*3*	*4*	*4*	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*2*	*2*	*3*	*3*	*4*	*2*	*2*	*3*	*3*	*3*
75	*2*	*3*	*2*	*2*	*4*	*3*	*4*	*1*	*1*	*0*	*1*	*2*	*3*	*3*	*2*	*0*	*4*	*2*	*1*	*4*	*3*
76	*3*	*2*	*3*	*3*	*3*	*3*	*2*	*2*	*2*	*1*	*2*	*2*	*3*	*3*	*3*	*3*	*2*	*2*	*3*	*4*	*3*
77	*0*	*0*	*0*	*2*	*0*	*4*	*0*	*0*	*0*	*2*	*3*	*2*	*2*	*2*	*4*	*4*	*4*	*4*	*4*	*3*	*1*
78	*3*	*4*	*3*	*0*	*4*	*0*	*1*	*3*	*3*	*1*	*3*	*2*	*2*	*0*	*0*	*0*	*0*	*0*	*1*	*0*	*4*
79	*1*	*3*	*0*	*3*	*1*	*2*	*0*	*0*	*0*	*1*	*2*	*2*	*1*	*3*	*2*	*2*	*1*	*1*	*2*	*2*	*1*
80	*3*	*4*	*4*	*4*	*4*	*4*	*4*	*3*	*3*	*4*	*3*	*3*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*3*	*3*
81	*4*	*3*	*3*	*2*	*3*	*4*	*2*	*3*	*2*	*2*	*3*	*3*	*3*	*3*	*4*	*3*	*4*	*3*	*2*	*4*	*3*
82	*2*	*0*	*3*	*2*	*0*	*4*	*2*	*0*	*0*	*3*	*2*	*2*	*4*	*3*	*4*	*4*	*4*	*4*	*4*	*4*	*4*
83	*3*	*3*	*2*	*3*	*1*	*4*	*3*	*0*	*0*	*2*	*2*	*3*	*4*	*4*	*4*	*4*	*3*	*4*	*3*	*4*	*0*
84	*2*	*2*	*3*	*4*	*3*	*3*	*3*	*2*	*2*	*3*	*2*	*3*	*3*	*4*	*3*	*4*	*4*	*4*	*3*	*4*	*3*
85	*3*	*4*	*0*	*4*	*4*	*4*	*3*	*3*	*3*	*2*	*4*	*4*	*3*	*3*	*4*	*4*	*4*	*4*	*3*	*4*	*4*
86	*0*	*1*	*4*	*4*	*3*	*4*	*2*	*3*	*3*	*2*	*4*	*3*	*4*	*4*	*4*	*4*	*4*	*4*	*3*	*4*	*0*
87	*0*	*4*	*0*	*3*	*4*	*4*	*4*	*0*	*0*	*0*	*1*	*3*	*4*	*4*	*3*	*4*	*4*	*4*	*4*	*4*	*3*
88	*4*	*2*	*2*	*2*	*4*	*2*	*4*	*3*	*1*	*2*	*2*	*2*	*2*	*2*	*2*	*4*	*3*	*2*	*3*	*4*	*2*
89	*3*	*3*	*2*	*0*	*4*	*0*	*0*	*2*	*2*	*2*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*3*
90	*4*	*4*	*3*	*0*	*4*	*0*	*4*	*3*	*3*	*1*	*1*	*1*	*1*	*0*	*0*	*0*	*0*	*0*	*1*	*0*	*4*
91	*3*	*3*	*2*	*2*	*2*	*2*	*3*	*1*	*1*	*2*	*2*	*2*	*2*	*2*	*2*	*3*	*3*	*2*	*2*	*2*	*2*
92	*2*	*2*	*3*	*2*	*2*	*2*	*2*	*1*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*3*	*2*	*4*	*2*	*3*	*3*
93	*0*	*0*	*4*	*4*	*0*	*4*	*0*	*0*	*0*	*0*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*0*
94	*4*	*1*	*4*	*3*	*4*	*3*	*2*	*2*	*2*	*1*	*1*	*2*	*2*	*2*	*3*	*2*	*4*	*3*	*2*	*4*	*2*
95	*4*	*4*	*3*	*1*	*4*	*4*	*4*	*3*	*3*	*1*	*3*	*2*	*2*	*4*	*3*	*3*	*4*	*2*	*3*	*4*	*2*

. . . continued

	Argument																				
ID	*23*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	*43*
96	*4*	*4*	*2*	*2*	*4*	*2*	*2*	*4*	*4*	*1*	*2*	*2*	*2*	*2*	*2*	*3*	*2*	*2*	*2*	*3*	*3*
97	*1*	*0*	*0*	*4*	*0*	*4*	*0*	*4*	*3*	*3*	*3*	*3*	*4*	*4*	*3*	*4*	*3*	*4*	*3*	*4*	*1*
98	*3*	*3*	*1*	*1*	*3*	*4*	*1*	*2*	*2*	*1*	*3*	*2*	*3*	*3*	*3*	*3*	*3*	*4*	*4*	*4*	*3*
99	*4*	*3*	*3*	*3*	*4*	*3*	*1*	*3*	*3*	*0*	*4*	*2*	*0*	*3*	*3*	*3*	*0*	*1*	*3*	*3*	*0*
100	*3*	*3*	*3*	*2*	*4*	*4*	*4*	*3*	*3*	*3*	*2*	*3*	*3*	*4*	*4*	*4*	*1*	*2*	*3*	*2*	*3*
101	*2*	*3*	*4*	*3*	*4*	*3*	*3*	*1*	*1*	*1*	*1*	*1*	*3*	*4*	*3*	*4*	*3*	*4*	*4*	*3*	*2*
102	*4*	*0*	*2*	*2*	*4*	*4*	*4*	*2*	*2*	*0*	*2*	*1*	*1*	*2*	*1*	*3*	*0*	*0*	*2*	*0*	*1*
103	*4*	*2*	*4*	*2*	*4*	*3*	*2*	*2*	*2*	*0*	*3*	*2*	*4*	*3*	*3*	*3*	*2*	*4*	*3*	*3*	*4*
104	*4*	*4*	*4*	*3*	*4*	*3*	*3*	*1*	*1*	*3*	*2*	*2*	*3*	*3*	*3*	*3*	*3*	*4*	*4*	*4*	*3*
105	*2*	*3*	*3*	*2*	*4*	*3*	*2*	*2*	*2*	*1*	*2*	*2*	*4*	*2*	*2*	*3*	*2*	*3*	*2*	*3*	*2*

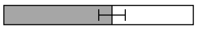
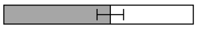
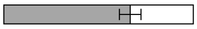
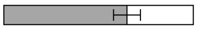
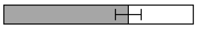
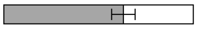
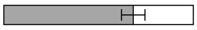
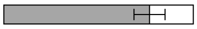
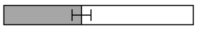
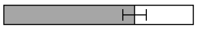
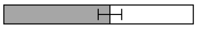
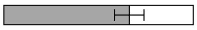
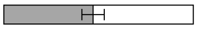
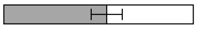
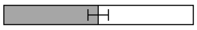
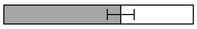
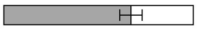
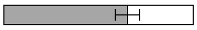
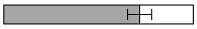
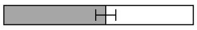
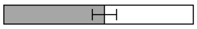
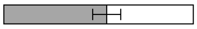
B.1.2 Statistics

- **Description.** This data set captures the perceived strength of the 43 arguments of the *SemEval-2015 Task 12* Test Data set (Pontiki et al., 2015) assessed by 105 participants using a 5-point Likert scale. This appendix summarizes the statistics for each argument.

- **Data.** The argument's id (column 1) raw and normalized means μ (columns $2+3$), the normalized standard deviation σ (column 3), the 95% confidence interval denoted by error bars (column 5), and the argument's stance (column 6).

- **Source.** This data was collected through an online crowd-sourcing survey.

- **Cross-Reference.** Section 5.2.3 provides additional context and analysis.

Argument	Mean μ	Normalized μ	StDev σ	Confidence	Stance
1	2.286	0.571	0.379	 0.571	-
2	2.248	0.562	0.375	 0.562	-
3	2.667	0.667	0.309	 0.667	+
4	2.600	0.650	0.337	 0.650	+
5	2.629	0.657	0.311	 0.657	+
6	2.524	0.631	0.325	 0.631	+
7	2.733	0.683	0.311	 0.683	+
8	3.076	0.769	0.287	 0.769	+
9	1.638	0.410	0.288	 0.410	-
10	2.762	0.690	0.299	 0.690	+
11	2.238	0.560	0.299	 0.560	-
12	2.648	0.662	0.360	 0.662	-
13	1.886	0.471	0.324	 0.471	-
14	2.171	0.543	0.414	 0.543	-
15	1.990	0.498	0.289	 0.498	-
16	2.467	0.617	0.342	 0.617	-
17	2.686	0.671	0.271	 0.671	+
18	2.610	0.652	0.291	 0.652	+
19	2.867	0.717	0.306	 0.717	+
20	2.152	0.538	0.285	 0.538	+
21	2.124	0.531	0.327	 0.531	-
22	2.171	0.543	0.346	 0.543	-

. . . continued

Argument	Mean μ	Normalized μ	StDev σ	Confidence	Stance
23	2.352	0.588	0.331	0.588	-
24	2.286	0.571	0.341	0.571	-
25	2.410	0.602	0.330	0.602	-
26	2.371	0.593	0.305	0.593	+
27	2.638	0.660	0.357	0.660	-
28	2.933	0.733	0.293	0.733	+
29	2.257	0.564	0.340	0.564	-
30	1.867	0.467	0.327	0.467	-
31	1.857	0.464	0.328	0.464	-
32	1.457	0.364	0.298	0.364	-
33	2.486	0.621	0.264	0.621	+
34	2.248	0.562	0.268	0.562	+
35	2.571	0.643	0.294	0.643	+
36	2.781	0.695	0.286	0.695	+
37	2.781	0.695	0.296	0.695	+
38	2.943	0.736	0.308	0.736	+
39	2.686	0.671	0.335	0.671	+
40	2.752	0.688	0.314	0.688	+
41	2.829	0.707	0.291	0.707	+
42	3.019	0.755	0.286	0.755	+
43	2.190	0.548	0.320	0.548	-

B.2 Persuasive Effectiveness

- **Description.** This data describes the collected data of the stance estimation model study conducted with 48 participants.

- **Data.** The participant's id (column 1), the positive feedback f^+ (column 2), the negative feedback f^- (column 3), the computed persuasive effectiveness z_{Φ_0} (column 4), whether or not the system's prediction match the user's decision (column 5) and the computed confidence (column 6).

- **Source.** The data was collected as part of a controlled experiment conducted in a laboratory setting.

- **Cross-Reference.** Section 5.2.4 provides additional context and analysis.

ID	f^+	f^-	Effectiveness z_{Φ_0}	Prediction	Confidence
1	0.42	0.42	0.55	-	0.62
2	0.47	0.35	0.65	-	0.82
3	0.49	0.35	0.65	✓	0.82
4	0.28	0.53	0.22	✓	0.94
5	0.26	0.49	0.30	✗	0.88
6	0.26	0.4	0.30	✗	0.88
7	0.44	0.44	0.48	-	0.55
8	0.49	0.35	0.61	✗	0.75
9	0.47	0.37	0.66	✓	0.83
10	0.26	0.4	0.33	✗	0.85
11	0.21	0.49	0.31	✓	0.87
12	0.35	0.44	0.66	-	0.83
13	0.49	0.33	0.68	✓	0.86
14	0.53	0.35	0.68	-	0.86
15	0.33	0.42	0.43	-	0.67
16	0.42	0.28	0.64	✓	0.80
17	0.23	0.56	0.23	-	0.94
18	0.28	0.42	0.37	-	0.79
19	0.63	0.12	0.83	✓	0.96
20	0.51	0.14	0.80	✓	0.95
21	0.56	0.16	0.79	✓	0.95

. . . continued

ID	f^+	f^-	Effectiveness z_{Φ_0}	Prediction	Confidence
22	0.16	0.6	0.20	✓	0.95
23	0.33	0.37	0.30	-	0.88
24	0.35	0.42	0.38	✗	0.77
25	0.35	0.4	0.48	-	0.55
26	0.37	0.3	0.58	✓	0.69
27	0.56	0.37	0.66	-	0.83
28	0.4	0.33	0.52	-	0.55
29	0.33	0.35	0.48	-	0.55
30	0.28	0.47	0.34	✗	0.83
31	0.47	0.21	0.64	✗	0.80
32	0.47	0.16	0.68	✓	0.86
33	0.4	0.23	0.53	-	0.57
34	0.3	0.6	0.30	✓	0.88
35	0.42	0.44	0.47	✓	0.57
36	0.44	0.35	0.58	✗	0.69
37	0.33	0.33	0.52	✗	0.55
38	0.3	0.47	0.38	✓	0.77
39	0.44	0.26	0.64	✗	0.80
40	0.32	0.53	0.37	✗	0.79
41	0.58	0.21	0.70	✗	0.88
42	0.16	0.47	0.25	✓	0.92
43	0.49	0.21	0.62	✗	0.77
44	0.3	0.49	0.36	✓	0.80
45	0.4	0.37	0.39	-	0.75
46	0.26	0.53	0.40	✗	0.73
47	0.42	0.26	0.58	✓	0.69
48	0.19	0.56	0.27	✓	0.91

B.3 Argument Visitation Quotient

B.3.1 First Study

- **Description.** This data set describes the collected data during the first study conducted with 84 participants to evaluate the effects of intervention strategies and embodiment condition on the computational metric (AVQ) and conversational user engagement (CE). The primary purpose of the first study was to filter out unnecessary conditions in preparation for the second study.

- **Data.** The `data set's id` (column 1), the `condition` (column 2) co-variate `gamification` (column 3), co-variate `intervention` (column 4), co-variate `embodiment` (column 5), the `computational metric` AVQ (column 6), the `number of system interventions` (column 7), and the `success rate` (column 8).

- **Source.** The data was collected through an online crowd-sourcing study.

- **Cross-Reference.** Section 7.1 provides additional context and analysis.

						interventions	
ID	*Group*	*Gamif.*	*Intervene*	*Embod.*	*AVQ*	#	*success*
1	*G0*	*no*	*no*	*no*	*0.92*	*0*	*0.00*
2	*G0*	*no*	*no*	*no*	*0.78*	*0*	*0.00*
3	*G0*	*no*	*no*	*no*	*0.94*	*0*	*0.00*
4	*G0*	*no*	*no*	*no*	*0.76*	*0*	*0.00*
5	*G0*	*no*	*no*	*no*	*0.81*	*0*	*0.00*
6	*G0*	*no*	*no*	*no*	*0.57*	*0*	*0.00*
7	*G1*	*no*	*yes*	*no*	*1*	*9*	*1.00*
8	*G1*	*no*	*yes*	*no*	*0.94*	*4*	*1.00*
9	*G1*	*no*	*yes*	*no*	*0.85*	*3*	*1.00*
10	*G1*	*no*	*yes*	*no*	*0.99*	*6*	*0.83*
11	*G1*	*no*	*yes*	*no*	*0.99*	*27*	*0.56*
12	*G1*	*no*	*yes*	*no*	*0.93*	*4*	*1.00*
13	*G2*	*yes*	*yes*	*no*	*0.98*	*4*	*1.00*
14	*G2*	*yes*	*yes*	*no*	*0.96*	*1*	*1.00*
15	*G2*	*yes*	*yes*	*no*	*0.95*	*5*	*0.60*
16	*G2*	*yes*	*yes*	*no*	*0.98*	*8*	*1.00*

. . . continued

						interventions	
ID	*Group*	*Gamif.*	*Intervene*	*Embod.*	*AVQ*	#	*success*
17	G2	yes	yes	no	0.95	5	1.00
18	G2	yes	yes	no	0.99	2	1.00
19	G2	yes	yes	no	0.83	2	1.00
20	G2	yes	yes	no	0.92	8	1.00
21	G2	yes	yes	no	0.86	3	0.67
22	G3	no	no	yes	0.98	0	0.00
23	G3	no	no	yes	0.84	0	0.00
24	G3	no	no	yes	0.81	0	0.00
25	G3	no	no	yes	0.95	0	0.00
26	G3	no	no	yes	0.74	0	0.00
27	G3	no	no	yes	0.92	0	0.00
28	G3	no	no	yes	0.84	0	0.00
29	G4	no	yes	yes	0.99	15	0.73
30	G4	no	yes	yes	0.98	4	1.00
31	G4	no	yes	yes	0.98	20	0.60
32	G4	no	yes	yes	0.98	1	1.00
33	G4	no	yes	yes	0.8	5	0.60
34	G4	no	yes	yes	0.96	7	0.86
35	G4	no	yes	yes	0.84	2	1.00
36	G4	no	yes	yes	0.98	15	1.00
37	G4	no	yes	yes	1	4	1.00
38	G4	no	yes	yes	0.87	7	0.86
39	G4	no	yes	yes	0.95	30	0.10
40	G4	no	yes	yes	0.93	28	0.11
41	G4	no	yes	yes	0.99	7	0.57
42	G5	yes	yes	yes	0.93	4	1.00
43	G5	yes	yes	yes	0.9	4	1.00
44	G5	yes	yes	yes	0.98	7	1.00
45	G5	yes	yes	yes	0.99	10	1.00
46	G5	yes	yes	yes	0.89	19	1.00
47	G5	yes	yes	yes	0.92	8	1.00
48	G5	yes	yes	yes	0.91	4	0.50
49	G5	yes	yes	yes	1	5	1.00
50	G5	yes	yes	yes	0.9	20	0.85

. . . continued

						interventions	
ID	*Group*	*Gamif.*	*Intervene*	*Embod.*	*AVQ*	#	*success*
51	*G5*	*yes*	*yes*	*yes*	*0.88*	*7*	*0.29*

B.3.2 Second Study

B.3.2.1 Collected Data

- **Description.** This data set describes the collected data during the second study conducted with 58 participants to evaluate the effects of intervention strategies on the computational metric (AVQ) and exploration behavior. This data set summarizes the collected data. User trust was measured using the trust scale questionnaire developed by Körber (2019).

- **Data.** The `data set's id` (column 1), the `argument visitation quotient` AVQ (column 2), percentage of visited `challanger arguments` (column 3), the `trust score` (column 4), the `user stance` z_{Φ_0} (column 5), the number of `visited pro arguments` $Args_v^+$ (column 6), the number of `visited con arguments` $Args_v^-$ (column 7), the total number of `visited arguments` (column 8), and the `success rate` (column 9).

- **Source.** The data was collected through an online crowd-sourcing study.

- **Cross-Reference.** Section 7.2 provides additional context and analysis.

							interventions	
ID	*AVQ*	*Chall. Args.*	*Trust*	z_{Φ_0}	$Args_v^+$	$Args_v^-$	#	*success*
				Control Condition				
1	0.78	0.73	0.27	0.25	11	4	0	0.00
2	0.94	0.43	0.68	0.56	8	6	0	0.00
3	0.76	0.38	0.67	0.71	8	5	0	0.00
4	0.81	0.33	0.68	0.76	10	5	0	0.00
5	0.57	0.45	0.48	0.92	6	5	0	0.00
6	0.91	0.46	0.72	0.5	29	25	0	0.00
7	0.82	0.73	0.71	0.28	8	3	0	0.00
8	0.99	0.51	0.5	0.48	24	23	0	0.00
9	0.91	0.56	0.52	0.5	39	31	0	0.00
10	0.93	0.63	0.59	0.41	15	9	0	0.00
11	0.77	0.52	0.48	0.72	11	12	0	0.00
12	0.8	0.44	0.76	0.6	36	28	0	0.00
13	0.99	1	0.39	0.4	11	0	0	0.00
14	0.95	0.83	0.58	0.5	10	2	0	0.00
15	0.78	0.4	0.64	0.71	6	4	0	0.00

. . . continued

							interventions	
ID	*AVQ*	*Chall. Args*	*Trust*	z_{Φ_0}	$Args_v^+$	$Args_v^-$	#	*success*
16	0.95	0.62	0.45	0.46	13	8	0	0.00
17	0.93	0.36	0.57	0.56	7	4	0	0.00
18	0.89	0.4	0.72	0.62	9	6	0	0.00
19	0.96	0.57	0.55	0.43	38	29	0	0.00
20	0.97	0.9	0.61	0.41	9	1	0	0.00
21	0.94	0.6	0.5	0.48	6	4	0	0.00
22	0.71	0.2	0.43	0.77	8	2	0	0.00
23	0.91	0.44	0.71	0.55	14	11	0	0.00
24	0.99	1	0.32	0.4	10	0	0	0.00
25	0.93	0.5	0.4	0.45	5	5	0	0.00
26	0.73	0.1	0.5	0.7	9	1	0	0.00
27	0.98	0.57	0.57	0.5	21	16	0	0.00
28	0.9	0.42	0.54	0.42	5	7	0	0.00
Experimental Condition								
29	1	0.41	0.66	0.56	16	11	9	1.00
30	0.85	0.5	0.57	0.35	5	5	3	1.00
31	0.99	0.87	0.64	0.39	13	2	7	0.83
32	0.99	0.54	0.53	0.45	35	30	49	0.56
33	0.93	0.85	0.45	0.32	11	2	4	1.00
34	0.99	0.57	0.35	0.45	8	6	4	1.00
35	0.99	0.47	0.54	0.47	14	16	32	0.47
36	0.84	0.73	0.64	0.71	3	8	3	1.00
37	1	0.56	0.32	0.47	10	8	9	1.00
38	1	0.47	0.53	0.5	24	27	28	0.76
39	0.98	0.6	0.64	0.42	21	14	19	0.88
40	0.91	0.55	0.52	0.59	18	22	21	1.00
41	0.99	0.55	0.34	0.43	17	14	17	1.00
42	0.95	0.69	0.52	0.42	11	5	8	1.00
43	1	0.36	0.39	0.49	4	7	5	1.00
44	0.89	0.9	0.59	0.34	9	1	1	1.00
45	0.87	0.45	0.56	0.55	39	32	82	0.29
46	0.92	0.85	0.63	0.61	2	11	4	1.00
47	0.88	0.4	0.82	0.41	4	6	5	1.00

. . . continued

							interventions	
ID	*AVQ*	Chall. Args	Trust	z_{Φ_0}	$Args_v^+$	$Args_v^-$	#	success
48	0.86	0.71	0.78	0.26	24	10	33	0.43
49	0.97	0.8	0.65	0.41	8	2	4	1.00
50	0.91	0.9	0.86	0.36	9	1	5	0.67
51	0.97	0.5	0.55	0.55	5	5	5	1.00
52	0.81	0.71	0.58	0.29	10	4	21	0.38
53	0.69	0.75	0.2	0.88	3	9	4	1.00
54	0.94	0.73	0.69	0.6	3	8	3	1.00
55	0.97	0.5	0.58	0.47	6	6	8	0.86
56	0.97	0.73	0.41	0.43	8	3	5	1.00
57	0.91	0.9	0.61	0.35	9	1	5	0.67
58	0.93	0.8	0.33	0.4	8	2	3	1.00

B.3.3 Eye Tracking Study

B.3.3.1 Dependent Measure: AVQ

- **Description.** This data set describes the dependent measures AVQ and focus on challenger arguments, collected during the eye-tracking study. The table consists only of participant IDs who completed the full study after prior registration.
- **Data.** The `data set's id` (columns 1 + 4), and the `argument visitation quotient` AVQ (columns 2 + 5), focus on challenger arguments (columns 3 + 6).
- **Source.** The data was collected as part of a controlled experiment conducted in a laboratory setting.
- **Cross-Reference.** Section 7.3.4 provides additional context and analysis.

Control Condition			Experimental Condition		
ID	AVQ	Chall. Args	ID	AVQ	Chall. Args
2	*1*	*0.6*	5	*0.85*	*0.61*
4	*0.86*	*0.43*	7	*0.95*	*1*
6	*0.95*	*0.5*	9	*1*	*0.53*
8	*0.93*	*0.33*	11	*0.87*	*0.5*
10	*0.81*	*0.6*	15	*0.98*	*0.67*
12	*0.92*	*0.48*	19	*1*	*0.33*
16	*0.86*	*0.4*	21	*0.98*	*0.62*
18	*0.93*	*0.5*	23	*0.95*	*0.51*
20	*0.98*	*0.31*	29	*0.97*	*0.67*
22	*0.89*	*0.53*	31	*0.98*	*0.83*
24	*0.98*	*0.4*	33	*0.96*	*0.65*
26	*0.86*	*0.45*	39	*0.92*	*0.52*
30	*0.81*	*0.63*	41	*0.97*	*0.6*
34	*1*	*0.64*	43	*0.99*	*0.76*
36	*0.86*	*0.43*	44	*0.98*	*0.79*
40	*0.98*	*0.56*	51	*0.96*	*0.9*
42	*0.8*	*0.67*	55	*0.86*	*0.64*
46	*0.86*	*0.59*	56	*0.95*	*0.49*
49	*0.92*	*0.48*	57	*0.91*	*0.93*

Control Condition			Experimental Condition		
ID	AVQ	Chall. Args	ID	AVQ	Chall. Args
52	*0.98*	*0.62*	59	*1*	*0.33*
54	*1*	*0.35*	60	*1*	*0.52*
		0	61	*0.99*	*0.82*
		0	63	*0.93*	*0.5*

B.3.3.2 Eye-Gaze Data: ANF and ADF

- **Description.** This data set describes the dependent eye gaze measures for both *Graph* and *Argument* AOI collected during the eye-tracking study. The table consists only of participant IDs who completed the full study after prior registration.

- **Data.** The `participant's id` (columns 1 + 5), the `areas of interest` (columns 2 + 6), the `average number of fixations` ANF (columns 3 + 7), the `average duration of fixations` ADF (columns 4 + 8).

- **Source.** The data was collected as part of a controlled experiment conducted in a laboratory setting.

- **Cross-Reference.** Section 7.3.4 provides additional context and analysis.

Control Condition				Experimental Condition			
`ID`	*`AOI`*	*`ANF`*	*`ADF`*	`ID`	*`AOI`*	*`ANF`*	*`ADF`*
6	*Argument*	*81.05*	*237*	5	*Argument*	*63.86*	*246*
	Graph	*34.42*	*289*		*Graph*	*19.36*	*255*
8	*Argument*	*98.52*	*125*	9	*Argument*	*125.41*	*243*
	Graph	*27.6*	*116*		*Graph*	*29.59*	*253*
10	*Argument*	*76.17*	*247*	11	*Argument*	*110.54*	*233*
	Graph	*23.32*	*229*		*Graph*	*31.57*	*247*
12	*Argument*	*5.94*	*122*	15	*Argument*	*113.88*	*215*
	Graph	*9.08*	*119*		*Graph*	*11.38*	*253*
16	*Argument*	*97.12*	*214*	19	*Argument*	*116.49*	*231*
	Graph	*17.66*	*209*		*Graph*	*28.74*	*256*
20	*Argument*	*51.41*	*198*	21	*Argument*	*135.19*	*219*
	Graph	*22.07*	*249*		*Graph*	*37.67*	*257*
22	*Argument*	*60.94*	*205*	23	*Argument*	*93.74*	*205*
	Graph	*7.98*	*235*		*Graph*	*4.58*	*298*
24	*Argument*	*86.87*	*193*	29	*Argument*	*80.23*	*238*
	Graph	*37.82*	*264*		*Graph*	*28.56*	*355*
26	*Argument*	*85.68*	*157*	31	*Argument*	*99.36*	*213*

. . . continued

Control Condition				Experimental Condition			
ID	AOI	ANF	ADF	ID	AOI	ANF	ADF
	Graph	*38.91*	*216*		*Graph*	*26.07*	*270*
30	*Argument*	*18.29*	*119*	33	*Argument*	*100.98*	*237*
	Graph	*4.68*	*114*		*Graph*	*24.18*	*267*
34	*Argument*	*80.89*	*130*	39	*Argument*	*110.25*	*233*
	Graph	*28.1*	*135*		*Graph*	*27.67*	*224*
36	*Argument*	*83.16*	*182*	41	*Argument*	*94.08*	*197*
	Graph	*40.48*	*231*		*Graph*	*40.1*	*291*
40	*Argument*	*68.82*	*223*	43	*Argument*	*87.14*	*194*
	Graph	*25.03*	*254*		*Graph*	*41.49*	*265*
42	*Argument*	*93.94*	*222*	44	*Argument*	*88.52*	*174*
	Graph	*5.25*	*242*		*Graph*	*39.33*	*196*
46	*Argument*	*75.48*	*214*	51	*Argument*	*80.63*	*178*
	Graph	*17.5*	*214*		*Graph*	*37.99*	*190*
49	*Argument*	*95.5*	*224*	55	*Argument*	*103.42*	*224*
	Graph	*27.7*	*223*		*Graph*	*19.36*	*239*
52	*Argument*	*89.74*	*266*	56	*Argument*	*111.84*	*201*
	Graph	*27.92*	*285*		*Graph*	*15.86*	*256*
54	*Argument*	*2.95*	*114*	57	*Argument*	*79.03*	*264*
	Graph	*6.1*	*139*		*Graph*	*29.71*	*344*
				59	*Argument*	*76.16*	*142*
					Graph	*26.14*	*163*
				60	*Argument*	*93.59*	*201*
					Graph	*19.38*	*246*
				61	*Argument*	*46.81*	*221*
					Graph	*33.63*	*278*
				63	*Argument*	*93.24*	*251*
					Graph	*70.89*	*225*

B.3.3.3 Independent Measures - UCs

- **Description.** This data set describes the independent measures collected during the eye-tracking study. The table consists only of participant IDs who completed the full study after prior registration. An overview of the employed questionnaires and tests is described in detail in Sec. 7.3.2.
- **Data.** The `participant's id` (column 1), the `need for cognition` NFC (column 2), the `deprivation sensitivity` DS (column 3), the `social curiosity` SC (column 4), the `openness` OP (column 5), the `conscientiousness` CS (column 6), the `affinity for technology` ATI (column 7), the `reading proficiency` RP (column 8), the `perceptual speed` PS (column 9), the `visual working memory` VWM (column 10), the `locus of control` LOC (column 11).
- **Source.** The data was collected as part of a controlled experiment conducted in a laboratory setting. Certain measures were taken in the lab (RP, PS, VWM), while the rest were acquired from participants before the study commenced.
- **Cross-Reference.** Section 7.3.4 provides additional context and analysis.

ID	*NFC*	*DS*	*SC*	*OP*	*CS*	*ATI*	*RP*	*PS*	*VWM*	*LOC*
Control Condition										
2	0.71	0.33	0.53	0.92	0.83	0.62	0.83	0.68	2.73	0.78
4	0.69	0.57	0.77	0.75	0.5	0.6	0.82	0.63	1.96	0.48
6	0.43	0.5	0.87	0.75	0.67	0.51	0.84	0.71	2.96	0.96
8	0.68	0.63	0.83	0.92	0.25	0.67	0.98	0.63	3	0.65
10	0.61	0.57	0.53	0.75	0.83	0.44	0.76	0.68	1.03	0.65
12	0.75	0.63	0.73	0.83	0.5	0.71	0.97	0.54	1.33	0.22
16	0.72	0.73	0.87	0.67	0.75	0.71	0.78	0.56	2.17	0.61
18	0.75	0.7	0.27	0.92	0.92	0.69	0.65	0.69	2.66	0.39
20	0.5	0.37	0.67	0.75	0.33	0.69	0.73	0.53	1.66	0.74
22	0.64	0.97	0.73	1	0.83	0.27	0.84	0.47	1.2	0.39
24	0.69	0.7	0.77	1	0.5	0.78	0.93	0.51	1.83	0.52
26	0.47	0.47	0.8	0.67	0.33	0.67	0.9	0.76	2.6	0.91
30	0.76	0.87	0.67	1	0.75	0.51	0.88	0.63	0.76	0.7
34	0.57	0.43	0.3	0.67	0.83	0.58	0.71	0.61	3.57	0.7

. . . continued

ID	NFC	DS	SC	OP	CS	ATI	RP	PS	VWM	LOC
36	0.6	0.53	0.73	1	0.92	0.67	0.93	0.68	3.93	0.74
40	0.61	0.9	0.77	0.92	0.42	0.62	0.86	0.78	1.7	0.74
42	0.76	0.73	0.83	0.75	0.58	0.73	0.89	0.65	0.3	0.3
46	0.5	0.6	0.87	0.83	0.75	0.51	0.66	0.72	2.56	0.7
49	0.53	0.6	0.87	0.58	1	0.73	0.55	0.76	3.2	0.52
52	0.65	0.53	0.77	0.67	0.67	0.71	0.93	0.74	2.16	0.35
54	0.81	0.9	0.73	1	0.58	0.98	0.58	0.5	1.63	0.7
Experimental Condition										
5	0.53	0.73	0.73	0.42	0.83	0.47	0.93	0.67	3.36	0.52
7	0.63	0.6	0.97	0.83	0.75	0.93	0.92	0.51	1.83	0.91
9	0.61	0.43	0.67	0.58	0.83	0.6	0.79	0.61	2.06	0.61
11	0.72	0.93	0.7	0.83	0.92	0.42	0.91	0.63	2.76	0.78
15	0.71	0.63	0.53	1	0.75	0.69	0.92	0.4	3.03	0.52
19	0.89	0.7	0.87	1	0.83	0.78	0.95	0.53	2.2	0.61
21	0.33	0.37	0.57	0.75	0.25	0.38	0.95	0.65	4	0.43
23	0.57	0.7	0.3	0.67	1	0.78	0.96	0.43	4.1	0.91
29	0.35	0.43	0.67	0.75	0.75	0.38	0.85	0.75	1.9	0.74
31	0.44	0.7	0.67	0.67	0.83	0.56	0.92	0.85	3.03	0.87
33	0.61	0.63	0.7	0.83	0.92	0.42	0.87	0.4	2.13	0.7
39	0.54	0.7	0.67	0.92	0.5	0.47	0.73	0.63	2.96	0.48
41	0.74	0.8	0.67	0.67	0.83	0.49	0.88	0.64	2.3	0.35
43	0.51	1	0.4	0.5	0.42	0.78	0.84	0.61	3.2	0.52
44	0.65	0.3	0.83	0.5	0.92	0.27	0.97	0.75	2.36	0.74
51	0.57	0.6	0.97	0.75	0.25	0.33	0.95	0.65	2.1	0.74
55	0.46	0.43	0.87	0.75	0.75	0.24	0.76	0.58	2.8	0.83
56	0.68	0.57	0.77	0.83	0.83	0.62	0.87	0.74	1.93	0.74
57	0.79	0.53	0.7	0.5	1	0.56	0.86	0.58	0.93	0.7
59	0.56	0.67	0.27	0.83	0.67	0.89	0.9	0.61	1.63	0.57
60	0.46	0.93	0.8	0.5	0.25	0.62	0.91	0.68	2.13	0.87
61	0.49	0.5	0.7	0.5	0.5	0.36	0.87	0.64	3.43	0.57
63	0.67	0.43	0.8	0.83	0.75	0.78	0.89	0.75	2.9	0.48

B.4 Perception Questionnaires (ITU-T, Trust, CE)

B.4.1 ITU-T Questionnaire

- **Description.** This data set describes the data collected in the first study to investigate the effects of the embodiment condition on the user's perception of the system (Möller, 2003) (1 = *strongly disagree*, 5 = *strongly agree*, , (**) 1 = *Bad*, 2 = *Poor*, 3 = *Good*,4 = *Fair*, 5 = *Excellent*) grouped by the following categories: information provided by the system (IPS), communication with the system (COM), system behavior (SB), dialogue (DI), user's impression of the system (UIS), acceptability (ACC), argumentation (ARG, and overall quality QLT). Significant values are check-marked [✓]. (*) Items have to be inverted.

- **Data.** The `question category` (column 1), the `textual question` (column 2), the `embodiment condition` (column 3), the `means` ($\mu_{embodied}$, μ_{chat}) with 95% `confidence interval` denoted by error bars (column 4), the `p-value` (column 5), the `effect size` (column 6), the `statistical significance` (column 7).

- **Source.** This data was collected through an online crowd-sourcing study.

- **Cross-Reference.** Section 7.1.3 provides additional context and analysis.

Cat.	Question	Embod.	$\mu_{embodied}/\mu_{chat}$	p	r	
IPS	*1. The system has provided you with the desired information.*	*Yes*	**3.44**	*.781*		✗
		No	3.36			
	2. The system's answers and proposed solutions were clear.	*Yes*	3.54	*.749*		✗
		No	**3.67**			
	3. You would rate the provided information as true.	*Yes*	**3.67**	*.585*		✗
		No	3.61			
	4. The information provided by the system was complete.	*Yes*	3.25	*.794*		✗
		No	**3.36**			
COM	*1. The system always understood you well.*	*Yes*	2.67	*.848*		✗
		No	**2.75**			
	*2. You had to concentrate to understand what the system expected from you.**	*Yes*	3.71	*.501*		✗
		No	**3.83**			
	3. The system's responses were well understandable.	*Yes*	**3.75**	*.316*		✗

. . . continued

Cat.	Question	Embod.	$\mu_{embodied}/\mu_{chat}$	p	r	
		No	3.54			
	4. You were able to interact efficiently with the system.	*Yes*	2.92	.533		✗
		No	**3.08**			
SB	*1. You knew what the system expected from you at each point of the interaction.*	*Yes*	**2.71**	.774		✗
		No	2.69			
	2. In your opinion, the system processed your responses (specifications) correctly.	*Yes*	3.23	.665		✗
		No	**3.28**			
	3. The system's behavior was always as expected.	*Yes*	**3.02**	.410		✗
		No	2.89			
	*4. The system often failed to understand you.**	*Yes*	**3.27**	.341		✗
		No	3.00			
	5. The system reacted naturally.	*Yes*	**3.27**	**.037**	.227	✓
		No	2.78			
	6. The system reacted flexibly.	*Yes*	**3.02**	.255		✗
		No	2.75			
	7. You were able to control the interaction in the desired way.	*Yes*	2.81	.580		✗
		No	**2.92**			
	*8. The system reacted too slowly.**	*Yes*	3.29	**<.001**	.485	✓
		No	**2.19**			
	9. The system reacted politely.	*Yes*	**4.37**	.615		✗
		No	4.31			
	*10. The system's responses were too long.**	*Yes*	2.58	**.016**	.263	✓
		No	**2.03**			
DI	*1. You perceived the dialogue as natural.*	*Yes*	**3.52**	**.032**	.234	✓
		No	3.03			
	2. It was easy to follow the flow of the dialogue.	*Yes*	3.31	.869		✗
		No	**3.36**			
	*3. The dialogue was too long.**	*Yes*	2.42	.093		✗
		No	**2.11**			
	4. The course of the dialogue was smooth.	*Yes*	**3.44**	.864		✗
		No	3.44			
	5. You and the system could clear misunderstandings easily.	*Yes*	2.79	.445		✗
		No	**3.00**			
	*6. You would have expected more help from the system.**	*Yes*	3.75	**.014**	.269	✓

. . . continued

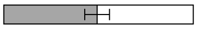
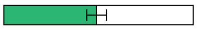
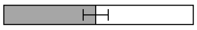
Cat.	Question	Embod.	$\mu_{embodied}/\mu_{chat}$	p	r	
		No	**3.19**			
UIS	*1. Overall, you were satisfied with the dialogue.*	*Yes*	**3.40**	*.431*		✗
		No	3.22			
	2. The dialogue with the system was useful.	*Yes*	3.27	*.686*		✗
		No	**3.39**			
	3. It was easy for you to obtain the information you wanted.	*Yes*	**2.92**	*.804*		✗
		No	2.86			
	4. You have perceived the dialogue as pleasant.	*Yes*	**3.90**	*.154*		✗
		No	3.56			
	5. You felt relaxed during the dialogue.	*Yes*	3.42	*.146*		✗
		No	**3.69**			
	6. Using the system was fun.	*Yes*	**3.19**	*.203*		✗
		No	2.83			
ACC	*1. In the future, you would use the system again.*	*Yes*	3.83	*.892*		✗
		No	**3.86**			
	2. You would recommend the system to a friend.	*Yes*	**3.21**	*.067*		✗
		No	2.75			
ARG	*1. I felt motivated by the system to discuss the topic.*	*Yes*	**3.64**	*.068*		✗
		No	2.94			
	2. I would rather use this system than read the arguments in an article.	*Yes*	**3.15**	*.504*		✗
		No	2.94			
	3. The possible options to respond to the system were sufficient.	*Yes*	3.00	*.806*		✗
		No	**3.06**			
	4. The arguments the system presented are conclusive.	*Yes*	**3.21**	*.419*		✗
		No	3.06			
	5. I felt engaged in the conversation with the system.	*Yes*	**3.40**	***.039***	*.226*	✓
		No	2.83			
	*6. The interaction with the system was confusing.**	*Yes*	**2.73**	*.149*		✗
		No	3.14			
	*7. I do not like that the arguments are provided incrementally.**	*Yes*	3.04	*.111*		✗
		No	**2.67**			
QLT**	*1. What is the overall impression of the system?*	*Yes*	**3.67**	*.047*	*.216*	✓
		No	3.17			

B.4.2 Trust Questionnaire

- **Description.** This data set describes the data collected in the first study to investigate the effects of the embodiment condition on the user trust using the trust scale questionnaire (Körber, 2019) (1 = *strongly disagree*, 5 = *strongly agree*) grouped by the following categories: understanding/predictability (UP), familiarity (F), propensity to trust (PT) and trust in automation (TA). Significant values are check-marked [✓]. (*) Items have to be inverted.
- **Data.** The `question category` (column 1), the `textual question` (column 2), the `embodiment condition` (column 3), the `means` ($\mu_{embodied}$, μ_{chat}) with 95% `confidence interval` denoted by error bars (column 4), the `p-value` (column 5), the `effect size` (column 6), the `statistical significance` (column 7).
- **Source.** This data was collected through an online crowd-sourcing study.
- **Cross-Reference.** Section 7.1.3 provides additional context and analysis.

Cat.	Question	Embod.	$\mu_{embodied}/\mu_{chat}$	p	r	
UP	*1. The system state was always clear to me.*	*Yes*	**2.90**	.244		✗
		No	2.67			
	*2. The system reacts unpredictably.**	*Yes*	**2.50**	.455		✗
		No	2.44			
	3. I was able to understand why things happened.	*Yes*	3.10	.522		✗
		No	**3.25**			
	*4. It's difficult to identify what the system will do next.**	*Yes*	3.33	.281		✗
		No	**3.06**			
F	*1. I already know similar systems.*	*Yes*	**2.58**	.704		✗
		No	2.50			
	2. I have already used similar systems.	*Yes*	2.50	.821		✗
		No	**2.56**			
PT	*1. One should be careful with unfamiliar automated systems.**	*Yes*	**3.46**	**.025**	.250	✓
		No	3.97			
	2. I rather trust a system than I mistrust it.	*Yes*	3.19	.969		✗
		No	3.19			
	3. Automated systems generally work well.	*Yes*	**3.25**	**.022**	.244	✓
		No	2.72			

. . . continued

Cat.	Question	Embod.	$\mu_{embodied}/\mu_{chat}$		p	r	
TA	*1. I trust the system.*	*Yes*		***3.44***	***.039***	*.244*	✓
		No		*2.97*			
	2. I can rely on the system.	*Yes*		***2.96***	*.951*		✗
		No		*2.94*			

B.4.3 User Engagement Questionnaire

- **Description.** This data set describes the data collected in the first study study to investigate the effects of the embodiment condition on user experience using the short user engagement questionnaire (O'Brien et al., 2018) (1 = *strongly disagree*, 5 = *strongly agree*) grouped by the following categories: Focused attention (FA), perceived usability (PU), aesthetic appeal (AE) and reward factor (RW). Significant values are check-marked [✓]. (*) Items have to be inverted.

- **Data.** The `question category` (column 1), the `textual question` (column 2), the `embodiment condition` (column 3), the `means` ($\mu_{embodied}$, μ_{chat}) with 95% `confidence interval` denoted by error bars (column 4), the `p-value` (column 5), the `effect size` (column 6), the `statistical significance` (column 7).

- **Source.** This data was collected through an online crowd-sourcing study.

- **Cross-Reference.** Section 7.1.3 provides additional context and analysis.

Cat.	Question	Embod.	$\mu_{embodied}$ / μ_{chat}	p	r	
FA	*1. I lost myself in this experience.*	*Yes*	**2.63**	.121		✗
		No	2.19			
	2. The time I spent using the application just slipped away.	*Yes*	**2.60**	.451		✗
		No	2.44			
	3. I was absorbed in this experience.	*Yes*	**3.08**	.144		✗
		No	2.69			
PU	*1. I felt frustrated while using the application.**	*Yes*	**3.10**	.641		✗
		No	3.22			
	*2. I found this application confusing to use.**	*Yes*	**2.98**	.385		✗
		No	3.19			
	*3. Using this application was taxing.**	*Yes*	**2.71**	.379		✗
		No	2.97			
AE	*1. The application was attractive.*	*Yes*	**3.19**	.875		✗
		No	3.14			
	2. The application was aesthetically appealing.	*Yes*	**3.25**	.174		✗
		No	3.06			
	3. This application appealed to my senses.	*Yes*	**2.98**	.615		✗

. . . continued

Cat.	Question	Embod.	$\mu_{embodied}/\mu_{chat}$	p	r	
		No	2.89			
RW	*1. Using the application was worthwhile.*	*Yes*	**3.38**	**.022**	.250	✓
		No	2.86			
	2. My experience was rewarding.	*Yes*	**3.40**	.076		✗
		No	2.94			
	3. I felt interested in this experience.	*Yes*	**3.71**	.586		✗
		No	3.56			

Annex I

Publications

I contributed to several publications during the thesis research. An overview is given below (chronological order).

I. Weber, K. (2017). *Adaption eines Sozialen Roboters auf Basis von Bestärkendem Lernen mit Linearer Funktionsapproximation und Sozialen Signalen* [Master's Thesis]. University of Augsburg

II. Weber, K., Ritschel, H., Lingenfelser, F., & André, E. (2018b). Real-Time Adaptation of a Robotic Joke Teller Based on Human Social Signals. *Proceedings of the 17th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2259–2261

III. Weber, K., Ritschel, H., Aslan, I., Lingenfelser, F., & André, E. (2018a). How to Shape the Humor of a Robot - Social Behavior Adaptation Based on Reinforcement Learning. *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI)*, 154–162

IV. Rach, N., Weber, K., Pragst, L., André, E., Minker, W., & Ultes, S. (2018b). EVA: A Multimodal Argumentative Dialogue System. *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI)*, 551–552

V. Rach, N., Weber, K., Aicher, A., Lingenfelser, F., André, E., & Minker, W. (2019). Emotion Recognition Based Preference Modelling in Argumentative Dialogue Systems. *Proceedings of the 2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 838–843

VI. André, E., Bayer, S., Benke, I., Benlian, A., Cummins, N., Gimpel, H., Hinz, O., Kersting, K., Maedche, A., Muehlhaeuser, M., et al. (2019). Humane Anthropomorphic Agents: The Quest for the Outcome Measure. *Proceedings of the Pre-ICIS Workshop "Values and Ethics in the Digital Age"*, *12*(4), 1–16

VII. Weber, K., Tinnes, L., Huber, T., Heimerl, A., Pohlen, E., Reinecker, M.-L., & Andé, E. (2020c). Towards Demystifying Subliminal Persuasiveness: Using XAI-Techniques to Highlight Persuasive Markers of Public Speeches. *Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS)*, 113–128

VIII. Weber, K., Janowski, K., Rach, N., Weitz, K., Minker, W., Ultes, S., & André, E. (2020a). Predicting Persuasive Effectiveness for Multimodal Behavior Adaptation using Bipolar Weighted Argument Graphs. *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 1476–1484

IX. Weber, K., Rach, N., Minker, W., & André, E. (2020b). How to Win Arguments: Empowering Virtual Agents to Improve Their Persuasiveness. *Datenbank-Spektrum*, *20*, 161–169

X. Ritschel, H., Kiderle, T., Weber, K., & André, E. (2020a). Multimodal Joke Presentation for Social Robots Based on Natural-language Generation and Nonverbal Behaviors. *Proceedings of the 2nd Workshop on Natural Language Generation for Human–robot Interaction (NLG4HRI)*, 1–3

XI. Ritschel, H., Kiderle, T., Weber, K., Lingenfelser, F., Baur, T., & André, E. (2020b). Multimodal Joke Generation and Paralinguistic Personalization for a Socially-aware Robot. *Proceedings of the 22nd International International Conference on Practical Applications of Agents, Multi-Agent Systems, and Trustworthiness. (PAAMS)*, 278–290

XII. Rach, N., Weber, K., Yang, Y., Ultes, S., André, E., & Minker, W. (2021). EVA 2.0: Emotional and Rational Multimodal Argumentation between Virtual Agents. *it - Information Technology*, *63*(1), 17–30

XIII. Weber, K., Aicher, A., Minker, W., Ultes, S., & André, E. (2023a). Fostering User Engagement in the Critical Reflection of Arguments. *Proceedings of the 13th International Workshop on Spoken Dialogue Systems (IWSDS)*, 1–16

XIV. Aicher, A., Weber, K., Minker, W., André, E., & Ultes, S. (2023). The Influence of Avatar Interfaces on Argumentative Dialogues. *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents (IVA)*, 1–8

XV. Weber, K., Tinnes, L., Huber, T., & André, E. (2023b). Exploring the Effect of Visual-Based Subliminal Persuasion in Public Speeches Using Explainable AI techniques. *Proceedings of the 25th International Conference on Human-Computer Interaction (HCII)*, 381–397

XVI. Aicher, A., Weber, K., André, E., Minker, W., & Stefan, U. (2024). BEA: Building Engaging Argumentation. *Proceedings of the 1st International Conference on Robust Argumentation Machines (RATIO)*, 1–17

XVII. Weber, K., Hogh, N., Conati, C., & André, E. (2024). A Gaze into Argumentative Chatbots: Exploring the Influence of Challenger Arguments on Reflection and Attention. *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents (IVA)*, 1–10

Annex II

Awards

- PLACE 2nd **Klaus Weber**, Hannes Ritschel, Ilhan Aslan, Florian Lingenfelser and Elisabeth André. 2018. **How to shape the humor of a robot - social behavior adaptation based on reinforcement learning.** In Proceedings of the 20th International Conference on Multimodal Interaction - ICMI '18, Boulder, CO, USA — October 16 - 20, 2018

- PLACE 2nd **Klaus Weber**, Kathrin Janowski, Niklas Rach, Katharina Weitz, Wolfgang Minker, Stefan Ultes and Elisabeth André. 2020. **Predicting persuasive effectiveness for multimodal behavior adaptation using bipolar weighted argument graphs**. In Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS' 20). ACM, New York, NY, 1476-1484