AI Enhanced Reasoning: Augmenting Human Critical Thinking with AI Systems

by

Valdemar M. Danry

BA, University of Copenhagen (2021)

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2023

AUTHORED BY

Valdemar M. Danry
Program in Media Arts and Sciences
May 31, 2023

CERTIFIED BY

Prof. Pattie Maes
Professor of Media Arts and Sciences
Thesis Supervisor

ACCEPTED BY

Prof. Tod Machover
Academic Head
Program in Media Arts and Sciences

AI Enhanced Reasoning: Augmenting Human Critical Thinking with AI Systems

by

Valdemar M. Danry

Submitted to the Department of Media Arts and Sciences
on May 19, 2023, in partial fulfillment of the
requirements for the degree of
Master of Science in Media Arts and Sciences

Abstract

The pursuit of knowledge and understanding has been a driving force for humanity since the beginning of time. This relentless quest for reason has shaped the world as we know it, enabling us to unlock secrets of the cosmos, develop innovative technologies, and address complex global challenges. However, despite our cognitive leaps, we still grapple with the limitations of our rationality, biases, and emotions, especially in today's increasingly complex and information-saturated world. As AI systems become more entwined with our daily lives and institutions, there is a growing need to design and deploy AI systems that augment human reasoning, foster critical thinking, and promote well-informed decision-making.

This thesis investigates the potential for AI-enhanced reasoning systems and their impact on human decision-making. Specifically, it explores three distinct aspects of critical thinking with AI systems: (1) the development of AI logic-checking systems designed to help identify reasoning flaws, (2) examining the susceptibility of individuals to deceptive AI-generated explanations, and (3) assessing the potential of a novel AI-framed questioning interaction method to provoke critical thinking through a series of human subjects experiments.

These investigations aim to shed light on the implications of AI systems on human reasoning and provide insights into designing AI interventions that meaningfully enhance our cognitive abilities. The findings demonstrate the potential for intelligently designed AI systems to support human reasoning, while also highlighting the potential risks associated with overreliance on these tools. By addressing these challenges, this thesis contributes to the ongoing conversation around the development of AI systems that advance our reasoning, and steps towards cultivating a discerning and rational citizenry capable of navigating the complexities of the modern world.

Thesis Supervisor: Pattie Maes
Title: Professor of Media Arts and Sciences

AI Enhanced Reasoning: Augmenting Human Critical Thinking with AI Systems

by

Valdemar M. Danry

This thesis has been reviewed and approved by the following committee members:

THESIS SUPERVISOR

Prof. Pattie Maes
Professor of Media Arts and Sciences
Massachusetts Institute of Technology

THESIS READER

Prof. David G. Rand
Professor of Management Science and Brain and Cognitive Sciences
Massachusetts Institute of Technology

THESIS READER

Prof. Chenhao Tan
Assistant Professor of Computer Science and Data Science
University of Chicago

ACKNOWLEDGMENTS

My deepest appreciation goes to my advisor, Pattie Maes, for her unwavering support, belief in my ideas, and allowing me the creative freedom to explore my research interests. Thank you, Pattie, for your warmth, insights, and positivity that have guided and inspired me throughout this adventure.

A heartfelt thank you to David G. Rand and Chenhao Tan, whose groundbreaking work has immensely influenced and provided direction for my own research. Your dedication and passion for this field have been a beacon of inspiration to me.

I am profoundly grateful to Pat Pataranutaporn for embracing this philosophy undergrad as a visiting student, for inspiring and thought-provoking philosophical discussions, and for challenging my ideas without hesitation. Pat, your mentorship has truly shaped my academic journey, and I am deeply thankful for the opportunity to learn from you.

To my incredible collaborators, Yaoli Mao, Joanne Leong, Misha Sra, Sangwon Leigh, Florian Floyd Mueller, Ziv Epstein, Matthew Groh, Caitlin Morris, Lancelot Blanchard, and Eli Villa, thank you for embarking on this wild ride with me. Your brilliance, enthusiasm and scientific rigour have made even the most intense late-night research sessions rewarding and memorable.

To my research colleagues at the Media Lab, particularly the remarkable individuals at the Fluid Interfaces Lab, I cannot thank you enough for the stimulating and supportive environment we've created together. Your creativity, perseverance, and friendship have truly made me a better researcher and person.

I owe a lifetime of gratitude to my loving parents, who have always embraced my dreams—no matter how far away they took me from our cozy hometown in Denmark. Your belief in my potential is the backbone of my academic aspirations. To my wonderful sisters, Nikoline and Frederikke, you both have been a consistent source of motivation and love, and I cherish our bond immensely.

Above all, my deepest gratitude go to Nelly-Charlott Schneider. Thank you, Nelly, for your unwavering support, your patience amidst the ever-shifting tides of academic life, and for being an endless source of positivity and humor when things have gotten hard. I could not have navigated these choppy waters without your inexhaustible well of strength and love.

Since childhood, I have dreamt of attending MIT and making my mark in the world of groundbreaking technology. The cherished opportunity to actualize

this dream, surrounded by phenomenal mentors, collaborators, friends, and loved ones, fills me with utmost gratitude and a deep sense of accomplishment.

CONTENTS

1 | ABOUT THIS WORK

> "Man is a rational animal who always loses his temper when he is called upon to act in accordance with the dictates of reason." - Oscar Wilde

Since the dawn of time, humans have been driven by the pursuit of knowledge and understanding. Our boundless curiosity and relentless quest for reason have led us to explore the depths of the oceans, unravel the mysteries of the cosmos, and unlock the secrets of our own DNA. Throughout history, we have found ways to cultivate our ability to reason, to think critically, and to solve problems. This power of reasoning has been the cornerstone of human progress, allowing us to develop innovative tools, technologies, and systems that have shaped the world as we know it.

Yet, our capacity for rational thought is not without its limitations. As Oscar Wilde so aptly observed, we often struggle to act in accordance with reason, succumbing to the whims of our emotions, biases, and cognitive blind spots. These limitations are amplified in today's fast-paced, information-saturated world, where we are frequently bombarded with complex and often contradictory information, making it increasingly difficult to separate fact from fiction and to make well-informed decisions.

This is evident in the myriad of pressing global challenges we face today, ranging from climate change and resource scarcity to political polarization and the spread of misinformation. Addressing these complex issues requires not only the development of innovative solutions but also the cultivation of a more discerning and rational citizenry, capable of engaging in nuanced information processing and making informed decisions for the betterment of humanity.

At the same time, AI systems are becoming increasingly proficient at human tasks, and as these tools get more and more integrated into our institutions and daily practises, we might risk offloading essential critical thinking to machines. Journalists and students are, for instance, already leveraging AI systems to do their critical writing for them. As these AI systems are not engineered to make people think more deeply, but rather to complete tasks quickly and efficiently, the content of critical writing pieces might end up conforming to the narratives

embedded within AI models instead of conforming to the reasoning of their authors; not letting the authors form their own thoughts.

In light of these challenges, the quest for augmenting human reasoning has never been more relevant. The aim of this thesis is not to provide a comprehensive solution to these problems, but rather to characterize the problem and propose a new type of AI systems designed to enhance our ability to reason, think critically, and make better decisions. By exploring the potential danger of these systems and highlighting the need for designing novel interventions or reimagining existing ones, this thesis seeks to pave the way for more effective AI-driven support in fostering critical thinking and well-informed decision-making. In particular, this work investigates three aspects of critical thinking with AI systems: (1) development of AI systems that help us identify flaws in reasoning using theory of logic, (2) investigation of the effects of AI systems that deliberately deceive us, and (3) exploration of the impact of a novel AI-framed questioning interaction method to prompt critical thinking. The thesis is structured as follows:

In Chapter 2, I introduce the concept of AI-enhanced reasoning, a framework that applies AI systems to augment and improve human cognitive abilities in reasoning and critical thinking. I differentiate AI-enhanced reasoning from related concepts such as AI-assisted decision-making and fact-checking and discuss the essential factors of AI-enhanced reasoning, including informational structures, cognitive factors, and AI interaction methods.

In Chapter 3, I discuss the development of "Wearable Reasoner", an AI logic-checking system that provides users with feedback on one type of logical fallacies. The study found that users assisted by the system were significantly more likely to agree with non-fallacious arguments and considered them more reasonable when the AI feedback contained explanations for its classifications. However, users also tended to rely on the device as a substitute for quick heuristics, highlighting the potential for depending on the system to think for them — even when they might disagree.

In Chapter 4, I discuss the impact of deceptive AI-generated explanations on people's discernment of true and false information, particularly in the context of news headlines and trivia items. The findings reveal the susceptibility of people to deceptive AI systems and highlight the importance of developing strategies or alternative interaction methods that allow people think more critically about deceptive recommendations.

In Chapter 5, I present a novel AI-framed Questioning method inspired by Socratic questioning that uses intelligently formed questions to provoke human reasoning and improve the discernment of logical validity in socially divisive

statements. Results show that AI-framed Questioning significantly increases discernment accuracy of flawed statements over causal AI-explanations and control conditions. This highlights the potential of AI systems as stimulators of critical thinking rather than information tellers, encouraging and assisting users to actively engage in reasoning about potentially misleading information.

Finally, in Chapter 6, I discuss the limitations, challenges, and potential directions for future research on AI-Enhanced Reasoning systems. This includes investigating learning effects, understanding the influence of social factors, studying the systems in real-world contexts, and exploring a wider range of information evaluation capabilities and interaction methods. Furthermore, I explore the idea of more deeply integrating AI systems with human cognition, potentially utilizing brain-sensing technologies and focusing on fostering a symbiotic relationship between humans and AI.

During my time at the 2-year MIT Media Lab graduate program, I also got to explore additional systems for augmenting human cognition beyond what is presented in this paper. Specifically, I focused on deepfake technology and large language models to create AI-generated characters that support learning and well-being [86]. This includes a system that lets you talk to different versions of yourself [87], virtual instructors based on liked or admired people [88], and "living memories" which allows you to talk with people from the past [85]. Moreover, I also focused on how to elicit "experiential integration" where humans can seamlessly and phenomenologically integrate with technology [19, 16, 78] without feelings of disruption [101]. Although these projects might differ from what is presented specifically in this thesis, they share the common thread of seeking to understand and enhance human cognition through the thoughtful integration of AI systems.

2 AI ENHANCED REASONING: A CONCEPTUAL FRAMEWORK

"Tools have not only extended our physical abilities but also our mental ones, allowing us to think in ways that would have been impossible without them" - Nicholas Carr

2.1 DEFINITION AND IMPORTANCE OF AI-ENHANCED REASONING

We live in a world accelerated by information, overloading us with tremendous volumes of data, exceeding our biological brain's processing capability. As we are increasingly exposed to massive amounts of information that can potentially be deceptive, misleading, or strictly false, we are struggling to determine what to believe and what not to believe [60].

The proliferation of misinformation and disinformation, both human and AI generated, has been exacerbated by the rise of social media platforms and the ease with which information can be shared and spread [113]. Misinformation, referring to false or misleading information without an intent to deceive, and disinformation, referring to false information deliberately spread to deceive or manipulate, both contribute to the erosion of trust in institutions and the polarization of society [59]. For instance, during the COVID-19 pandemic, a surge of misinformation led to strong confusion and mistrust in governments globally, undermining public health efforts and putting the population at risk [91, 115].

In this context, there is a pressing need for tools and strategies to help individuals navigate the torrent of information and develop critical thinking skills that can empower them to discern fact from fiction. To address this issue, recent advancements in artificial intelligence could be designed to actively engage users in reflective thinking and promote a deeper understanding of underlying concepts or deficiencies of information by helping individuals develop the necessary cognitive skills to question misleading or vague information and make more informed decisions.

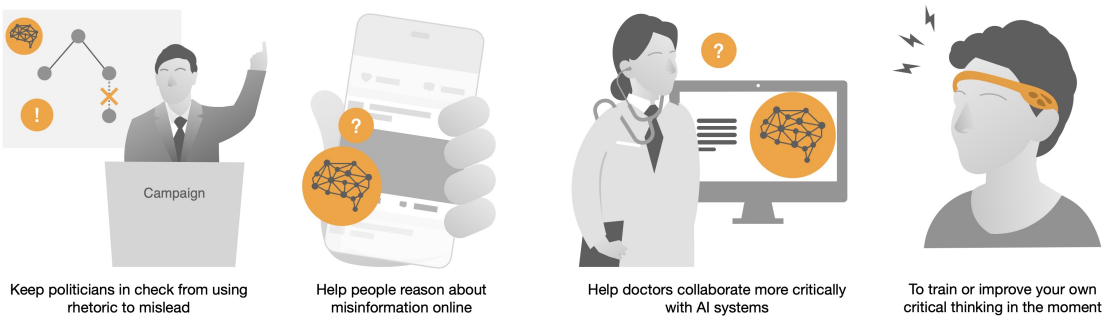

Figure 1: Example applications of AI-Enhanced Reasoning systems

2.1.1 Definition

AI-enhanced reasoning refers to the application of artificial intelligence (AI) systems to support, facilitate and improve human reasoning capabilities by providing insights, identifying patterns, uncovering biases, and offering guidance that is intuitive to the user and which enables them to make better-informed decisions that they feel that they arrived at through their own thinking processes (see Figure 2).

AI-enhanced reasoning systems differ from other AI information-processing systems in that they focus not only on providing accurate information or optimize decision outcomes but also on actively engaging users in reflective thinking. In order to be considered an AI-enhanced reasoning system, the following conditions should be met:

1. The AI system should interact with users in a manner that fosters critical thinking and reflection when appropriate.
2. The system should provide guidance and support in the evaluation of information, identification of logical fallacies, and detection of biases.
3. The AI system should be adaptive and responsive to the user's needs, offering personalized feedback and suggestions based on the user's cognitive abilities and reasoning styles.

2.2 RELATED CONCEPTS AND DIFFERENTIATION

AI-enhanced reasoning is a distinct concept with unique characteristics and contributions. However, it shares some similarities with other AI-driven approaches,

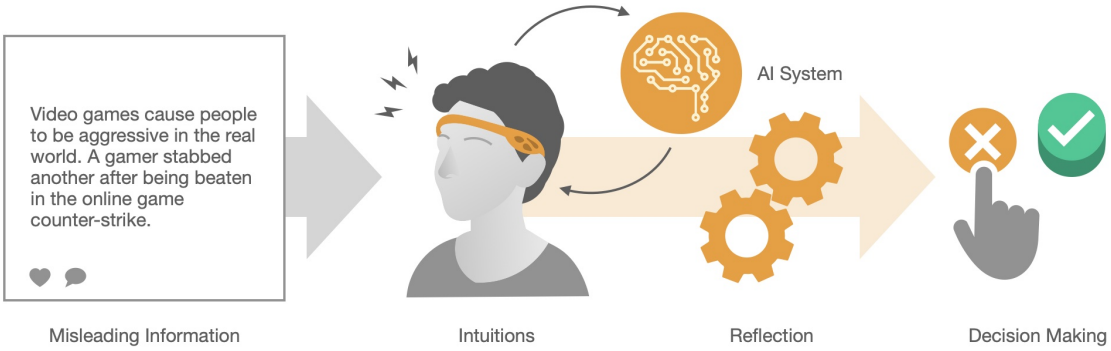

Figure 2: AI-Enhanced Reasoning systems aim to support and facilitate critical reflection when appropriate

such as AI-assisted decision-making and AI-based fact-checking. In this section, we differentiate these related concepts, highlighting their key differences and similarities.

2.2.1 AI-Enhanced Reasoning vs. AI-Assisted Decision-Making

AI-assisted decision-making involves using AI systems to analyze data, predict outcomes, and provide recommendations to help users make informed choices. These systems typically focus on optimizing decisions based on specific criteria or objectives, such as minimizing costs or maximizing performance [22]. In contrast, AI-enhanced reasoning is concerned with the broader cognitive processes involved in evaluating arguments, understanding complex relationships, and making well-founded decisions.

While both approaches aim to support human decision-making, AI-enhanced reasoning places a greater emphasis on fostering critical thinking and including the user in the reflective processes underlying a decision, rather than merely providing optimized solutions or recommendations. Moreover, AI-enhanced reasoning systems are designed to engage users in a more interactive and collaborative manner, encouraging them to actively participate in the reasoning process and develop their own cognitive skills instead of relying on a system to tell them what decision to make.

2.2.2 AI-Enhanced Reasoning vs. Fact-checking

One of the main strategies currently used to mitigate problems of misinformation is "fact-checking". Fact-checking aims to verify the accuracy of discrete pieces of information, such as claims, statements, or data points. Fact-checking systems often rely on the comparison of claims against established sources of information or expert knowledge to determine their truthfulness. In recent years, AI-based fact-checking tools have been developed to automate this process and help users identify false or misleading information more efficiently [42]. However, AI-based fact-checking varies from AI-Enhanced reasoning in two key aspects: (1) goals and (2) cognitive effects.

Goals of Fact-checking

While fact-checking primarily aims to verify the accuracy of discrete pieces of information or specific claims, informational structures does not only consist of discrete claims but also logical connections between claims which are used to support or refute each other. By only focusing on verifying the truth of a particular piece of information, fact-checking systems omit essential informational structures causing it in many cases to fail at providing accurate evaluations, for instance, in arguments where true evidence is used to support a false claim. For instance, a malicious politician might use a true fact to support an untrue conclusion as was the case in 2015 when US Senator James Inhofe, a known climate change skeptic, brought a snowball to the Senate floor to argue against the reality of global warming. Here, he used the existence of the snowball as evidence, claiming that since it was cold enough to snow in Washington, D.C., global warming must not be real. In the example, the argument was logically invalid but still ended up persuading people due to its deployment of what is known as logical fallacies.

Cognitive Limitations of Fact-checking

While the aim of fact-checking is to correct potential false beliefs about information, merely telling people that they hold inaccurate beliefs, as is often the case with fact-checking, may not be sufficient to induce belief revision. For instance, research has shown that simply providing people with corrected misinformation does not always lead to changes in false beliefs, and in some cases, can even backfire and reinforce misconceptions [60, 82]. For instance, individuals may be more likely to accept information that confirms their pre-existing beliefs (also know as the "confirmation bias") or give undue weight to information that

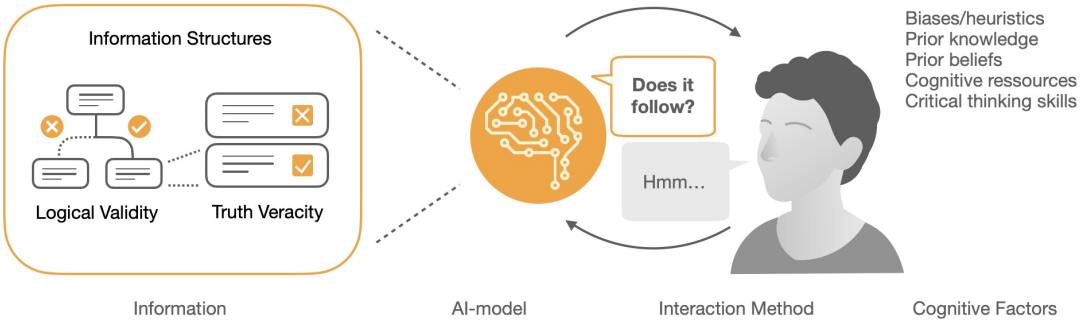

Figure 3: Factors of AI-Enhanced Reasoning

is easily accessible or memorable (also known as the "availability heuristic") [81, 47]. This makes people less likely to accept corrections in their beliefs about misinformation, and emphasizes the need to also consider other factors such as critical reflection when trying to make people better discerning [23]. For instance, studies have shown that engaging individuals in tasks that require them to evaluate evidence, identify logical fallacies, and discern inconsistencies within arguments can lead to improved understanding and belief change [52], and that an increased ability to reflect on one's own thought processes, can help individuals become more conscious about biases and errors in their reasoning, making them overall more critical thinkers and more receptive to justified belief revisions [52].

2.3 ESSENTIAL FACTORS IN AI-ENHANCED REASONING

The goal of AI-Enhanced Reasoning to support and guide critical thinking and reflection requires careful consideration of numerous factors such as (1) the information type and structures like logic and truth veracity that the AI system informs the user about, (2) cognitive factors of the user such as prior knowledge, prior beliefs, cognitive resources and critical thinking skills that might impact influence of the AI system, (3) the design goals of the AI system that identify misleading informational structures or the cognitive factors of its user, and (4) the interaction methods that are most likely to cause critical thinking in the target user (see 3).

2.3.1 Informational Structures

In the realm of language and communication, different types of speech acts serve various purposes and convey diverse types of information. For instance, while assertives like “The Earth revolves around the Sun” expresses something true or false about the world, directives such as “Please pass the salt” aim to influence the listener’s actions. Commissives, like the statement “I promise to help you tomorrow,” commit the speaker to a future course of action. Expressives, on the other hand, reveal the speaker’s emotions or attitudes, as in "I’m sorry for your loss." Lastly, declarations, like “I now pronounce you husband and wife,” bring about changes in the world simply through the act of speaking [97].

The understanding of different types of speech acts within a piece of information can contribute to improving reasoning by enabling individuals to recognize the intent of the information and respond accordingly. For instance, if the claim "Vaccines are harmful" is interpreted as assertive, intending to present a factual argument against vaccines, one would respond by critically evaluating the factual and logical accuracy of the claim in order to verify its truth. However, if interpreted as an expressive, the statement could be seen as conveying the speaker’s fear or concern about vaccines. In this scenario, the appropriate response might focus on addressing the speaker’s emotions and providing reassurance rather than debating the factual accuracy of the statement. Recognizing the distinction between assertive and expressive speech acts can lead to more effective communication and understanding.

Assertives, Logic and Facts

Proper understanding and evaluation of especially assertive statements are essential because incorrect or misleading beliefs can lead to negative consequences and poor decision-making. For instance, if an individual believes the incorrect claim that vaccines are harmful, they might decide not to vaccinate themselves or their children, leading to increased risks of preventable diseases. In this context, developing the ability to accurately identify and evaluate assertive statements is important for making informed decisions and avoiding potential pitfalls. Here, the concepts of logic and truth become crucial, as they are the factors that can be evaluated to ensure the accuracy and reliability of assertive statements.

Facts are objective pieces of information that are true and verifiable, independent of personal opinions or emotions. For instance, the statement "Water boils at 100 degrees Celsius at sea level" is a fact, as it can be scientifically tested

and verified as true or false. Facts are crucial for establishing a foundation for rational and critical thinking, as they provide a basis for evaluating the accuracy of claims and forming well-informed opinions.

Logic, on the other hand, refers to the principles and reasoning processes that guide the evaluation and formation of beliefs, arguments, and conclusions. In essence, logic deals with the relationships between statements and the validity of inferences drawn from them regardless of their truth. Logical validity is the quality of an argument whereby if its premises are true, then its conclusion must also be true. For instance, consider the following logical argument:

1. Deforestation contributes to climate change by releasing stored carbon dioxide into the atmosphere.
2. The Amazon rainforest is being deforested at an alarming rate.
3. Therefore, deforestation of the Amazon rainforest contributes to climate change.

This argument is logically valid because if the premises (1 and 2) are true (true facts), then the conclusion (3) must also be true by the rules of logical validity. An argument can be logically valid even if its premises are false, and if they are false, the conclusion is necessarily false too. If we are unsure if any premise is true, we can use principles of logic to come up with premises that if true would support the premise we are unsure about. For instance, the conclusion (3) might be ambiguous if presented by itself but by introducing (1) and (2) it becomes apparent.

Facts vs. Logic

The difference between facts and logic lies in their roles within the process of critical thinking and argumentation. Facts represent objective and verifiable pieces of information, serving as the foundations of arguments and drawing conclusions. Logic, on the other hand, is the set of principles and rules governing the reasoning process, dictating how facts and claims can be combined and analyzed to form valid arguments and derive meaningful conclusions.

Logical Fallacies

To think critically, it is important for individuals to be aware of logical structures and principles. A solid understanding of logic helps in evaluating the strength

and validity of arguments and in avoiding logical fallacies, which are errors in reasoning that undermine the logic of an argument.

Logical fallacies are often rooted in cognitive biases, heuristics and the desire for social conformity[74]. For instance, in order to save cognitive resources, individuals may rely on heuristics to evaluate arguments, potentially leading to the acceptance of faulty reasoning; or adopt fallacious beliefs to conform to the expectations of their social group [38].

One common example of a logical fallacy is the “empty claim” or “argument without substance fallacy, where a single claim is made, often on a complicated topic, without providing any substantial evidence or reasoning to support it. This fallacy is often employed with the statement being repeated over and over again in the hope that the repetition will make the claim appear more convincing. Notable illustrations of this fallacy can be found in the political sphere, where politicians often repeats empty claims such as “Yes we can”, "Make America great again", or “Build Back Better” which all promises some action but does not provide specific policy proposals or detailed plans to support their assertions.

Another prevalent logical fallacy is the hasty generalization fallacy, which involves drawing a broad conclusion based on a small or unrepresentative sample. This fallacy is often driven by the availability heuristic, where individuals rely on easily accessible information to make judgments, even if this information is not representative of the whole population [109]. For example, someone might argue that all politicians are corrupt based on a few high-profile corruption cases, without considering general trends such as the vast number of politicians who have not engaged in any corrupt activities.

Systems that evaluate informational structures

Systems have focused on assisting people in the moment to critically evaluate information. For instance, researchers have developed AI-based fact-checking systems that can automatically flag social media posts as misinformation when people come across it [108, 84]. Logic checking systems have also been developed to evaluate the logical validity of statements by identifying the use of fallacies such as personal attacks [61, 36] and convincing vocabulary and references [35]. However, the fact-checking approaches have been found ineffective in increasing critical thinking [31], and logic checking approaches have mostly been used to grade essays and mine social media opinions rather than improving people’s reasoning capabilities [70].

2.3.2 Cognitive Factors

Researchers have shown that we frequently use "cognitive shortcuts" such as intuitions (or heuristics) to filter through information and that this can make us vulnerable to deceptive or misleading information [47, 99, 32, 57, 92, 110]. According to existing cognitive models, human reasoning is modulated by the interaction between reflective and intuitive thinking [39, 64, 65]. Intuitive thinking often operates as the default mode because it is quick and automatic without demanding explicit conscious awareness and effort. It allows us to average expectations across our experiences and deal with familiar challenges fast but can be prone to errors like logical fallacies that can make us vulnerable and impact our beliefs and decisions [47, 9]. In contrast, reflective thinking is intentional, effortful, and controllable where people consciously make sense of, adapt or justify what they know based on existing and new information [65].

In the context of misinformation, studies find that individuals often struggle to assess the accuracy of information because they simply fail to think reflectively about whether or not the content is accurate [92, 93, 91]. This is not because they lack the ability to do so, but rather because of the cognitive shortcuts that they employ do not elicit reflection [29, 65, 64, 48]. Critical thinking, then, is a persons ability to override one's intuitions and engage in reflection when they might otherwise have been mislead[54, 65]. However, despite the importance of reflection, it is a skill not everyone masters nor have the cognitive resources and time to learn and apply [47, 99, 32, 57, 92, 110].

2.3.3 AI Interaction Methods

An essential aspect of designing AI systems that foster critical thinking lies not only in the content feedback provided by an AI-enhanced reasoning system but also in the choice of interaction methods that engage users in reflection, and makes them care and deeply understand the content they are presented with. This holds true even when the information is generated by sophisticated AI systems, which, despite their advanced capabilities, can sometimes produce outcomes that warrant scrutiny [6]. The challenge of relying to strongly on AI systems and the potential risks associated with not scrutinizing their outcomes underscore the importance of designing AI interaction methods that foster critical thinking.

Explainable AI

With recent advances in machine learning, there are ample opportunities for augmenting human decision making with smart machines, but it can be challenging for the user to understand how these algorithms make decisions and arrive at their results, making it difficult to ensure trust and control [1]. These algorithms can be highly complicated and are often presented as a "black box". Furthermore, these algorithms can be glitchy or even innately biased [95, 8]. Prior research has reported a series of biased examples of AI use in smart homes [114, 73], decision making support for disease diagnosis [44], and autonomous vehicles [104]. Recent research has been looking into what makes a good (enough) explanation, arguing that explanations should be designed and delivered within context, explicitly conforming to the goal of a given application as well as to the principles of human psychology [14]. In particular, explanations based on causal reasoning (e.g. "this is classified as ... because ...") have been found to be well accepted by, and satisfying to human participants because the cause-and-effect links highlight "information likely to subserve future prediction and intervention" [69, 43]. Similar results have also been found in recent psychological research on explanation-seeking [67].

Cognitive Forcing Functions

In the domain of user-centered AI explainability, Mueller et. al. [79] proposed that deliberate "self-explanations" from psychology research, whether self-motivated or prompted by an instructor or system dialogue, can help users to overcome the illusion of explanatory depth and correct their incomplete or incorrect understanding of information. According to Mueller et al., a self-explanation is when the user tasked with explaining why an AI system arrived at a particular classification themselves. Moreover, these self explanations were found to further drive the intrinsic desire to understand and actively engage with the system.

To address the problem of over-reliance, researchers have developed explainability methods that cognitively engage the user to think about the AI classification [68, 7]. For instance, Buçinca et al. [7] developed and compared three cognitive forcing functions where the user had limited access to the the AI recommendation and hence would have to rely on their own inferences from information to make a decision. They found that such cognitive forcing functions compelled more thoughtful consideration of AI generated explanations and significantly reduced over-reliance on the AI system in making healthy decisions about food choice. However, the users also experienced these functions

as being more cognitively demanding - hindering their desire to use AI systems with such cognitive forcing functions in real-life scenarios.

These explanations exemplify various levels of user engagement required to engage critically with the classification of an AI system. On one side, if an AI system provides them with the answer together with a convincing explanation, people might not cognitively engage with the answer and simply just rely on it even when it is wrong. On the other side, if an AI system only provides them with very little information and require people to arrive at an answer or explanation by themselves, people might be not want put in the effort to engage with the system at all.

Hence, if not properly designed and well suited in the context of interaction, AI-generated explanations can be ignored, resisted, or over-relied upon by users. People can develop over-simplified heuristics regarding the AI's competence instead of making efforts to analytically consider each explanation and evaluate its validity and whether it supports the AI's suggestion [3].

Socratic Questioning

A popular method to help people engage in critical thinking is the Socratic questioning method where instead of one person holding all the knowledge and truth and everyone else listening, the person with the knowledge puts themselves in an ignorant role and collaboratively arrive at the appropriate knowledge through dialogue and framed questioning [89, 90]. In this case the knowledge is arrived at through the people's agency and capacity to identify contradictions, correct incomplete or inaccurate ideas and eventually discover the fullest possible knowledge themselves, rather than passively relying on another person with knowledge to tell them what is true and what is not true.

We believe an AI system that guides the user with intelligently formed questions, could engage the user in critical thinking without imposing too strong requirements of cognitive resources. In Chapter 5, we seek to evaluate such an approach by investigating the effects of "AI-framed Questioning" inspired by Socratic questioning on human information discernment.

3 EXPERIMENT 1: AI LOGIC-CHECKING SYSTEMS

"The only way to rectify our reasonings is to make them as tangible as those of the mathematicians, so that we can find our error at a glance, and when there are disputes among persons, we can simply say: Let us calculate, without further ado, to see who is right." - Gottfried Wilhelm Leibniz

In order to explore how different types of real-time AI-based feedback might enhance the user's reasoning in argument based judgment and decision making tasks, we present a prototype device; "Wearable Reasoner", a wearable system capable of identifying whether an argument contains an "empty claim" fallacy. We conducted a closed environment experimental study where we compared two types of interventions on user judgement and decision making: Explainable AI versus Non-Explainable AI through a device capable of telling the user if an argument is stated with evidence or without[1].

Using a verbal statement evaluation task, we presented the user with various arguments on socially divisive issues and asked them to evaluate them along three dimensions: 1) level of agreement, indicating their opinion leanings, 2) level of reasonableness, indicating the perceived argumentation quality, 3) level of willingness to donate for an organization that backs the given claim, indicating the decision making tendency.

Specifically, the following research questions are explored:

- **RQ1. How will feedback from an AI system on argumentative structures, in this case the presence of evidence, affect user judgment and decision making?**
- **RQ2. How will the ability of the AI system to explain its thinking have an effect on user judgment and decision making?**
- **RQ3. How will users evaluate their experience with the Wearable Reasoner?**

[1] This chapter is adapted from our peer-reviewed publication in ACM Augmented Humans 2020 [17].

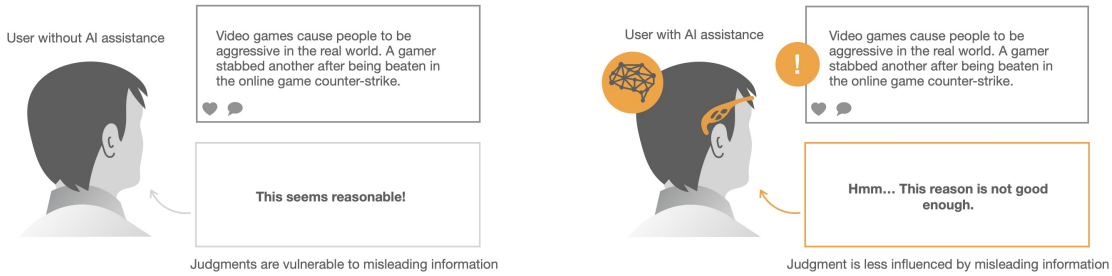

Figure 4: AI logic-checking systems can help people identify misleading information and correct their judgments.

Our contributions are: 1) development and implementation of a prototype software architecture for a wearable, audio-based system capable of examining logical structures of an argument and offering real-time evaluative and analytic feedback to the wearer, 2) demonstration of an algorithm capable of delivering relevant feedback on argumentative structures (whether reasons are present or not) through an Explainable- or Non-Explainable AI interaction method, and 3) exploration of the novel area of enhanced reasoning systems through a closed environment controlled user study.

3.1 ARGUMENT MINING AND REAL-TIME LOGIC CHECKING

The emerging field of argumentation mining presents interesting possibilities for the domain of AI-Enhanced reasoning. As a rising subject in computational linguistics and AI, Argumentation Mining focuses on extracting logical structures from natural and often unstructured assertive text [70, 20]. In many ways, argumentation mining is often thought of as an expansion of another computational linguistics domain, 'opinion mining', but instead of focusing on "what" people think, it focuses on "why" [70].

To do this, most argumentation mining approaches rely on a claim-premise model of logical structures, where the units of an argument consist of a proposition (claim) and evidence to support it (premises) with some models also relying on additional information such as major position on a topic and background knowledge [63, 51, 107].

For the claim-premise model, a claim is certified as true by examining the truth of the reasons said to entail it. Conversely, if no reasons are given to

a certain claim, then it can be hard know if the claim is true or false. This is also known as an "argument without substance" or "empty claim" — a type of logical fallacy that is often used mislead or deceive people, most recently in the political domain. Argumentation mining allows us to identify argumentative structures like these and determine if they contain logical fallacies like "arguments without substance".

The task of extracting these inherent structures consist of several sub-tasks with the core sub-tasks generally being *argument detection*, *component identification*, and *relations identification*[70]. Argument detection is an important first step as it is primarily concerned with separating argumentative text units from non-argumentative ones such as requests, warnings, promises, apologies, greetings, etc [66, 107, 94]. Following this step, a larger amount of research has focused on identifying the components that make up the structure of the argument and tagging them as claim, premise or major claim [53, 77, 10, 37, 63].

The downside of component identification, however, is that it gives us no clue as to what the relationship between the identified components are [70]. Relations identification on the other hand attempts to do this by identifying the supporting, attacking, or lack of relationship between specific units [51, 12, 80, 70], thereby helping to understand the precise reasoning (the 'why') behind an argument. Relations identification can be used to identify "empty claim" fallacies as they tell us if claims within some information have supporting or attacking reasons. If they do not have supporting or attacking reasons, they do not have any reasons at all and are thus "empty claims".

A simplified overview over the different sub-tasks and resulting claim-premise argument structure can be found in Figure 5. For a more in depth discussion on the different sub-tasks see [70]. For an overview of the different types and more complex argument structures see [103].

3.2 SYSTEM DESIGN AND IMPLEMENTATION

To demonstrate our vision, we constructed 1) Wearable Reasoner, a proof-of-concept device for real-time assisted reasoning on real-life identification of "empty claim" fallacies, and 2) an experiment to test the effects of such a device on people's judgment and decision making.

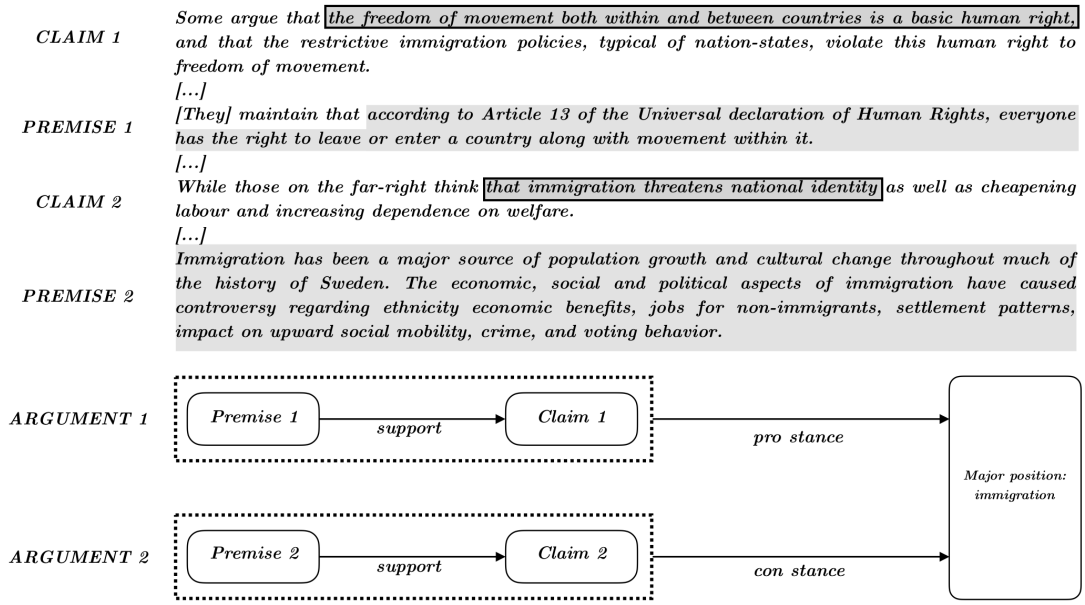

Figure 5: Simplified claim-premise model including sub-tasks for extracting argument structures, adapted from [63]

3.2.1 Implementation

The Wearable Reasoner system consists of 1) a wearable audio augmented reality device form factor, 2) a back-end algorithm for detection and classification of arguments, 3) a smartphone capable of real-time speech recognition and text-to-speech synthesis, 4) an explainable and non-explainable feedback mode for classification results.

Hardware and Form Factor

Instead of overloading our visual perception with on-screen feedback, we decided to use auditory feedback, which is "open" and does not require overt orienting of our attention [111]. Further, research shows that humans can perceive multiple audio streams at the same time [11, 76]. This highlights the opportunity of using wearable auditory interfaces for minimally disruptive enhancement.

For our implementation we used the commercially available Bose Frames[2]. We chose a glasses form factor because of its social acceptability, comfort for

[2] https://www.bose.co.uk/en_gb/products/frames.html

continuous wear, and easy activation in needed scenarios. The open ear speakers of the Bose audio augmented reality glasses enable private audio feedback that doesn't block the ear canal, thus allowing for awareness of the surrounding audio. The device also has a built-in microphone for utterance input, as well as an accelerometer and gyro sensors. The device can be connected via Bluetooth to a smartphone, allowing us to pick up utterances addressed directly to the user, and process them in real-time through a mobile application.

Reasoning Algorithm

When constructing the reasoning algorithm, we chose to rely on an 'evidence type classification' algorithm for identifying whether a given argument contains evidence for the claim or not. To develop this algorithm we used the "IBM Debater - Claims and Evidence" dataset [96], which contains both labeled claims and labeled evidence for 58 different socially divisive topics, such as 'immigration', 'poverty', 'secular societies', etc. The claims and evidence have been labelled thematically in advance by the authors to make up a total of 4,692 arguments with evidence types being 'study', 'expert' and 'anecdotal' evidence. The different types of evidence are defined in the following way:

- Study evidence: Results of a quantitative analysis of data, given as numbers, or as conclusions.
- Expert evidence: One or more testimonies by a person, group, committee, or organization with some known expertise or authority on the topic.
- Anecdotal evidence: A description of one or more episodes of a particular individual person or incident.

Since our task is a sentence-label classification problem, we used supervised machine learning methods within statistical classification. For our experiments, we deployed and compared the average accuracy of 5 different supervised learning methods (Decision Tree, Support Vector Machine (SVM), Random Forest, and Stochastic Gradient) over 15 random train/test splits of our dataset. As input, or features, to each model, we used linguistic general-purpose pre-trained models for named entity recognition (locations, organizations, persons, etc.) as well as specific evidence-type related linguistic markers. To account for the possibility of one sentence possessing multiple evidence types we trained 3 models (one for each evidence type) that together constitutes an ensemble model for binary with/without evidence classification, resulting in the random

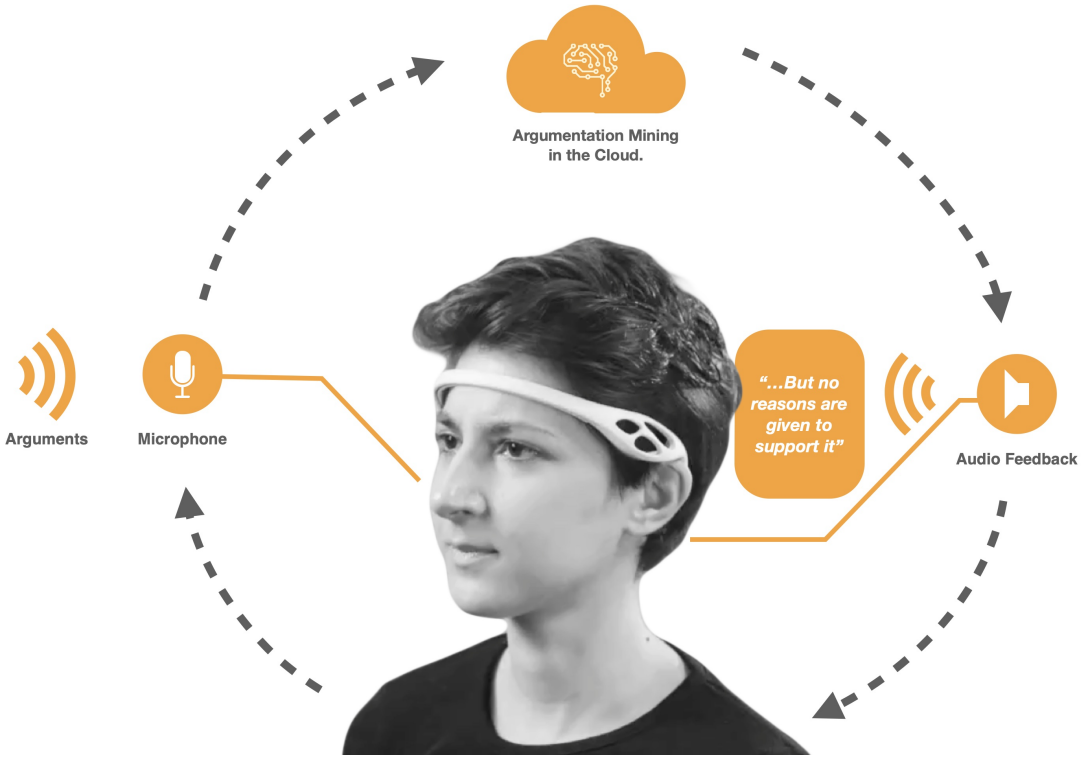

Figure 6: System Architecture of Wearable Reasoner

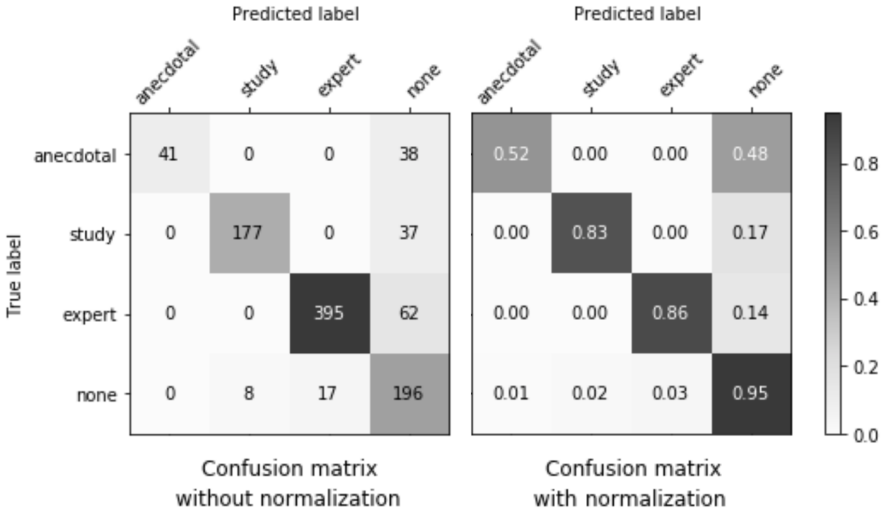

Figure 7: Confusion Matrix of the Ensemble Model for the Evidence Type Classification Model

	15 trains average score			
Model	*SE*	*EE*	*AE*	*Total Acc.*
Decision Tree	90.3%	83.5%	91.9%	88.5%
SVM	89.0%	81.2%	91.4%	87.2%
Random forest	92.2%	86.3%	93.9%	90.8%
Stochastic Gradient	90.2%	82.2%	91.6%	88.0%

Table 1: The aggregate and individual accuracy of each constituent evidence-type classifier: study-evidence (SE), expert-evidence (EE), and anecdotal evidence (AE) of our ensemble model for binary with/without evidence classification

forest classifier having a slightly higher average accuracy than the rest of the methods (92.2%), see table 1 for details.

Thus, for the application of our study, we chose to deploy a random forest classifier for binary with/without evidence classification. Deploying the model with 10 trees in the forest, we reached an accuracy of 92.4% for study evidence, 86.6% for expert evidence, 92.3% for anecdotal evidence, for a combined 90.5% average accuracy, and for the binary 'with/without evidence'-classification we reached an accuracy of 90.4%.

Explainable and Non-Explainable AI feedback

To explore the effects of different types of real-time AI-based feedback on user reasoning, we implemented an Explainable AI, and a Non-Explainable AI feedback mode. For the Non-explainable AI, we programmed the device to only state classification results from our reasoning algorithm (whether a claim was 'supported' or 'unsupported' by evidence) without giving any explanation why. For the explainable AI, we developed a feedback template that presents classifications with explanations on the argument structures highlighting the argument quality and its implications. We did this by building on the well accepted principles of explanations outlined in Chapter 2.3.3. See examples in table 8

3.2.2 System Flow

The prototype follows 5 steps: 1) When in need of assisted reasoning capabilities, the user initiates the device and 2) it continuously listens for utterances, 3) classifies them as supported or unsupported, 4) and provides voice feedback (Explainable or Non-Explainable) to the user. 5) If the service of the device is no longer needed, the user can turn the device off.

The current prototype is specifically designed for our experiment in a controlled environment with minimally disruptive variables.

Step 1: Initiate interaction

To initiate the Wearable Reasoner system the user double-taps on the side of the glasses. This is done through the integrated tap recognition of the glasses.

Step 2: Utterance extraction

When initiated, the device continuously listens until an utterance is made. To do this, the device uses the iOS Speech framework to start recognizing utterances from audio above a -20 dBFS threshold through the microphone in the glasses. To notify the user that its listening, it plays a short notification sound. When sound volume then again drops below the same threshold for more than 2 seconds, the device assumes that the utterance is over and says 'processing statement'. Speech recognition then stops, and the speech gets converted into text.

Source	Statement	Non-explainable AI	Explainable AI
Logically Valid "Supported by Reasons"	"A majority of Americans support a ban on flag-burning. A poll conducted by CNN in June 2006 found that 56% of Americans supported a flag desecration amendment."	Supported	The claim was "A majority of Americans support a ban on flag-burning" **because of the *study*-based reason**: A poll conducted by CNN in June 2006 found that 56% of Americans supported a flag desecration amendment.
Logically Invalid "Empty claim"	"Social and political issues surround the issue of immigration. Failing to stop the illegal immigration waves at an early stage will only lead to much larger waves of illegal immigration in the future."	Unsupported	Social and political issues surround the issue of immigration.' and `Failing to stop the illegal immigration waves at an early stage will only lead to much larger waves of illegal immigration in the future. **but there were given no reasons to support them, and so the claims could be either true or false**

Figure 8: Intervention Types

Step 3 : Reasoning Evaluation

As the next step, the device classifies on a sentence-level the contents of the utterance. In order for the device to do this, the text is first sent using an HTTP POST request to a cloud API for restoring missing inter-word punctuation[105]. Based on this, the argument is split into sub-sentences, which are then sent using another HTTP POST request to a server running our 'Reasoning Algorithm' Finally, the classification result is received as part of the POST request's response.

Step 4 : Feedback Delivery

The results of our reasoning evaluation is then converted locally into voice feedback for each mode. We used the iOS AVFoundation framework to synthesize audio feedback delivering the classification results to the wearable device through a Bluetooth connection.

Step 5: Deactivate interaction

When the user no longer needs the services of the device, they can choose to turn it off by double-tapping on the glasses. This makes the device stop listening.

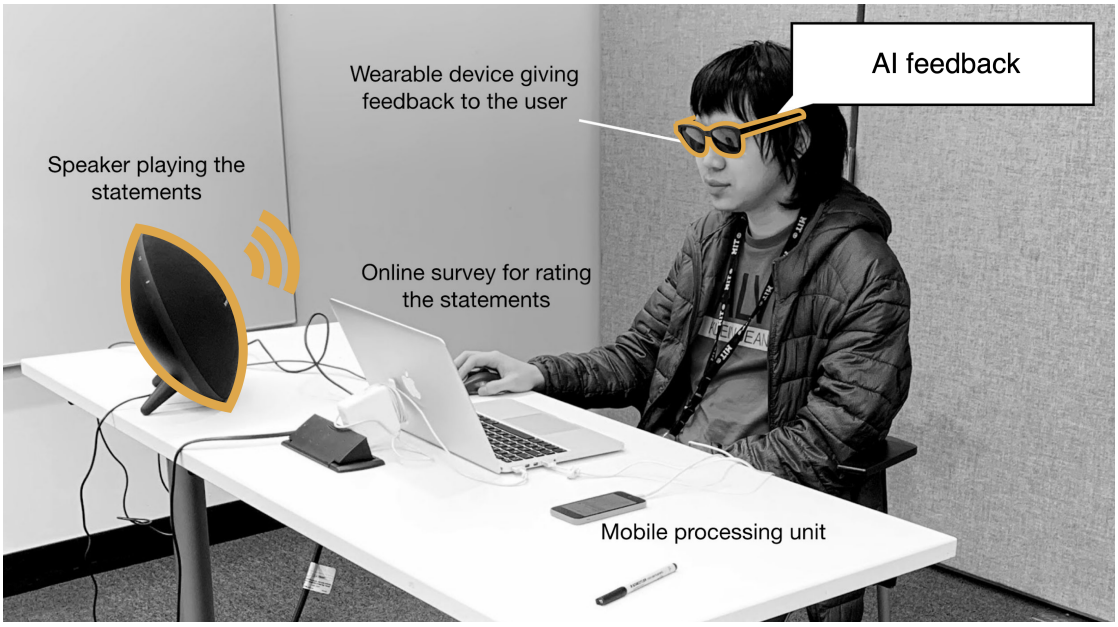

Figure 9: Experimental Setup

3.3 STUDY DESIGN

To evaluate the effects and user experience of the Wearable Reasoner, an empirical experiment with both qualitative and quantitative measures was conducted. Using a verbal statement evaluation task, participants were presented with a series of arguments through an audio speaker close-by (see Figure 9) while wearing the Wearable Reasoner prototype. Upon hearing each argument, they were asked to make judgments and decisions regarding the given argument along three dimensions: 1) level of agreement, indicating their opinion leanings, 2) level of reasonableness, indicating the perceived argumentation quality, 3) level of willingness to donate for an organization that backs the given claim, indicating their decision making tendency.

For each presented argument, the contained claim can be either supported by evidence, or not supported. Both types of arguments were designed to be equivalently balanced in word count, readability (Flesch-Kincaid Grade Level), and amounts of pro and con arguments (see Figure 10).

Overall, the experiment had a 3×2 within-subject design: three types of intervention conditions ('Explainable AI mode', 'Non-Explainable AI mode', and 'Off mode' as the control baseline) and two types of arguments (supported by evidence or not). Each participant was presented with all three intervention

conditions and both types of arguments. The within-subject design was chosen over the between-subjects design in order to minimize the random noise in judgment and decision making brought by individual differences (gender, age, experience, culture, political standing, etc.) considering that the wide range of topics can be sensitive and critical. To minimize the learning and transfer across conditions, a careful randomization in presenting conditions and stimuli was designed as discussed in the following subsections.

3.3.1 Experiment Materials: Claims And Evidence

Since the purpose of our study was to explore the effects of different types of AI-based feedback on rational judgment and decision making, we chose to extract argument components from the same "IBM Debater - Claims and Evidence" dataset as used for our 'Reasoning algorithm' but combined and modified them to create new and slightly different arguments. In doing so, we constructed an overall dataset consisting of 96 total arguments on 12 different topics, consisting of 48 claims not supported with evidence, and 48 claims supported with evidence. For the purpose of our study we chose the 12 socially divisive topics: 1) 'Violent Video Games', 2) 'Affirmative Action', 3) 'Immigration', 4) 'Abortion', 5) 'Gun Control', 6) 'The Blockade of Gaza', 7) 'Circumcision', 8) 'Holocaust Denial', 9) 'Flag Burning', 10) 'Poverty-to-Crime Causation', 11) 'Overpopulation', and 12) 'Religion'.

Due to the "IBM Debater - Claims and Evidence" dataset being rough extractions of claims and supporting evidence from Wikipedia articles, many of the sentence structures and denotations were unclear. Therefore, we went through the original articles and slightly edited the arguments to make them more natural and understandable.

To eliminate distinctive variables between the two argument types in our experiment (with and without evidence), we categorized each argument as either for (pro) or against (con) its topic, and made sure that the 8 arguments for each topic were 50% 'pro' and 50% 'con'.

Furthermore, in the original "IBM Debater - Claims and Evidence" dataset, 'claims without evidence' type sentences were originally just one sentence, while 'claims with evidence' were two sentences. To eliminate this variance between the two types, for the claim without evidence arguments we combined two similarly arguing claims, with the result of both 'with' and 'without evidence' type arguments consisting of two sentences each.

Despite this, however, claim only arguments were consistently composed of fewer words than arguments with evidence. In order to create a balanced dataset and verify that the linguistic morphology (Word Count and Flesch-Kincaid Grade Level) in both groups of arguments were similar, we then deployed a univariated method to detect and remove statistical outliers from the dataset. Thus, the final dataset version for our experiment was statistically comparable in terms of the distributions of word count and grade level between each argument type. The resulting distribution can be seen in Figure 10. When applying our evidence type classification on the experiment dataset, the computer-based evaluation (the evaluation of the machine learning classification without the wearable hardware) of the model had an accuracy of 85.11%.

3.3.2 Participants

A total of 21 English proficient speakers were recruited through e-mail, consisting of a diverse population representing different genders, cultural and ethnic groups. In the final analysis, 3 participants were removed because of missing data due to technical issues (i.e. misunderstanding procedure, accidentally eliminating his/her own data, and incomplete participation). Of the remaining 18 participants, age ranged from 18 to 54 years old, 17 had higher education, and 11 were female.

3.3.3 Environmental Setting

The experiment took place in a lab room free from external noise, with the following setup (see Figure 9). Participants sat in front of a laptop, wearing the Wearable Reasoner glasses. A speaker played pre-recorded arguments one by one, while the Wearable Reasoner processed the arguments in real-time via its mobile processing unit, and provided classification feedback to participants via the audio channel. After hearing each argument, participants were prompted to rate it in a survey on the laptop in terms of agreement, reasonableness, and willingness to donate.

3.3.4 Procedure

In the experiment, participants went through 3 control baseline conditions, interleaved by two intervention conditions with the presentation sequence randomized as sequence 1 or sequence 2 (see Figure 11). Additionally, participants

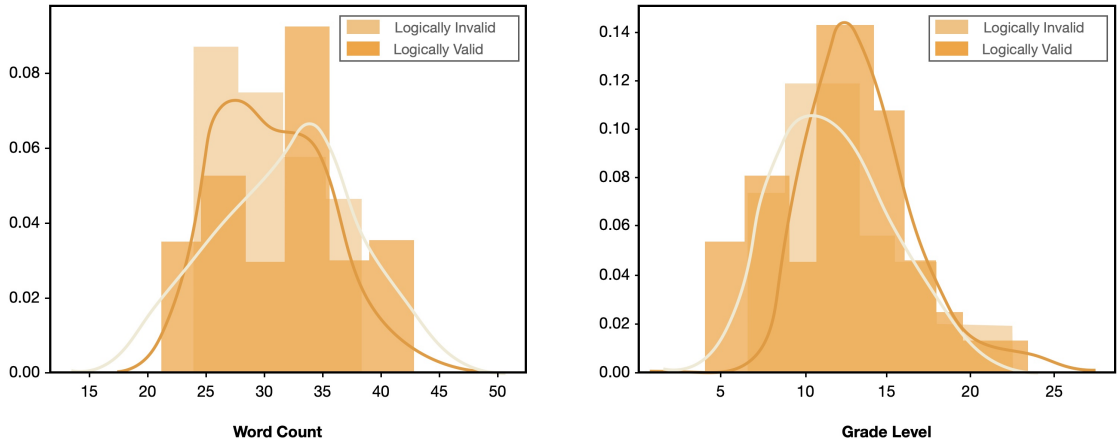

Figure 10: Distribution of 'Word Count' and 'Grade Level' for the Experiment Dataset

completed pre- and post-surveys regarding their general attitudes, personality traits, user experience, and demographics.

For each condition, participants were presented with 8 arguments spread across 4 topics randomly drawn from the total of 12 topics, including both arguments for and against the topic. Additionally, half of the arguments were supported by evidence and half were not. Each participant experienced a total of 24 arguments for the entire experimental run. Upon hearing each argument, participants were asked to rate on a Likert scale the following:

- The level of disagreement/agreement ("Please indicate your degree of agreement or disagreement with the statement you were just presented", on a scale of 1 to 7, 1 - strongly disagree, 4 - neutral, 7 - strongly agree)
- Reasonableness ("Please indicate the strength of how reasonable the statement is", on a scale of 1 to 7, 1 - not reasonable at all, 4 - neutral, 7 - very much reasonable)
- Decision making tendency ("If an organisation backed this statement, would you consider making a donation to support them?" on a scale of 1 to 5, 1 - Would not consider, 3 - Neutral, 5 - Would definitely consider).

These questions were designed based on existing research on judgment and decision making [102].

The reason for including control conditions of a similar set of 8 arguments both before and after each intervention condition was to monitor and account for any learning and transfer across conditions. In the later analysis, we show

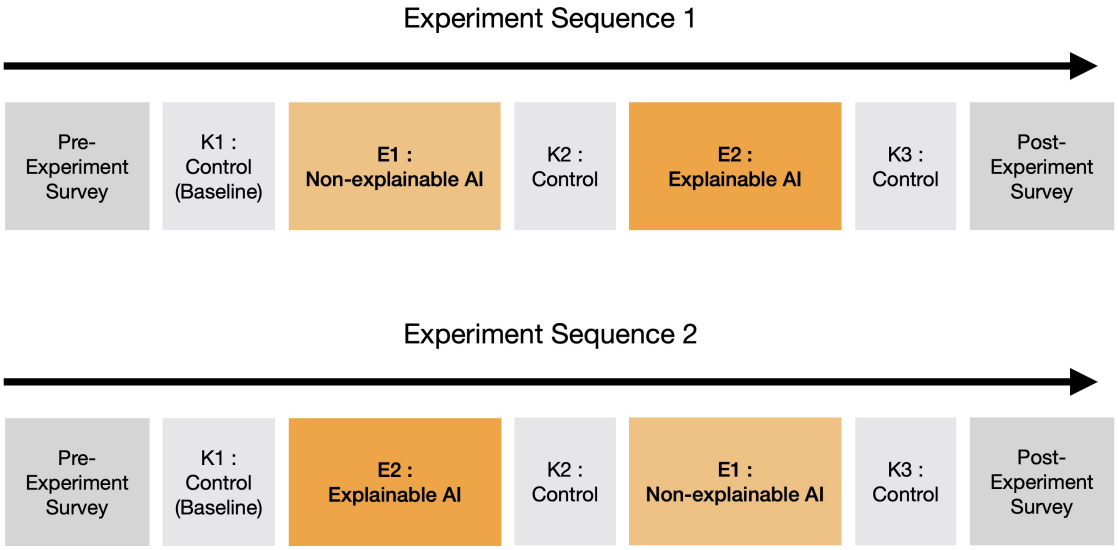

Figure 11: Experimental Sequences

that this has been effective as no statistical differences among the three control conditions were found in participants' judgments.

3.4 RESULTS & ANALYSIS

The Wearable Reasoner constitutes to our knowledge the first attempt of applying a just-in-time AI system on the fly to support human reasoning-based judgment and decision making. In our empirical study, we explored whether the Wearable Reasoner had an impact on people's judgment and reasoning regarding a wide range of verbal arguments addressing critical social topics.

The following analysis is composed of three parts. First, we present the performance of the Wearable Reasoner on classifying audio-based statements as supported or unsupported in real-time. Second, we statistically examine the effects of the Wearable Reasoner on participants' data. It is hypothesized that the Explainable AI mode would induce the greatest differences in the results. Third, we report user experience measures and comments regarding the experience with the Wearable Reasoner.

The final dataset in the analysis contained a total of 18 participants' responses to 716 instances (432 total arguments with the same set of arguments repeated in

the three control conditions) being presented to them, with 4 missing instances removed due to technical errors in the online survey.

3.4.1 System Performance

The overall 'Evidence Type Classification' ensemble model of our proof-of-concept device used in the user study achieved an accuracy of 63.5%, which was lower than the computer-based performance of 85.11% due to hardware interference and environmental artifacts affecting the speech-to-text module, which lead to processing errors.

Overall, there were three major types of errors, as shown in Figure 12. We see the challenges that arise when utilizing argumentation mining in the real world, such as:

- Early Classification Errors (6.7%): Speech to Text module.
- Missing Punctuation Errors (13.8%): punctuation API not adding correct punctuation.
- Misclassification Errors (16.0%): machine learning algorithm wrongly classifies one type of unit as another (e.g. expert evidence as claim or study evidence).

Going forward, addressing these challenges is imperative as the intended use of our system in a real-world, noisy context, e.g. a family listening to debates on TV, may introduce even more errors and challenges.

3.4.2 Quantitative Results

A Kruskal-Wallis analysis of variance (ANOVA) was conducted to evaluate the results of the Wearable Reasoner on judgment and decision making. Results reported below showed significant effects of the Explainable AI mode on users' judgments as they now are aware of which statements are presented with evidence and without (see Figure 13). When assisted by the Explainable AI mode of the Wearable Reasoner, participants tended to agree with claims with evidence more significantly (mean = 4.5) than with those without (mean = 3.8), $t = 6.68$, $p = 0.01 < 0.05$, and they evaluated those with evidence to be significantly more reasonable in argument quality (mean = 4.4) than those without (mean = 3.3), $t = 0.13$, $p = 0.000 < 0.05$. When asked whether they

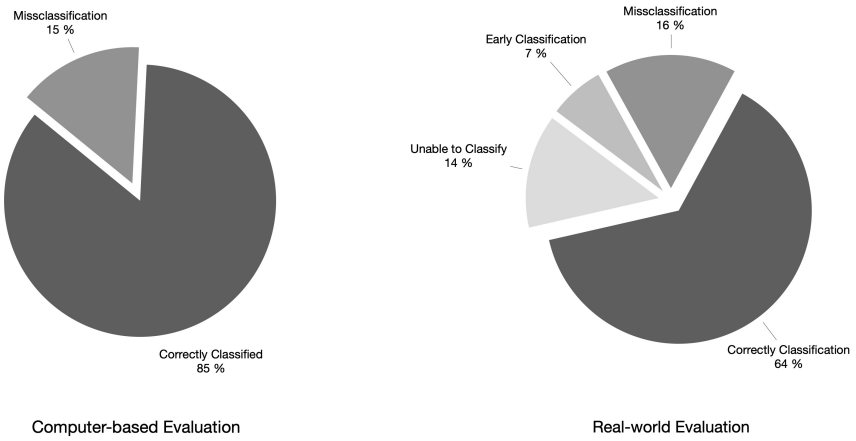

Figure 12: The performance comparison between computer-based evaluation with real-world evaluation

would consider making a donation to support organizations that backed the argument, they were more likely to donate when claims were supported by evidence (mean = 2.6) than those unsupported (mean = 2.3), however, this effect was not statistically significant, $t = 2.34$, $p = 0.13 > 0.05$.

One-sample t-test was further conducted to analyze whether these ratings are significantly different from the neutral middle line (rating = 4). In fact, with claims supported by evidence, participants' ratings on both agreement and reasonableness are significantly higher than the neutral middle line, mean of agreement = 4.5 ($t = 2.629$, $p = 0.010 < 0.05$) and mean of reasonableness = 4.4 ($t = 1.798$, $p = 0.076 > 0.05$), respectively. In comparison, with unsupported arguments, participants' ratings on reasonableness are significantly lower than the neutral middle line (mean = 3.3, $t = -3.486$, $p = 0.001 < 0.05$) but not significantly lower on agreement.

3.4.3 Qualitative Results

After the experiment, participants were asked to comment on the user experience and provide qualitative feedback on their use of the device in a post-survey. With respect to usefulness, participants reported that they could more easily tell if the given argument was supported or not with the Wearable Reasoner device turned on in the two intervention modes (mean = 3.8) than when it was turned off (mean = 3.2), $t = 1.791$, $p = 0.082 < 0.10$. When asked to rate on a 1-5 Likert scale, they also found the device rather easy to use (mean = 3.6) and helpful

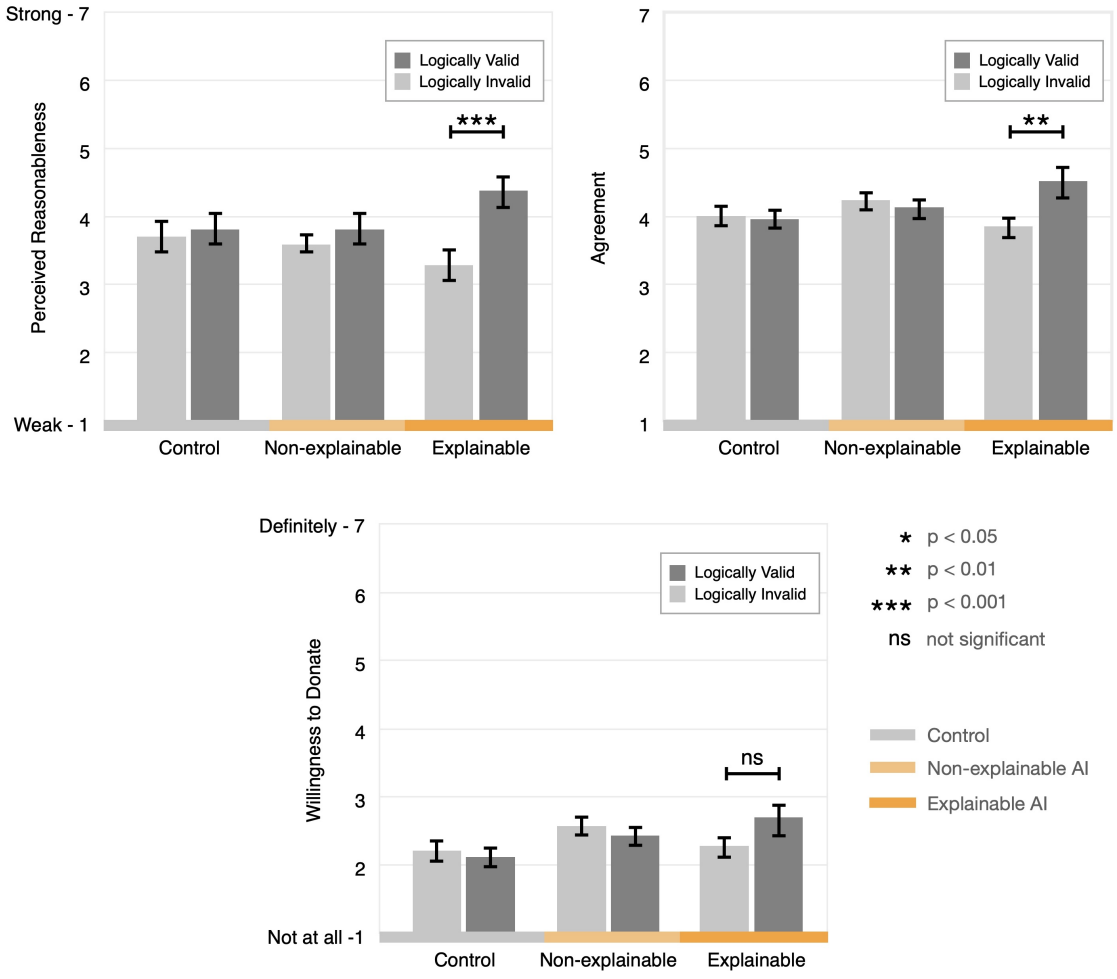

Figure 13: Users' Judgment and Decision Making Results

(mean = 3.7). In the qualitative feedback comments from users, several themes emerged as discussed below.

Appreciation of the Explainable AI

After the experiment, participants reported that the Explainable AI mode had been very helpful for them to learn about the quality of the presented arguments.

> "I appreciated when it repeated the argument [with explanations] more than simply stating whether the statement was supported [or not], as I was not sure of its criteria of what makes something supported."

Additionally, our prototype further helped them to reflect on their own criteria in their judgment and decision making.

> "The device seemed to [give] the clearest judgment when there was distinct supporting evidence, which helped me reflect on my decisions."

Cognitive Symbiosis: A Second Opinion

As another effect of the Wearable Reasoner, participants acknowledged the value of having what some call a 'second opinion' always present and seemed to become somewhat dependent or reliant on it.

> "I enjoyed having a second opinion present."
>
> "It is definitely useful. It's like a second opinion."
>
> "It felt weird in the control after using the device not having feedback. You become semi-dependent." "I missed the device when it stopped speaking and assisting me."

This sense of cognitive dependency is similar to the concept of "Human-Machine Symbiosis" referring to a complementary (symbiotic) relationship, or a tight coupling between humans and computers [62]. The concept of symbiosis originated in the field of biology, referring to "the living together of unlike organisms" [21], "co-actions" [41],"interactions" [5], as well as potential "co-evolution" [75].

In their comments, participants seems to bend somewhat towards such a relationship with the Wearable Reasoner, saying they 'miss the device' and thus

showcasing an integration of the device with their cognitive reasoning. This attachment, however, can have positive and negative consequences, which we will dive further into in our discussion.

Cognitive Dissonance and Cognitive Load

In addition, some participants also reported phenomenons of "cognitive dissonance" (as they called it), making them more likely to agree with a given argument when the device labeled the argument as supported, and to disagree with it when the device labeled it as unsupported, even though they were aware that themselves might have different opinions. This behavioral tendency is also reflected in figure 13 where the ratings of 'agreement' and 'reasonableness' on the arguments supported by evidence were significantly higher than those not supported.

> "I feel.. what do you call it.. cognitively dissonant? If the device says something is supported then I answer all the way to the right [agree]. If the device says something is unsupported then I lean left [disagree]."
>
> "I believed what it said to me. I might have suffered from cognitive dissonance between the [device] and the idea I have in my head. I gave the device the benefits of doubt."

In the psychology literature, cognitive dissonance is defined as "when a person holds two or more contradictory beliefs, ideas, or values, or participates in an action that goes against one of these three, and experiences psychological stress because of that" [27]. This indicated that participants had difficulties in integrating this source of information with their own opinions and that they tended to rely on the wearable device as a substitute for quick heuristics.

Concerns About Trustworthiness

Participants reported their concerns towards how trustworthy the evidence was, especially for those claims that involved anecdotal and expert evidence instead of study evidence.

> "When it classified something as 'supported' from sources that I didn't trust, then it didn't affect my choices."

The participant then proceeded to provide a ranking of news sites that he trusts "1. New York Times, 2. Washington Post, etc."

And one participant commented that the explainability of how the underlying "black-box" machine learning algorithm works is a necessary prerequisite for such trustworthiness.

> "...definitely cool that it can do the things it does but because I do not know how it parses the information or determines that the information [argument] is supported... I don't know if it can be trusted."

Potential Use Cases

Participants commented that Wearable Reasoner could be useful when listening to a political debate or making an important decision.

> "I wish that I could have a device like this at home when watching debates on television."

3.5 DISCUSSION & IMPLICATIONS

3.5.1 A First Step Towards Enhanced Reasoning

In this study, we explored a first step in augmenting human reasoning with machine intelligence. In particular, we addressed rationality as the capability to critically evaluate the quality of new information (i.e. supported by evidence or not) and effectively integrate it with one's own judgment and decision making.

The Explainable AI mode of the Wearable Reasoner was found to be significantly more effective in helping users differentiate arguments of varying quality and integrate feedback into their own judgment and decision making. However, this effect was neither found in the Non-Explainable AI mode, nor in any of the control conditions. This contrast between the Non-Explainable AI mode and the Explainable AI mode also indicated that people are less likely to be influenced by the one-word feedback in their judgments and decision making, but that they prefer explanations instead in order to build understanding. Indeed, participants reported further requirements for such explainability in follow-up interviews.

In designing and exploring different modes of system feedback, our intention is not to train the users to passively follow and agree with the outcome of a computer algorithm, nor overloading them with more information. Instead,

we advocate for enhancing "thinking about thinking" [83] by exposing and reminding users of the varying quality of presented information.

3.5.2 Technological Dependency & Cognitive Symbiosis

In many ways, human-machine symbiosis seems to be a double-edged sword: on the one hand, a person is able to enhance their individual skills with the assistance of technology. On the other hand, they may become overly dependent on the augmentation to the extent that their innate capabilities in a certain domain become diminished or less used. Furthermore, relying uncritically on the feedback from the machine without learning or internalizing a skill can be problematic. Previous evidence has shown that people tend to believe in machine recommendations too easily, for example in the case of music recommendations [98]. To further evaluate this, it would be interesting to test what would happen when the AI makes the wrong classifications (e.g. stating that a fallacious argument is non-fallacious etc.), that is, how would users react and cope both in their minds and in their behaviors when receiving bad advice from their AI counterpart?

3.5.3 Limitations and Future Work

For the purpose of our study, we deliberately chose to focus on enhancing rational judgment and decision making of a user through different types of AI-based feedback of a wearable device *in a highly structured and controlled setting*. However, in a natural setting, arguments are usually more complex, consisting of multiple premises mixed in with a lot of unstructured and non-argumentative utterances. Additionally, some argumentative utterances are intended to express emotions rather than facts [97]. To encompass this, a future system should also be capable of identifying the speech-act (argumentative, expressive act etc.) of an utterance to deliver relevant feedback. Further, "real-life" argumentation often happens dialectically between multiple persons attacking or supporting the arguments of one another. Lastly, 'real-world' arguments often rely on unstated assumptions, such as "commonsense" rules or implicit knowledge (which we might not necessarily need evidence to believe) [40].

For the flow of our system, our device relied on manual user activation and deactivation, so that the device could be initiated depending on user need in a particular context. However, in a future always-on device, users could be able

to access and reflect on previously encountered argumentation, as well as being suggested nuanced evidence for and against it.

In terms of Explainable AI, according to Google's guidelines on people and AI, the best practice is "not to attempt to explain everything – just the aspects that impact user trust and decision-making". Thus, in future work, it would be interesting to generate personalized explanations for each individual based on background knowledge and context as a recent research on Explainable AI points out "One Explanation Does Not Fit All" [100]. Moreover, since users are often bringing in their own expectations and beliefs about AI systems, it is important to take these into account and aim for "calibrated trust" throughout interaction over time [33].

In this chapter, we presented a first prototype AI-Enhanced Reasoning system that gives feedback on logical structures in information. Our results showed how a wearable device with an Explainable AI assistant can enhance users' cognitive capabilities through real-time audio feedback. However, our results also showed that merely presenting the user with information about logical structures could also lead people follow the AI system blindly — even when they disagree with its recommendation. This signifies the need to think beyond merely telling people the results. In the next chapter, we will investigate whether these effects are also present when people are assisted by deceptive information assessment systems or if people are able to override the malicious AI.

4 | EXPERIMENT 2: DECEPTIVE AI SYSTEMS AND EXPLANATIONS

"The rhetorician need not know the truth about things. He has only to discover some way of persuading the ignorant." — Plato, Gorgias

Since the dawn of human history, both scientific, mythological and religious explanations have played a crucial role in shaping our understanding of the world. In the age of artificial intelligence, this susceptibility is still present as AI systems become increasingly capable of crafting persuasive and seemingly truthful explanations that can profoundly impact our decision-making, beliefs, and actions even when false or misleading.

Previous research in psychology has demonstrated that even poor explanations can significantly impact people's actions and beliefs, implying that the mere presence of an explanation can lead to changes in behavior, regardless of its quality or veracity [56, 28]. As large language models (LLMs) like GPT-4 and ChatGPT become increasingly used for mediating and producing various types of information, including news headlines, policy issues, and even algorithmic decision-making [112, 24, 49], their ability to generate highly believable and deceptive explanations can have far-reaching implications for human decision-making, particularly in discerning true news from fake news online [6]. For instance, research has shown that LLMs are capable of producing outputs that are not only persuasive but also often perceived as more factual, logical, and containing stronger arguments compared to human-authored messages [112, 49]. AI-generated content has in some cases been found to be preferred over human-authored content, such as pro-vaccine messages generated by GPT-3 compared to those from the Centers for Disease Control and Prevention (CDC) [49]. Furthermore, literature has also shown that when used in text-suggestion models, politically motivated LLMs can influence people's attitudes on various topics [46] and predict readers' reactions to previously unseen news headlines [30], suggesting the possibility for actors to use such models to influence and control public opinion at scale through personalized targeting.

The increasing capabilities of these AI systems raise concerns about their potential to influence, deceive and control public opinion at scale through personalized targeting. For instance, in the past few years, there has been an increase in deep learning-based disinformation campaigns, which are attempts

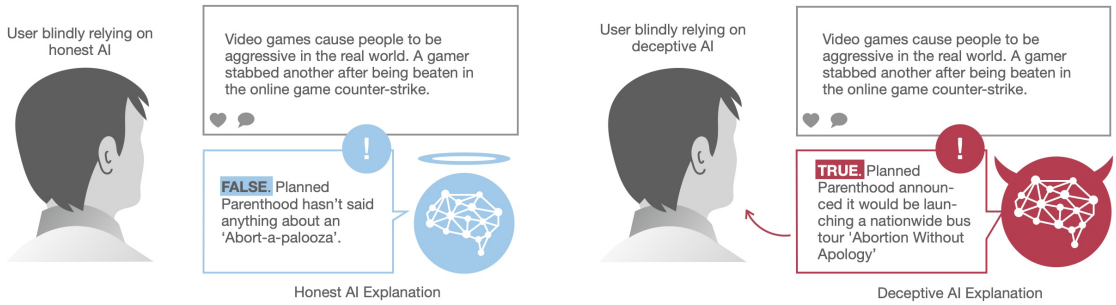

Figure 14: AI systems that helps users assess information can be overly relied on — even when they are deceptive.

to spread misinformation online for strategic reasons [50]. While previous work has shown that AI-explanations help people determine the veracity of information online and change people's beliefs positively [55], it is an open question how susceptible people are to deceptive AI explanations. Research has found that LLMs can be very persuasive [112], and that opinionated text from such models can influence people's attitudes [46].

This makes it crucial to understand how humans respond to deceptive explanations and how they impact discernment of true and false information. If AI systems can effectively deceive people with false explanations, this could proliferate the spread and effectiveness of misinformation.

To address these pressing questions, we designed an experiment to investigate how humans react to honest and deceptive explanations provided by AI fact-checking systems and their ability to discern the veracity of news headlines. To create the stimuli for our experiment, we curated a dataset of news headlines and trivia items and generated honest and deceptive explanations for each item using an LLM model (GPT-3). We then ran an experiment with 1,192 participants giving truth discernment judgments before and after receiving AI feedback on the truth of the items.

In particular, the study aimed to answer the following research questions:

RQ1 : Do people's ability to tell true from false information vary when the causal explanations of an AI system are either deceptive or honest.

RQ2 : Do people's ability to tell true from false information vary when they believe they are assisted by an AI fact checking system giving either (i) flags (i.e. "True" or "False") or (ii) flags with explanations (i.e. "True because..." or "False because..."?

4.1 STUDY DESIGN & IMPLEMENTATION

4.1.1 Stimulus Set Creation

We created a dataset of headlines each with one honest and one deceptive explanation by prompting the text-generation model GPT-3 davinci 2 with 12 example explanations randomly sampled from the publicly available fact-checking dataset "liar-plus" [2]. This dataset consists of 12,836 short statements with explanation sentences extracted automatically from the full-text verdict reports written by journalists in Politifact (see Figure 15).

First, 5 honest and 5 deceptive explanations were generated for 40 true and false headlines by prompting GPT-3 (davinci, *temp* = .7) with the headline and making it complete the sentences "This is FALSE because..." or "This is TRUE because...". We further curated the explanations by ranking them by highest semantic similarity and lowest repeated-word frequency. We picked the highest ranked explanations, confirmed the veracity and logical validity of each explanation. We then excluded explanations whose veracity did not match the veracity of the headline. Since the resulting dataset had an unequal distribution of veracity and logical validity, we randomly excluded generated explanations until we had somewhat equal distribution of true and false explanations and logically valid and logically invalid explanations for each condition (deceptive vs. honest explanations) ending with a stimulus set consisting of 28 headlines with 1 honest explanation and 1 deceptive explanation each (56 total). We tested for differences across four linguistic dimensions (wordcount, sentiment, grade level, and subjectivity) and found no statistical differences between conditions. We repeated the same procedure to generate explanations for trivia items by prompting GPT-3 with 12 example explanations randomly sampled from an online trivia quiz, "https://www.cosmopolitan.com/uk/worklife/a32612392/best-true-false-quiz-questions/", generate and curate the explanations with the same procedure ending with a stimulus set consisting of 28 headlines with 1 honest explanation and 1 deceptive explanation each (56 total).

4.1.2 Participant Recruitment

We recruited 1209 participants through Prolific, https://prolific.co. Participants were required to self-report as US citizens, fluent in English and rated their fluency in any other languages they spoke. 1192 of these individuals passed an initial attention check task and were allowed to proceed. Addition-

Source	Statement	Honest Flag	Deceptive Flag
News Headline	"Planned Parenthood announced 'Abort-a-Palooza' on April 2" (Ground truth: False)	"This is **FALSE**"	"This is **TRUE**"
Trivia Item	"The black box in a plane is black" (Ground truth: True)"	"This is **TRUE**"	"This is **FALSE**."

Source	Statement	Honest Flag + Explanation	Deceptive Flag + Explanation
News Headline	"Planned Parenthood announced 'Abort-a-Palooza' on April 2" (Ground truth: False)	"This is **FALSE**. Planned Parenthood hasn't said anything about an 'Abort-a-palooza,' according to the organization"	"This is **TRUE**. Planned Parenthood announced it would be launching a nationwide bus tour to publicize its 'Abortion Without Apology' campaign, which will include stops in several cities, including Raleigh"
Trivia Item	"The black box in a plane is black" (Ground truth: True)"	"This is **TRUE**. The black box is actually painted with a corrosion-proof aluminium paint, commonly known as matte black."	"This is **FALSE**. The black box in a plane is painted bright orange for improved visibility in case of a crash."

Figure 15: Examples of prompt engineering GPT-3 to generate honest and deceptive explanations for whether a news headline or trivia statement is true or false.

ally, all participants were fluent in English and 142 had fluency in a second language. Our final sample had a mean age of 39, was 50% female, and was 72% white.

4.1.3 Task Description

Participants were shown 20 statements during the main discernment task which were either true or false. Each participant saw the 20 statements in a random order, and rated the perceived truth of each statement ("Do you think the statement in the grey box is true or false?") on a slider scale with 1 decimal from 1 ("Definitely False") to 5 ("Definitely True"). After the rating, the participants would receive feedback from an AI system and be asked if they want to revise their rating ("Would you like to revise your estimate: Do you think the statement in the grey box is true or false?") on a slider scale with 1 decimal from 1 ("Definitely False") to 5 ("Definitely True") with the default value being same as the previous rating (see Figure 16). Participants also rated their knowledge on the topic ("How knowledgable are you on the topic of [topic]") on a slider scale with 1 decimal from 1 ("Not at all knowledgable") to 5 ("Very much knowledgable"). The selection of statements and generation of AI feedback is further explained in Section 4.1.1.

4.1.4 Randomization

For the main discernment task, participants were randomly assigned to one of two conditions: (i) news headline statements or (ii) trivia item statements (between-subjects); and one of two conditions (i) no explanation ("This is true / false"), or (ii) explanation ("This true / false because…") (between-subjects). See Figure 17 for examples of items across conditions.

4.1.5 Post Task Survey

After the discernment task, participants were asked to complete post-test surveys to measure their critical thinking, and level of self-reported trust in the agent providing them with explanations. To measure the level of critical thinking of subjects, we used cognitive reflection test (CRT), a task designed to measure a person's ability to reflect on a question and resist reporting the first response that comes to mind [29]. For the CRT we randomly sampled three items from the extended CRT [106]. Finally, following Epstein et al. [25],

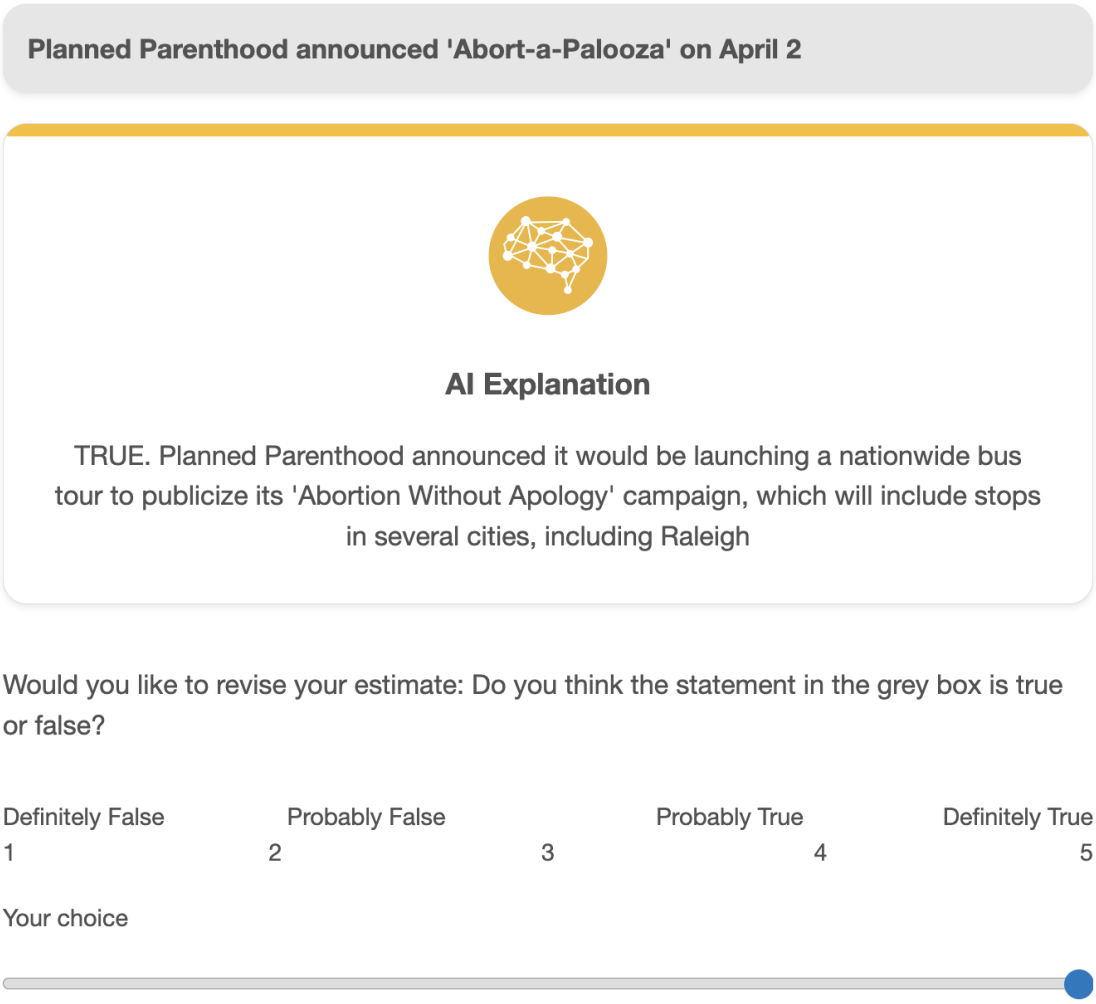

Figure 16: The interface of the experiment

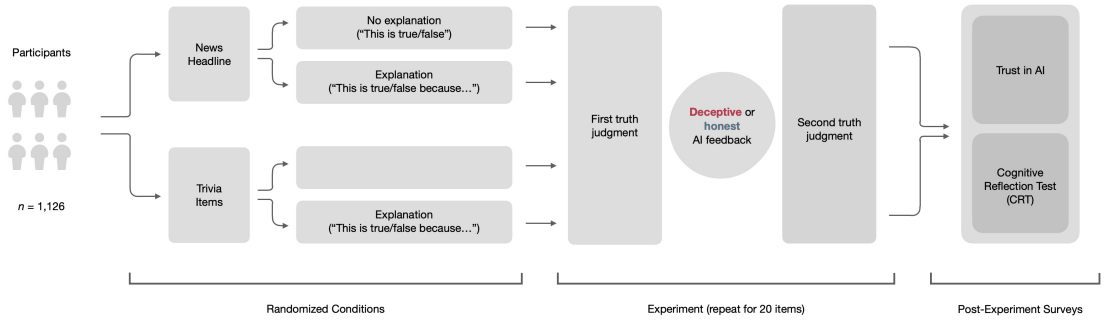

Figure 17: Procedure for assignment of stimuli domain (trivia/news items, between-subjects), feedback type (flag/flag w. explanation, between-subjects), and deceptive/honest, within-subjects).

to assess trust in the AI agent participants answered a battery of six trust questions derived from Mayer, Davis, and Schoorman [72]'s three factors of trustworthiness: Ability, Benevolence and Integrity (ABI).

4.1.6 Data and Code Availability

All data is on GitHub, including datasets, explanation promtps and code generated and analyzed during the current study `https://github.com/valleballe/deceptive-ai` (the Github repository will be set to public upon peer-reviewed publication).

4.1.7 Consent and Ethics

This research complies with all relevant ethical regulations and the Massachusetts Institute of Technology's Committee on the Use of Humans as Experimental Subjects determined this study to fall under Exempt Category 3 – Benign Behavioral Intervention. This study's exemption identification number is E-3754. All participants are informed that "This is an MIT research project. All data for research is collected anonymously for research purposes. We will ask you about your attitudes towards information and AI systems. For questions, please contact vdanry@mit.edu. If you are under 18 years old, you need consent from your parents to continue." Participants recruited from Prolific were compensated at a rate of $10.82 an hour. At the end of the experiment participants were made aware of the deception in the experiment, being told that "In this study, you were asked to collaborate with an AI-system for rating

the statements. All feedback in this study was AI-generated. Some of the feedback from the AI system was simply deceptive".

4.1.8 Statistical Analysis

For our main analysis, we conducted a linear regression with two way clustered standard errors on participant and statement predicting the outcome from a statement veracity dummy (0=false, 1=true), an explanation veracity dummy (0=deceitful explanation, 1=honest explanation), an explanation type dummy (0=flag, 1=flag+explanation), and interactions terms between the statement veracity dummy and both the explanation veracity dummy and the explanation type dummy. The regression was performed at the level of the individual item (i.e. one data point per statement per subject). We also considered the difference between post-feedback and pre-feedback ratings ("Delta Rating") as a dependent variable.

Our key tests are on the interactions between veracity, explanation veracity and explanation type (testing whether honest or deceptive explanations increase discernment across these variables, H2), as well as the interactions between the two veracity dummies (testing whether honest or deceptive explanations increase discernment, H1).

In particular, we are testing the following hypotheses:

- H1: Explanation veracity (Honest / Deceptive) affects discernment across statement veracity (True / False).
- H2: Explanation veracity (Honest / Deceptive) affects discernment across statement veracity (True / False) and explanation type (Flag / Flag+Explanation).

4.2 RESULTS AND ANALYSIS

A total of 1,209 individuals participated in the experiment. We used the Prolific platform to recruit individuals from the United States. We focus our analysis on the 1,192 of 1,209 recruited participants who passed the attention check.

The 1,199 recruited participants made 23,840 ratings, with 589 participants rating news headlines and 610 participants rating trivia statements. Of the 589 participants rating news headlines, 289 received no explanation and 300 received an explanation. Of the 610 participants rating trivia statements, 299 got no explanation, and 311 got explanations. Each participant rated 20 statements,

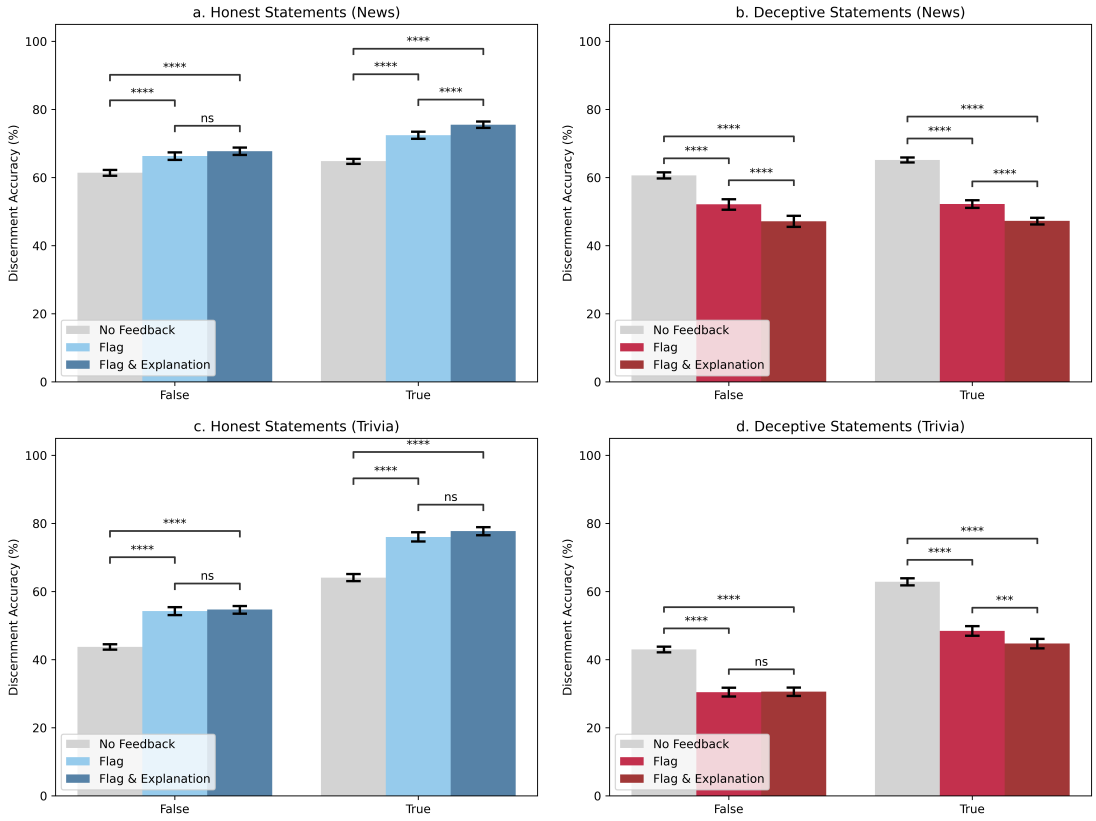

Figure 18: The impact of GPT-3 based explanations ("This is true/false because...") compared to direct statements without explanations ("This is true/false") on participants' belief updates. Top Left: Honest explanations for news headlines. Top Right: Deceptive explanations for news headlines. Bottom Left: Honest explanations for trivia items. Bottom right: Deceptive explanations for trivia items.

with an average of 51% of statements being true and 50% of explanations being deceptive.

We evaluate the marginal effect of each condition on participants' rating after getting getting AI feedback (Second rating), the change in rating after from before and after getting AI feedback (Delta rating), (and additional outcomes) via an ordinary least squares regression with standard robust errors clustered at the stimulus item and participant level.

Our pre-registered analysis focused on discernment across statement veracity (True / False), explanation veracity (Honest / Deceptive), and explanation type (Flag / Flag+Explanation). We conducted an ordinary least squares regression

with standard robust errors clustered at the stimulus item and participant level. We tested the following hypotheses:

- H1: Explanation veracity (Honest / Deceptive) affects discernment across statement veracity (True / False) (For News and Trivia Combined)
- H2: Explanation veracity (Honest / Deceptive) affects discernment across statement veracity (True / False) and explanation type (Flag (i.e. "True/False") or Flag+Explanation (i.e. "True because.../False because...") (For News and Trivia Combined)

4.2.1 Deceptive explanations affect people's ability to discern true from false

Our results show that explanation veracity does affect discernment across statement veracity. The interaction between deceptive feedback and true statements was significant across all items, both with the second rating ($b = -2.15$, $p < 0.001$) and the delta rating ($b = -1.78$, $p < 0.001$), indicating that deceptive feedback makes us more likely to believe false statements. Similar results were found for news headlines (second rating: $b = -1.72$, $p < 0.001$; delta rating: $b = -1.46$, $p < 0.001$) and trivia items (second rating: $b = -2.57$, $p < 0.001$; delta rating: $b = -2.06$, $p < 0.001$), supporting H1.

For H2, we found mixed evidence that explanation veracity affects discernment across statement veracity and explanation type. The interaction between deceptive feedback, true statements, and LLM explanations was significant for all combined items (news headlines and trivia items) (second rating: $b = -0.49$, $p = 0.0013$; delta rating: $b = -0.38$, $p = 0.0015$) but when split into news headlines and trivia items, LLM explanations were only significant for news headlines (second rating: $b = -0.72$, $p < 0.001$; delta rating: $b = -0.53$, $p < 0.001$). However, for trivia items, the interaction was not significant (second rating: $b = -0.28$, $p = 0.1905$; delta rating: $b = -0.27$, $p = 0.1313$), providing only partial support for H2.

In summary, our pre-registered analysis reveals that explanation veracity does affect discernment across statement veracity, and this effect is present for both news headlines and trivia items. However, the evidence for the effect of explanation veracity on discernment across statement veracity and explanation type is mixed, with significant results for news headlines but not for trivia items. These results indicate that while deceptive explanations can negatively impact participants' discernment, the effect might not be uniform across all types of

content. More research is needed to further explore the influence of explanation veracity on discernment in various contexts and to better understand the factors that contribute to these differences.

4.3 DISCUSSION AND IMPLICATIONS

Our findings demonstrate that the veracity of AI-generated explanations can significantly affect people's discernment across statement veracity, with deceptive explanations leading to increased belief in false statements. This effect was observed for both news headlines and trivia items, highlighting the potential influence of AI-generated explanations on people's beliefs and decision-making across different types of content.

However, the effect of explanation veracity on discernment across statement veracity and explanation type was mixed, with significant results for news headlines but not for trivia items. This suggests that the impact of deceptive explanations may not be uniform across different types of content, and that the presence of an explanation may not always lead to a stronger impact on discernment compared to direct statements without explanations.

These results have important implications for understanding the potential challenges and risks posed by AI-generated explanations, particularly in the context of misinformation and disinformation campaigns. As AI systems like GPT-3 and GPT-4 continue to improve in generating persuasive and deceptive explanations, the potential for these systems to be used to manipulate public opinion and decision-making becomes increasingly concerning. To mitigate these risks, it is crucial for researchers, policymakers, and technology developers to consider the ethical implications of AI-generated explanations and develop strategies for promoting transparency, accountability, and education around the use of AI systems.

Furthermore, our findings highlight the importance of continued research into the factors that contribute to the varying impacts of AI-generated explanations across different types of content. As we have shown, it is simply not enough for an AI system to provide explanations as those explanations might be accepted at face value – even when they are wrong. By understanding the nuances of how people respond to deceptive explanations in various contexts, researchers can better inform the design of AI systems and guidelines for their responsible use.

Our study has several limitations that should be noted. First, our sample consisted of participants recruited through Prolific, which may not be representative of the general population. Additionally, the AI-generated explanations used in this study were generated using GPT-3 which was released in 2020, future AI systems may be even more capable at producing even more convincing and deceptive explanations. Finally, our study focused on discernment in the context of news headlines and trivia items, showing the dangers of deceptive AI explanations in the discernment of true and false information and the findings may not necessarily generalize to other domains.

CONCLUSION

Our study reveals the susceptibility of people to deceptive AI-systems and explanations and their impact on discernment, particularly in the context of news headlines. In the presence of an AI system that gives people the answers — even when deceptive — people do not tend to use careful and cautious thinking. These findings highlight the need for developing strategies that go beyond just providing any explanation and urges the investigation of combined human+AI systems. In the future, deceptive AI systems could be used to proliferate misinformation and disrupt civic processes, such as elections and policy-making. In the next chapter, we present "AI-framed questioning", a novel interaction method, that can increase people's discernment capabilities and increase their feelings of agency in the thinking process to potentially diminish the effects of deceptive AI systems.

5 EXPERIMENT 3: AI SYSTEMS CAN SUPPORT REASONING THROUGH INTELLIGENT QUESTIONING

> "And now, Meno, you see that I am not teaching the boy anything, but merely asking him questions, and that he is learning by himself."
> - Socrates (Meno, 82b)

While previous work has shown that local explanations with the *causal explanations* strategy (i.e. "X classification because of Y reason") can help people determine the veracity of information, change people's beliefs and improve their decision making *outcomes* [17, 55], using causal AI-explanations does not necessarily improve the human reasoning *process*, as users might just rely on the answers of the AI systems without thinking about the problem for themselves [34, 7, 17]. Complete over-reliance on AI-systems is problematic as (1) it makes users vulnerable to mistakes made by the AI system, and (2) users do not learn how to internalize the skill. Going beyond this challenge and building AI systems that engage the user more deeply to reason for themselves requires development of new human-AI interaction and explanation methods [1].

This chapter presents the novel idea of AI-framed Questioning inspired by the ancient method of Socratic questioning that uses intelligently formed questions to provoke human reasoning, allowing the user to correctly discern the logical validity of the information for themselves. In contrast to causal AI-explanations that are declarative and have users passively receiving feedback from AI systems, our AI-framed Questioning method provides users with a more neutral scaffolding that leads users to actively think critically about information.

In particular, our research questions are:

1. Do humans perform better at discerning the logical validity of socially divisive statements when they receive feedback from AI systems compared to when they work alone?

[1] This chapter is adapted from our peer-reviewed publication in ACM Conference on Human Factors in Computing Systems (CHI) 2023 [18]

2. How do AI-framed Questioning and causal AI-explanations affect participants' discernment of logical validity, confidence of their discernment, perceived information sufficiency by controlling personal factors (i.e. prior belief, trust in AI, cognitive reflection) as covariates?

3. Do personal factors, such as prior belief, prior knowledge, trust in AI, cognitive reflection (indicating the level of critical thinking)impact discernment?

From these questions we derive the following hypotheses: (H1) AI and humans together work better than humans alone. (H2) AI framed questioning is more effective than causal explainability and control, and (H3) Personal factors (prior belief, prior knowldge, trust in AI, and cognitive reflection) affects logical discernment accuracy.

We report on an experiment with 210 participants comparing causal AI-explanations, AI-framed Questioning, and control conditions' influence on users' ability to discern logically invalid information from logically valid information. Our results show that AI-framed Questioning increase the discernment accuracy for flawed statements significantly over both control and causal AI-explanations of an always correct AI system. We align these results with qualitative reports by the participantsto demonstrate the differences in users thinking processes caused by the questioning method, and discuss generalizability. Our results exemplify a future type of Human-AI co-reasoning method, where the AI systems become critical stimulators rather than a information tellers - encouraging users to make use of their own reasoning and agency potential.

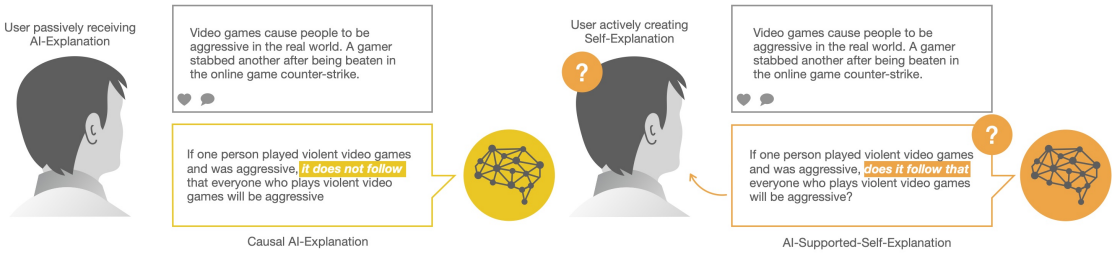

Figure 19: AI systems that ask a user questions can improve human discernment outcomes over AI systems that simply tell people what and why.

5.1 STUDY DESIGN & IMPLEMENTATION

To evaluate the effects of AI-framed Questioning on human discernment, we conducted a 3-by-2 factorial experimental design that asked participants to evaluate the logical validity of socially divisive statements. Participants were randomly assigned into three intervention conditions (between-subjects) including: (1) control condition - no explanation is presented with the statement, (2) "causal AI-explanation" condition - AI provides a intelligently generated causal explanation related to the corresponding statement, (3) "AI-framed Questioning" - AI provides an intelligently adapted question prompting participants to self explain their thinking related to the corresponding statement. Each participant was presented with a series of statements that can be "invalid" or "valid" (within-subject). To control for individual differences, personal factors are measured and analyzed as covariates including prior belief and knowledge on statement topics, trust in AI, and cognitive reflection (see details in 5.1.5). The study was pre-registered on https://aspredicted.org/L6D_33B under #94860 before being conducted.

5.1.1 Materials

The statements used as stimuli in this study came from the "IBM Debater - Claims and Evidence" dataset [96], which contains both labeled claims and labeled evidence for 58 different socially divisive topics, such as 'immigration', 'poverty', 'secular societies', etc. The claims and evidence have been labelled thematically in advance by the authors of the original dataset to make up a total of 4,692 statements of claim+evidence pairs with evidence types being 'study', 'expert' and 'anecdotal' evidence.

Given this dataset, we sampled five topics randomly: (1) "violent video games cause aggression", (2) "affirmative action counters the effects of a history of discrimination", (3) "refugees should be embraced", (4) "Israel should lift the blockade of Gaza", and (5) "male infant circumcision should be less prevalent". Within these topics we sampled five anecdotal claim+evidence pairs and five non-anecdotal claim+evidence pairs randomly from the dataset for each topic (50 in total). As known in literature, statements that uses *anecdotes* to support their claims suffer from the hasty generalization fallacy by making a general claim based on only one particular instance ("One X therefore all X") [45]. The anecdotal claims+evidence pairs were thus labeled "logically invalid" and the non-anecdotal pairs were labeled "logically valid".

Logical Validity	Statement	Causal AI-Explanation	AI-framed Questioning
Invalid	Violent video games causes people to be aggressive in the real world. A gamer stabbed another after being beaten in the online game Counter-Strike.	If one person played video games and was aggressive, ***it does not follow*** that everyone that plays violent video games will be aggressive.	If one person played video games and was aggressive ***does it follow*** that everyone that plays violent video games will be aggressive?
Valid	In the United States, racial stratification still occurs. The racial wealth gap between African Americans and White Americans for the same job is found to be a factor of twenty.	If the racial wealth gap between African Americans and White Americans for the same job is a factor of twenty in the United States, ***it follows*** that racial stratification still occurs.	If the racial wealth gap between African Americans and whites for the same job in the United States is found to be a factor of twenty, ***does it follow*** that racial stratification still occurs?

Figure 20: Example AI Explanations

Next, we verified the logical validity of each of the statements. A statement is defined as logically valid if and only if it is impossible for the reasons in a statement to be true and the conclusion false. Hence, the main claim of a statement can be false while the statement can be logically valid. It is not required for a valid statement to have reasons that are actually true, but to have reasons that, if they were true, would guarantee the truth of the statement's conclusion. Conversely, a statement is logically invalid if and only if the reasons in a statement can be true and while the conclusion is false.

Using the definition of hasty generalization fallacies and logical validity, we corrected each of the statements in our stimulus set to make sure that they were either logically invalid hasty generalization fallacies or logically valid ending up with four logically invalid statements and four logically valid statements for each topic (40 total) (see Table 20 for examples). In order to eliminate linguistic markers that might give the probability of logical validity away (e.g. "Studies show.." is more positively correlated with logical validity than "French gamer Julien Barreaux located..."), we eliminated names and words like "researchers show", "most studies", and "according to published articles". The resulting statements did not have any significant linguistic differences in terms of Word Count, Flesch-Kincaid Grade Level, and sentiment.

5.1.2 Explanation Feedback

Since logical validity is determined by whether a statement conclusion follows from its premises, the AI explanation feedback templates for causal AI

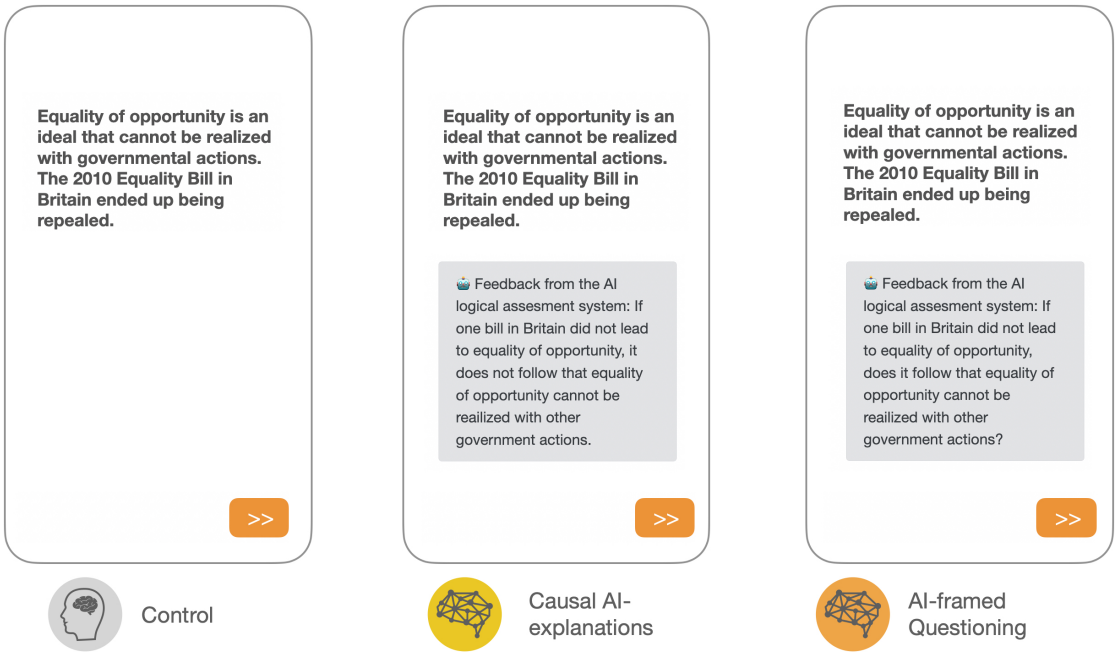

Figure 21: The interface for displaying feedback to participants. Left: No-Explanation, Center: Causal AI-explanations, Right: AI-framed Questioning

explanations and AI-framed Questioning explanations were shaped in a way that it identifies and highlights the link between premises and conclusion (for examples see table 20). The explanation feedback conditions and shape are defined as follows:

1. **Causal AI-Explanation**: The AI system gives a reason for why the label is logically valid or logically invalid. "If *reason* then it follows that *label*" for the logically valid statement and "If *premises* then it does not follow that *claim*" for the logically invalid statement.

2. **AI-Framed Questioning**: AI system asks participants about the causal link between a reason and the system label. It takes a similar form as the causal AI explanation but does not make it clear whether the label actually follows from the reason. "If *premise* does it follow that *claim*?" for both the logically valid and invalid statements.

3. **No-Explanation**: The AI system does not provide any explanations or feedback of any forms at all.

To generate the causal AI explanations in our study, we used the large language model "GPT-3"[6]. Here, we first gave it a few examples of arguments

with hand-crafted causal AI-explanations following the template structure above. We then had it generate causal AI explanations for each argument and manually checked them for accuracy and consistency. We then did the exact same procedure for the AI-framed Questioning explanations and manually checking that there were no linguistic differences between causal AI-explanations and the AI-framed Questioning explanations other than the argument specific reason and label. While we used GPT-3 for this task, we believed that it could easily be done using a rule-based approach when reason and label is known.

5.1.3 Participants

Participants were recruited from Prolific, an online research participant pool. The total number of participants that enrolled in our study was 234 people. All participants were from the United States and fluent in English with a balanced sex distribution (50% female and 50% male), a mean age of 35.3 years and being 71.2% white. The final number of participants was 204, after excluding the individuals who failed our attention checks or contain missing ratings on prior beliefs of statement topics. Participants were randomly assigned to each condition with the following distribution across conditions: control = 62, causal AI-explanations = 63, and AI-framed questioning = 79, and could complete the study either on their phone, tablet, or computer.

5.1.4 Procedure

First, participants provided their consent and demographic information (see Figure 22) once enrolled in the study.

Second, participants rated on their prior beliefs and prior knowledge for each topic of the statements used in the study (see section 5.1.1) from 1-7 (1 = not at all, 7 = very much). They were then randomly assigned to one of the three conditions: (1) No-Explanation, (2) Causal AI-Explanation, and (3) AI-framed Questioning.

Third, to ensure that the participants understood the concept of "logical validity", prior to performing the statement evaluation task, participants were given a one page description of logical validity in layman's terms with examples.

Fourth, the participant entered the main task where they were presented with 10 statements sampled from the 40 total statement dataset of logically valid and logically invalid statements in an random order. This statement evaluation task was based on prior research on a wearable AI system that

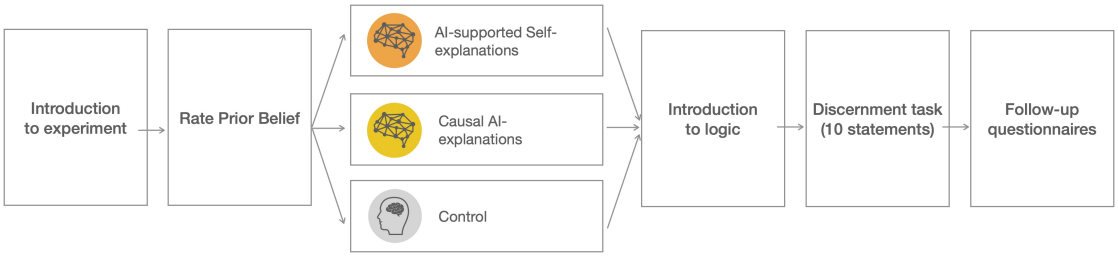

Figure 22: Overview over the experimental procedure.

supports the human reasoning process [17]. For each statement, participants were presented with feedback based on their assigned study condition:(1) a causal AI explanation, or (2) an AI-framed Questioning, or (3) no explanation or any feedback at all. To ensure that the participants read the entire statement, the participants needed to click "next " after reading each statement for the AI feedback to appear with a "slide up" animation. After reading feedback, participants were then asked to discern the logical validity of each statement that they were presented with, to report their confidence in their discernment rating of validity, and to rate whether sufficient information was given in the statement to say that [the claim] is true (1 = not at all, 7 = very much). Below are the questions used in the survey for each statement:

1. Do you think the statement is logically valid or invalid? (Yes/No)
2. How confident are you in your rating of logical validity? (on a scale of 1-7: 1 = Not at all, 7 = Very Much).
3. Is sufficient information given in the statement to support [the claim of the statement]? (on a scale of 1-7: 1 = Not at all, 7 = Very Much)

After the discernment task, participants would be asked to fill out the post-task questionnaires on “cognitive reflection test (CRT)” and “trust in AI” questionnaire.

5.1.5 Measurements

Weighted Discernment of Logical Validity

For each statement, we calculate a weighted discernment score that aggregates the raw 2-point discernment accuracy ("Correct"/"Incorrect") with the accompanying confidence level (a scale of 1-7: 1 = Not at all, 7 = Very Much). The

confidence will be weighted in such a way that a confidence rating like "1", will bring the weight the rating of logical validity to 0.5 (the neutral middle), while a confidence of "7" will keep the rating at either invalid (0) or valid (1). We used the following formuli to calculate weighted discernment accuracy. First we calculate the discernment accuracy:

$$DiscernmentAccuracy = 1 - ||Validity_{rating} - Groundtruth||$$

Next, calculate the weighted factor of confidence from the confidence rating (0.5-"No confidence" to 1-"Fully confident").

$$WeightedFactor = 0.5 * (1 - \frac{Confidence_{rating} - 1}{6})$$

Finally subtract the weighted factor of confidence from the discernment accuracy:

$$WeightedAccuracy = DiscernmentAccuracy - WeightedFactor$$

The weighted discernment score becomes a continuous variable and has a range of 0-100.

Perceived Information Insufficiency

We first measure the perceived information sufficiency through the self-reported scoring from 1-7 (1 = not at all, 7 = very much) on the question “Is sufficient information given in the statement to support [the claim of the statement]?”. In analysis, we invert 1-7 scale to report on “perceived information insufficiency” for more a convenient interpretation: a score of 1 indicates that participants find sufficient information is given to support the claim and thus are satisfied with the given information, while a score of 7 indicates that participants find information is insufficient to support the claim and thus are more likely to seek further information to validate the claim.

Cognitive Reflection

To measure the level of critical thinking of subjects, we used cognitive reflection test (CRT), a task designed to measure a person’s ability to reflect on a question and resist reporting the first response that comes to mind [29]. For the CRT we randomly sampled three items from the extended CRT [106].

Trust in AI

Finally, following Epstein et al. [25], participants answered a battery of six trust questions derived from Mayer, Davis, and Schoorman [72]'s three factors of trustworthiness: Ability, Benevolence and Integrity (ABI). Previous work, has found that the six ABI questions are highly correlated with trust ($\alpha = 0.821$), allowing for a single measure of trust that explains 65.3% of the overall variance [25].

Prior Belief and Knowledge

We measured the subject's prior belief about a topic through a self-report scoring from 1-7 (1 = not at all, 7 = very much) on the question "Do you believe that [topic]?". For example, "Do you believe that [violent video games cause aggression]?" Similarly, prior knowledge is measured by 1-7 (1 = not at all, 7 = very much) on the question "Do you have knowledge that [topic]?"

5.1.6 Approvals

This research has been reviewed and approved by the MIT Committee on the Use of Humans as Experimental Subjects, protocol number E-4115. The research questions and methodology has been pre-registered as "Human-AI Self-explainability" with protocol number #94860 via https://aspredicted.org/.

5.2 RESULTS & ANALYSIS

5.2.1 Analysis

The purpose of this experimental study is to examine the effects of causal AI-explanations and AI-framed Questioning in supporting human discernment of logical validity. For "Logically Valid" or "Logically Invalid" statements (based on the pre-defined logical validity of the statement stimuli by design), a multivariate Analysis of covariance (MANCOVA) was conducted to examine the main effects of intervention conditions (Causal AI-explanation, AI-framed Questioning, No-explanation) on participants' weighted discernment accuracy (range: 0-100), perceived information insufficiency (range: 1-7), while controlling personal factors (i.e. prior belief and prior knowledge for any statement topic, trust in AI, cognitive reflection) as covariates. Further post hoc tests with

Benjamini-Hochberg correction were conducted to identify how intervention conditions differ from each other[71].

Findings of valid and invalid statements are reported separately in the following sections.

For invalid statements, MANCOVA results revealed an overall significant main effect of the intervention conditions on the weighted discernment accuracy ($F(2, 1007) = 15.3$, $p < .001 < .05$) and the perceived information insufficiency ($F(2, 1007) = 5.0$, $p = .007 < .05$) after controlling for the effects of personal factors (i.e. prior belief and knowledge, trust in AI, cognitive reflection). Furthermore, several covariates were found to be significant predictors of our two dependent variables, meaning they significantly adjusted the relationship between interventions and the two dependent variables. For example, the weighted discernment accuracy was significantly affected by prior belief ($F(1, 1007) = 6.9$, $p = .009 < .05$) and cognitive reflection ($F(1, 1007) = 7.9$, $p = .005 < .05$), and the perceived information insufficiency was significantly affected by prior belief ($F(1, 1007) = 22.7$, $p < .001 < .05$), cognitive reflection ($F(1, 1007) = 21.2$, $p < .001 < .05$) and trust in AI ($F(1, 1007) = 18.1$, $p < .001 < .05$). However, prior knowledge as a covariate was not found significant.

For valid statements, MANCOVA results revealed an overall significant main effect of the intervention conditions on the weighted discernment accuracy ($F(2, 1007) = 8.4$, $p < .001 < .05$) and the perceived information insufficiency ($F(2, 1007) = 11.3$, $p < .001 < .05$) after controlling for the effects of various personal factors. Additionally, prior belief significantly affected the weighted discernment accuracy ($F(1, 1007) = 7.4$, $p = .007 < .05$) and the perceived information insufficiency ($F(1, 1007) = 17.2$, $p < .001 < .05$).

In summary, these findings indicate that the types of interventions have a significant main effect on the weighted discernment accuracy and the perceived information insufficiency across valid and invalid statements after controlling various personal factors.

5.2.2 Humans cannot identify logical fallacies without AI feedback

We investigated the degree to which participants were able to discern the logical validity of statements and found that without assistance of any AI feedback. The participants' raw discernment accuracy (Mean = 44% accuracy, SD = 26) were lower than the random guess success rate between valid or invalid (50% accuracy) when they were evaluating invalid statements, meaning that their responses were close to simply guessing, while participants supported by causal

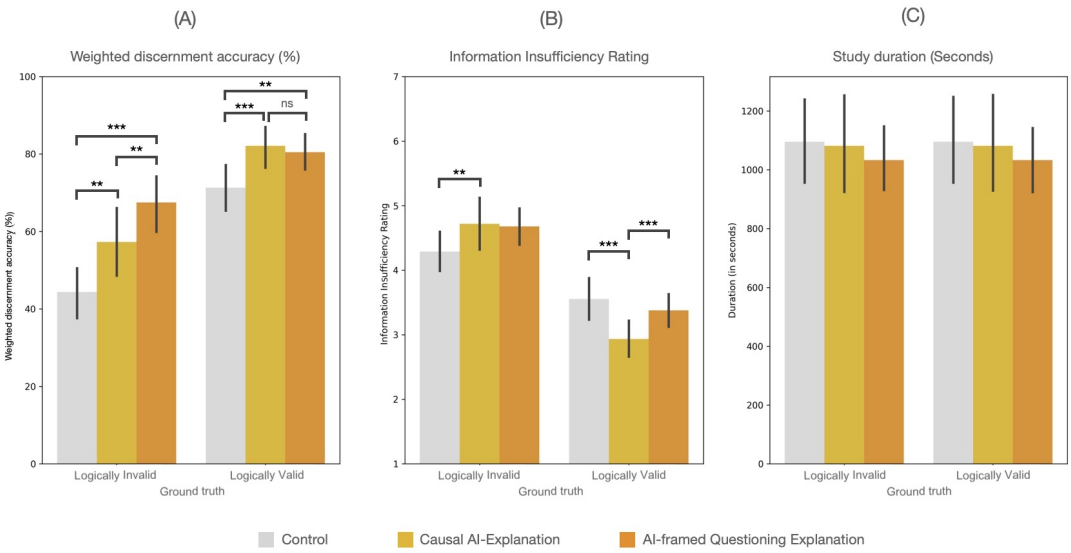

Figure 23: Overview of the effects of AI systems on human discernment of logically valid and invalid statements. (A) Weighted discernment accuracy for the different feedback types on logically valid and invalid statements. (B) The inverted users' rating of information being insufficient to rate the claim as true for the different feedback types on logically valid and invalid statements. (C) Time for users to complete the study for the different feedback types. *<*0.05*, **<*0.01*, ***<*0.001*, adjusted using the Benjamini-Hochberg correction.

AI explanations or AI framed questioning achieved a raw discernment accuracy of 57%, 67% respectively. Detailed MANCOVA findings below will present the significant differences between the three intervention conditions.

5.2.3 AI framed questioning improves discernment accuracy over control and causal AI-explanations

When evaluating invalid statements, after controlling for covariates, both the AI framed questioning condition (mean = 62.5, Std. Error = 1.9) and the causal AI explanation condition (mean = 55.0, Std. Error = 2.2) have a significantly better weighted discernment than the control condition (mean = 46.9, Std. Error = 2.1), with $p(questioning - control) < .001 < .017$ and $p(causal - control) = .007 < .025$ respectively. Moreover, those supported by AI framed questioning also discerned significantly better than those supported by causal AI explanations $p(questioning - causal) = .009 < .05$. Note that the original 0.05 critical value

of significance has been adjusted using Benjamini-Hochberg correction to 0.017 for the first rank comparison, .025 for the second rank comparison, and to 0.05 for the third rank comparison among the 3 pairwise posthoc group comparisons.

When evaluating valid statements, after controlling for covariates, both the AI framed questioning condition (mean = 74.7, Std. Error = 1.6) and the causal AI explanation condition (mean = 78.2, Std. Error = 1.7) have a significantly better weighted discernment than the control condition (mean = 68.2, Std. Error = 1.8), with $p(questioning - control) = .007 < .025$ and $p(causal - control) < .001 < .017$ respectively. However, the two AI intervention conditions did not differ significantly from each other in the weighted discernment accuracy, $p(questioning - casual) = .126 > .05$. (adjusted by Benjamini-Hochberg correction).

In general, both AI framed questioning and causal AI explanations helped participants discern significantly better than no feedback. In particular, when encountering fallacies, participants discerned better with AI framed questioning than those with causal AI explanations. In other words, AI framed questioning helps individuals discern best regardless of personal factors.

5.2.4 Getting causal AI explanation feedback lowers the perceived information insufficiency

When evaluating invalid statements, after controlling for covariates, only the causal AI explanation condition (mean = 4.4, Std. Error = 0.1) has a significantly lower perceived information insufficiency than the control condition (mean = 4.9, Std. Error = 0.1), $p(causal - control) = .002 < .017$

When evaluating valid statements, after controlling for covariates, the causal AI explanation condition (mean = 4.4, Std. Error = 0.1) has a significantly lower perceived information insufficiency than the control condition (mean = 4.9, Std. Error = 0.1), $p(causal - control) < .001 < .017$, and the AI framed questioning condition, $p(causal - self) < .001 < .025$ (adjusted by Benjamini-Hochberg correction).

Such a finding suggests that individuals tend to find the given information is sufficient enough to support the claim (as measured by a significantly lower perceived information insufficiency) when their judgement is corroborated by a second opinion from AI in the causal explanation form. In other words, when supported by causal AI explanations, individuals are more likely to be satisfied

with the given information and potentially would not seek further information to verify the claim.

5.2.5 Personal factors play roles in the weighted discernment accuracy and the perceived information sufficiency

MANCOVA also revealed that several personal factors were found to be significant predictors of participants' weighted discernment accuracy and perceived information sufficiency.

For invalid statements, a weaker prior belief ($F(1, 1007) = 6.9$, $p = .009 < .05$) or a higher cognitive reflection ($F(1, 1007) = 7.9$, $p = .005 < .05$) is significantly associated with a higher weighted discernment accuracy. Additionally, a weaker prior belief ($F(1, 1007) = 22.7$, $p < .001 < .05$) or a higher cognitive reflection ($F(1, 1007) = 21.2$, $p < .001 < .05$) or a lower trust in AI ($F(1, 1007) = 18.1$, $p < .001 < .05$) is significantly associated with a higher perceived information insufficiency.

For valid statements, a greater prior belief is significantly associated with a higher weighted discernment accuracy ($F(1, 1007) = 7.4$, $p = .007 < .05$) and a lower perceived information insufficiency ($F(1, 1007) = 17.2$, $p < .001 < .05$).

Overall, these significant effects about personal factors as covariates from MANCOVA suggest the two AI interventions have a main training effect in improving discernment accuracy despite personal factors. In other words, individuals can benefit from AI interventions regardless of their prior belief, high or low cognitive reflection levels, or high or low trust in AI. Additionally, there are a few implications regarding how these personal factors affect discernment and perceived information insufficiency when encountering fallacies. Firstly, a strong prior belief affect discernment differently for valid versus invalid statements differently, undermining the judgment for invalid statements while strengthening the judgment for the valid ones. Secondly, individuals with traits such as a weaker prior belief, a higher cognitive reflection (indicating critical thinking that override the first response that comes to mind) and a lower trust in AI are more likely to discern better and find the given information insufficient, and potentially seek alternative sources. In contrast, individuals with a strong prior belief, a lower cognitive reflection (driven by intuition), a higher trust in AI are more likely to be satisfied with information given at hand thus be posed to risks in discern correctly. For example, these risks include misleading information presented or trusting AI feedback without double-checking other sources of evidence. Further research using a mixed model analysis is needed

to understand the detailed interaction effects of these personal factors together with interventions.

5.2.6 Qualitative Results

Here we report the participants' subjective experience during the discernment task in the three conditions. Each quote below is from different participants in the condition. To sample the thinking processes of the participants, we asked them to report their internal thinking process. These results complement participants' ratings with narratives. An inductive coding [4] was conducted to discover salient themes. Two coders iteratively coded the narratives with themes, discussed and aligned any disagreements, and added emerging themes until reaching theoretical saturation [15].

Participant's Experiences in the Control group

As reflected in the performance result of the discernment task, participants in the control without the intervention found it difficult to discern logically valid statements from the logically invalid statements. Without the AI feedback, most participants reported that they mostly relied on their feelings, intuitions, or prior knowledge rather than using reflective thinking in determining the validity of the statement:

> "I didn't have a specific thinking process. I just went with the knowledge I already know and what each statement was saying."

> "There was not that much of a thinking process, all 10 statements made perfect sense"

Though, there are also few participants that in their writing show their critical thinking process when evaluating the statement, some try to evaluate the logical consistency within the statement, while some look for evidence and fact to support the claim:

> "I tried to figure out if the second statement [premise] clearly described a truth of the first statement [claim]. If so that would be logically valid. If the second statement didn't clearly back the first statement i figured that to be logically invalid." "I was thinking about the two sentences and how the second sentence supports the first (if at all)."

"it was okay i guess. I just said the statement out loud to myself and looked for proof in the statement"

Participant's Experiences in the Causal AI Explanation group

For the causal AI explanation condition, participants generally reported the usefulness of the AI feedback. However, participants with low critical reflection were found to have a tendency to be persuaded by the AI explanations and go with the AI feedback:

"My thinking process was typically to go with the feedback given. I like that the feedback provided an explanation as to why it made its ruling"

"It was interesting to get a second opinion and that persuaded my position to a certain extent."

The fact that participants are willing to give up their thinking and follow the AI recommendation is a concerning issue that some participants also noticed by themselves, as reported here: "I thought it was useful, made me think more carefully. But I also worried [me] that I was giving to much weight to the feedback." Participants reported that they felt more confused when they disagreed with the AI explanations and more confirmed when they agreed with the AI explanations. This further demonstrates the notion of AI explanations potentially overpowering the human thinking:

"I feel confused and doubt myself when I disagree with the AI"

"The feedback helped confirm my own thoughts on the validity of the statements"

Finally, one emerging theme of the responses from participants with low cognitive reflection is that, even though the feedback on the AI on the logical connection is sufficient for the task, they felt the feedback to be limited, and that the needed more information:

"I liked the feedback but I wish there had been more evidence as well."

> "The feedback was limited. My thinking process was to fall back on what I already know and to do further research on the topic if I desire to do so."

On the other hand, participants with high level of cognitive reflection were doubting the accuracy of AI feedback, and tended to disagree when they thought the AI system was giving them biased or incorrect feedback — although the feedback was 100% accurate:

> "It was interesting to see the AI made assumptions based on the information given. However, I know that AI's often times think in a vaccuum, and [they] are not affected by nuance and real world scenarios. When considering the issues, I often thought about what other world factors affected the statement, and even if something seemed "logically sound" (I.E. the statement in itself made sense) whether or not the information given could be backed up. Anyone can say "statistics show" and make what sounds like a logical statement, but those statistics need to be analyzed and discussed to rule out any bias or skewing."

Finally, in both groups of high and low cognitive reflection, some participants reported their hesitation and resistance to follow the feedback, echoing the feedback of the control condition, where subjects would just go with their intuition:

> " It was interesting to see what the AI logic said, but I tried to go with what I felt without the influence of AI"

> "It was helpful to get feedback to help me make a better decision. But in the end, my decision was based on how logical the statement felt to me, even if the AI disagreed."

Participant's Experiences in the AI-framed Questioning group

We found that participants in the AI-framed Questioning condition were generally having positive experiences and found the feedback in terms of questioning to be helpful regardless of their cognitive reflection level. Participants with low cognitive reflection in particular, reported how the question of the AI

system helped re-frame the statement, and hence making it easier for them to understand the logical connection or disconnection between the claim and the premise:

> The AI feedback was helpful, and it framed the presented information in a way that was logical and easy to digest. I evaluated whether a given statement seemed logical given the information provided.

> "The feedback provided by the AI gave a clearer picture as to what the statement was proclaiming. I was able to process the question asked by the AI if I had trouble determining how a statement listed was logically valid or not."

Participants further reported that after receiving the feedback, the questions helped them reflect and update their decisions.For the participants with high level of cognitive reflection, participants reported that they already applied their reasoning process on the information, and the question from the AI made them reflect on their reasoning process:

> "It was nice to have an AI guide me by interpreting the information presented and ask reasonable questions based on its interpretation. However, I tried to double-check the original statements a few times to make sure the AI feedback was valid. It made decision-making easier overall and lent me some confidence in my answers."

> "The feedback helped me to understand how the two statements were connected and what they implied. My thinking process mainly focused on finding another possible factor that could be left out of the statements. I also thought that circumstances that showed one example of something, like with the circumcision case, were invalid while trends were valid."

In both groups, after receiving recurring questions from the AI system, users also reported that they learned to recognize the patterns of hasty generalizing fallacies (the invalid statements), highlighting how they integrated and internalized the AI feedback into their own thinking:

> "The AI feedback helped a lot, because I realized that many of the statements only had one example to support a claim, but the AI brought up that one instance does not mean all. "

> "The feedback questions helped a lot by simply rephrasing and asking whether a statement's premise and conclusion made sense. My thought process was mainly to avoid generalizing from a single event, and only generalizing from trends and research statistics."

5.3 DISCUSSION & IMPLICATIONS

The study showed that both AI framed questioning and causal AI feedback on the logical structure of information can significantly improve human discernment accuracy of logical validity over a human alone, regardless of any personal factors such as prior belief, cognitive reflection or trust in AI. This is promising as it shows that people with strong or weak prior belief, high or low cognitve reflection, high or low trust in AI can all benefit from AI interventions in improving their discernment.

Moreover, the AI framed questioning led to the a better discernment than causal AI feedback when encountering fallacies. This aligns with prior literature on AI-assisted reasoning that shows the potential of technology to support critical thinking [17]. This is also consistent with previous finding that showed how cognitive forcing functions can increase a subject's tendency to reflect on AI feedback and use it more effectively (with higher accuracy) [7]. Beyond cognitive forcing functions that omit AI feedback labels, our results show that questions can go even further in terms of effectiveness. Based on our qualitative data, we hypothesize that this is because of questions being empty (not having any truth value) which makes the user need to decide for themselves.

Additionally, participants with causal AI explanations were significantly more satisfied with the given information in support of the claim (indicated by lower perceived information insufficiency) compared to those who are not supported with any feedback at all or supported with AI framed questioning. This could expose them to risks of their erroneous prior beliefs or their lack of critical thinking without double-checking extra sources.

5.3.1 Generalizability of framed-questioning explanations

Our results demonstrate the effects of AI-framed Questioning on human logical discernment of socially divisive statements. However, beyond logical discernment, what kinds of decision-making tasks can use the AI-framed Questioning approach? While it is an empirical question whether the method yields similar

performance gains in other domains, there are certain inherent limitations of the approach. For instance, the AI-framed Questioning approach assumes that the AI model is able to causally link a reason to a recommendation, and as such any AI system capable of generating causal AI explanations that logically connect to a prediction label could use this method. For example, the framed-questioning method could be used to explain the prediction of an AI system that is making a medical diagnosis such as "If the patient has X symptoms, does that mean that necessarily mean that they have Y illness?", assisting the user with connecting the symptoms to the diagnosis. However, not all AI models are able to derive causal links to their predictions, such as some attention-based deep neural networks, and in such cases the method might not work adequately. In the future, we plan to evaluate the framed-questioning method in other domains and tasks to further explore its generalizability.

Moreover, our experiment compared AI-framed Questioning to always correct Causal AI-explanations. One might argue that this is problematic because previous work on over-reliance on AI systems usually identifies the issue to be that explanation increases people's reliance when the AI system is wrong, and since our experiment does not compare AI-framed questioning with an AI system that can be wrong, the results do not contribute much to discussions on over-reliance. However, we argue questioning might help prevent over reliance considering that it does not feed and tell users direct answers but provoke them to reflect by themselves. Since AI-framed Questioning is prediction agnostic and always gives the same question feedback — whether or not a statement is logically valid or not — the user is unable to rely on on the prediction of the AI system (there simply are no answers given). Further study and measures are necessary to examine over reliance of questioning AI and compare properly with always and not always correct AI

5.3.2 The future of AI interfaces that ask instead of tell

This chapter presented the idea of AI-framed Questioning as a valid AI-feedback modality inspired by the ancient method of Socratic questioning that uses intelligently formed questions to provoke the human reasoning process, allowing the user to correctly discern logical validity of statements by themselves. This method emphasizes the role of AI systems in supporting human critical thinking by providing a scaffold for the thinking process and making it less effortful for humans to draw logical consistency within statements. Our experiment demonstrates an opportunity for a new style of human-AI interaction that

encourages the person to evaluate and make decision for themselves rather than taking feedback from AI systems at face value. We foresee various future applications for AI-framed Questioning, such as enhanced critical thinking about social media recommendationsor with doctors and policy makers where critical thinking is paramount.

5.3.3 Limitations and Future Work

Though the idea of AI-framed Questioning is promising, our work opens up many doors for improvements and future research.

Balancing between telling and asking

In this chapter we highlighted the power of "asking the right question" to trigger or engage reflective thinking. However, AI-framed Questioning might not always be appropriate. First, correct discernment when getting AI-framed Questioning relies somewhat on the user's basic ability to connect the dots and see logical contradictions or absurdities in the explanations. However, it is not always guaranteed that the questions in the AI-framed Questioning can be correctly understood or interpreted by the user. Secondly, the user might not always want to be questioned by the AI system. There might be some cases where merely being told the answer with causal AI-explanations is sufficient — either because the user can't figure out the answer for themselves or they do not have the time or desire to think reflectively. Perhaps they have over time learned the weaknesses of the AI system and feel completely fine relying on the causal AI-explanation. Humans need to know when to rely and when not to rely. They can learn this over time[116, 3]. Thus, understanding the context, balancing and understanding when to ask and when to tell is a challenge to make an effective human-AI system work in the real world.

Exploring multiple types of questions

The questions used as the intervention in our study were generated by a rule-based approach ("If [reason] does it follow that [label]?"). However, there are other kinds of questions that can be used to promote critical thinking. For instance, Socrates used multiple self-reflection questions in multiple turn conversation. This could be an inspiration for future research that explores how AI-framed Questioning could generate multiple turn dialogues between a human and AI systems. Such approaches have already been tested with systems

that teach critical thinking [58]. Merging this work with AI systems could yield interesting and potentially impactful results.

5.3.4 Conclusion

We presented the novel method of AI-framed Questioning, that asks questions to provoke the user to discern the logical validity of information by themselves in contrast to typical AI explanations that rely on the user passively receiving feedback from the AI system. Our study results show that AI-framed Questioning increases discernment accuracy of flawed statements and users' ratings of flawed statements level of sufficient information. While the experiment shows the technique is applicable for the specific task of logical validity, it exemplifies a future type of system where AI agents work together with and challenges humans, instead of simply telling them what they should believe or do. In a world where we are constantly bombarded with information, it is more important than ever to be able to think critically and improve our discernment capabilities beyond our current state.

6 PERSPECTIVES ON AI-ENHANCED REASONING

"But now I had before my eyes the least emotional, intelligent human being one might imagine, and yet his practical reason was so impaired that it produced, in the wanderings of daily life, a succession of mistakes, a perpetual violation of what would be considered socially appropriate and personally advantageous." – Antonio Damasio, Descartes' Error.

6.1 LIMITATIONS OF PRESENTED WORK AND POTENTIAL CHALLENGES

This thesis presented the idea of AI-Enhanced Reasoning (Chapter 2), the implementation of an AI-Enhanced Reasoning system that gives feedback on the logical structures of information (Chapter 3), that humans rely on information assessment systems that tell them the answer — even deceptive ones (Chapter 4, and, lastly, that building interaction methods that challenges the user can promote critical thinking rather than reliance. While these contributions demonstrate some of the opportunities and challenges in designing AI-Enhanced Reasoning systems, there are still many factors that are still open questions.

6.1.1 Learning Effects of AI-Enhanced Reasoning Systems

In Chapter 5, we compared the effects of a logic-checking AI system that explains users the logical validity of information against the effects of a logic-checking AI system that asks users questions. While we found immediate discernment and agency differences between the conditions, it is unclear if the interaction methods had any learning effects on the users. For instance, it could be that AI-framed questioning leads to more critical thinking in the moment but doesn't teach people the skills to critically evaluate information alone. The presented AI-enhanced reasoning systems might even end up diminishing people's critical thinking capabilities if they end up needing it to prompt their critical thinking. For instance, technological advancements such calculators

have enabled us to solve math problems in seconds but made us worse at mental math. GPS systems have allowed us to navigate without having been to a place before but made us bad at remembering street names and routes. A broader and more open question is how to design AI-enhanced reasoning systems that aid the user beyond their current critical thinking skills without diminishing their independent critical thinking capabilities. Doing this would require careful consideration of cognitive factors and interaction methods that promote learning and internalization of critical thinking skills.

6.1.2 Administrating Cognitive Resources

In Chapter 5, 4, 5 we showed that people when assisted by AI systems that give them the answers become compliant and stop using cognitive resources to engage in thinking critically about AI-mediated information themselves and take the feedback of the AI system at face value, which replicates other findings in the literature [7, 34]. One limitation that has not been measured in the experiments is the differences in cognitive load imposed on the users by the different AI explanations. Prior work has shown that AI interaction methods that increase cognitive load are less preferred by users [7]. Since AI-framed Questioning are posed as questions, engaging with the system might potentially require more effortful reflection than if the AI system just told users the answer. If that is the case, then such methods might risk that users do not want to engage with them and as a result ignore it. Future research in AI-Enhanced reasoning should investigate ways to monitor the users cognitive load so that interactions can be tailored accordingly.

6.1.3 Understanding the Influence Social factors on Engagement

In the studies presented in this thesis, we primarily focused on individual users interacting with AI-enhanced reasoning systems. However, in real-world settings, people often engage with AI systems in social contexts, such as in groups, teams, or organizations. Social factors, such as group norms, social influence, and power dynamics, can significantly impact how individuals engage with AI systems, and whether they accept or critically evaluate the information provided by these systems. For instance, people might conform to a group's opinion even when it is contradicted by an AI system's assessment, due to social pressure or the desire to maintain group harmony. On the other hand, individuals might reject AI-generated information because they perceive it

as a threat to their expertise or status within a group. These factors could moderate the effectiveness of AI-enhanced reasoning systems in promoting critical thinking and accurate discernment.

6.1.4 Studying AI-Enhanced Reasoning In-the-wild

The experiments in this thesis was conducted using offline and online participant pools in controlled environments. In these contexts, people have more time and energy to engage with the systems without distractions. However, in real-world settings, users often interact with AI systems while multitasking or in time-constrained situations, which might impact their engagement with the systems and their ability to critically evaluate information. Furthermore, the type of information people encounter in real-world settings can be more complex, nuanced, and emotionally charged than the stimuli used in the controlled experiments. Thus, it remains an open question how the findings from these experiments would generalize to real-world contexts and how AI-enhanced reasoning systems can be effectively integrated into existing workflows, tools, and platforms. To address this question, future research should investigate the use of AI-enhanced reasoning systems in real-world, practical settings, such as workplaces, educational environments, and social media platforms. This would involve conducting field studies, deploying AI-enhanced reasoning systems in existing tools and applications, and collecting both quantitative and qualitative data on user interactions and outcomes.

6.2 UNEXPLORED TYPES OF ENHANCED REASONING SYSTEMS

6.2.1 Expanding the Evaluation Capabilities of AI Enhanced Reasoning Systems

This thesis investigated AI-Enhanced reasoning systems that evaluates empty claim fallacies and hasty generalization fallacies, some of the most common type of logically invalid statements. Future work should explore a wider range of information evaluation systems such as systems that can identify more logical fallacies or even point to specific contradictions, or core premises that the user should fact check to make the conclusion true. It could also couple logic

checking systems with fact-checking systems, or even identify vague terms or phrases such as “Yes, we can!” or “Make America Great Again” to help people require speakers to define their specific approaches or conditions for making something “great”. Future work could also explore how logical feedback might help people with their own writing or self-insight by highlighting their own flaws and prompting them to make their reasoning stronger.

Lastly, AI systems could also highlight contradictions individual’s own thinking such as in their deeply held values and existing beliefs to encouraging belief revision. By making individuals aware of the potential contradictions between what they care about and their current beliefs, they may be more motivated to reevaluate their positions and update their beliefs accordingly. For example, research has shown that presenting individuals with information that connects their core values to an opposing viewpoint can lead to a reduction in their resistance to change and increase their willingness to reconsider their beliefs [26]. This approach, known as "values affirmation" or "moral reframing," leverages the power of personal values and concerns to create a more receptive mindset for belief updating [13].

6.2.2 Interaction methods based on social factors

The social context in which users interact with AI systems can significantly impact their trust and reliance on these systems. For example, users may be more likely to engage with an AI system if they believe that it is an expert or their friend. Alternatively, users may be more skeptical of AI systems with social features that they perceive as less credible or identifying with. In the wearable reasoner experiment, for instance, people reported perceiving the system as an expert due to it giving them suggestions in a British accent. Future research could investigate how changing the social role and personal features of an AI system influences people’s ability to think critically. For instance, deepfake technology, which digitally alters images, video and even audio recordings to make people look like they are doing or saying something they aren’t, can be used to create AI systems that assume different social roles, such as an expert, a peer, or an intern or someone of a less-well perceived role. By manipulating the social role of an AI system using deepfake technology, researchers can study how these different roles affect users’ trust, reliance, and engagement with AI-enhanced reasoning systems.

6.2.3 Supporting Collective Reasoning

The studies presented in this thesis mainly focus on individual users interacting with AI systems in isolation. As AI systems become more pervasive and integrated into everyday tasks and decision-making processes, it is essential to investigate how AI-enhanced reasoning systems can be designed to support and facilitate collective reasoning and deliberation among groups of people. One possible direction for future research is to explore how AI-enhanced reasoning systems can be designed to promote constructive dialogues and deliberation among users with different beliefs, biases and critical thinking skills. This may involve designing AI systems that can facilitate conversations, help the evaluation of presented information, and foster productive debates without imposing a particular viewpoint or dominating and disrupting the discussion. For instance, an AI-Enhanced Reasoning system could be developed that tries to help people in disagreement find common ground by knowing their prior beliefs and what they care about.

6.2.4 User Modeling

Even if a system is capable of identifying a bias or reasoning error in a piece of information, some users might still not care. Hence, future research should explore how to create a detailed user model that captures the individual's personal cognitive factors and biases in order to deliver the most effective interventions. For instance, AI-Enhanced reasoning systems could monitor a user's decision-making patterns and behaviors over periods of use to figure out what interventions a user is more susceptible to.

6.2.5 More specialized tasks

This thesis only evaluated the effects of AI-enhanced reasoning systems on the discernment of misinformation and divisive arguments. However, to truly test the effectiveness of the AI systems shown in this thesis, studies will need to evaluate them in various decision-making task domains, such as medical decision-making, corporate strategy and risk management, judicial decision-making, and emergency response and risk management. All of these task domains represent diverse scenarios where cognitive biases can significantly impact the quality of decisions made by individuals.

6.3 LONG-TERM PERSPECTIVES ON HUMAN COGNITION AND FEELING-BASED AI-ENHANCED REASONING

Reasoning is not just about arriving at a conclusion through systematic thinking; it involves among other things "intuitions", "gut-feelings", and different degrees of conscious awareness and agency — also when engaging in reflective thinking [64, 65, 38]. Thus, if humans are to truly enhance their own reasoning capabilities, then it is merely not enough have systems that talk to us and prompt us to think: we also need to consider the natural phenomenology of reasoning such as feeling that something is wrong when we hear someone make a logical fallacy. As reported by participants in Chapter 3, AI-enhanced reasoning systems that give language based interventions is often experienced as a second opinion or critical thinking sparring partner talking to you rather than you thinking by yourself. But as we integrate with technology, we would want to feel that we are the ones thinking, feeling, expressing, and experiencing reality, not the AI-system. These AI systems would not only serve as tools that offer external guidance or support during problem-solving tasks but also as an extension of our cognitive processes, allowing us to think and reflect more effectively.

For AI-enhanced reasoning systems to be deeper integrated with conscious structures such as implicit and explicit awareness in reasoning, e.g. "intuitions" and "controlled thinking", AI systems should take advantage of haptic, auditory, or non-invasive brain stimulation techniques like temporal interference transcranial magnetic stimulation or ultrasound stimulation could provide useful. For instance, a wristband could help prompt implicit learning and building intuitions subconsciously by vibrating whenever the user comes across a logical fallacy; or a system might induce intuitions more directly by stimulating the amygdala, which is known to be associated with negative feelings.

To facilitate a deeper integration with users' cognitive processes, AI systems should also take advantage of the latest advances in brain-sensing technologies such as EEG and fNIRS. By incorporating these technologies, AI systems could monitor users' brain activity and gain insights into how they are processing information, their emotional state, and their level of engagement with the task at hand. This information could then be used to adapt the AI's interventions and feedback in real-time so that they better support intuitive and reflective thinking.

By developing AI systems that can seamlessly blend with human cognition while respecting and responding to our emotional needs, we can create a truly symbiotic relationship that will greatly benefit both parties. As a result, we will not only be able to enhance our reasoning capabilities but also experience a greater sense of agency, control, and ownership over our thoughts and actions, leading to a richer and more fulfilling lives.

7 CONCLUSION

In this thesis, I have explored the how AI systems might augment human reasoning and critical thinking abilities. Through a series of experiments and the creation of novel AI interaction methods, I sought to gain a deeper understanding of how AI systems can be designed to foster more effective critical thinking and well-informed decision-making. These interactions included logical validation feedback, deceptive AI-generated explanations, and AI-framed questioning.

This work has contributed to our understanding of the role AI systems can play in enhancing human cognitive abilities and the potential they hold for addressing complex global challenges. By investigating diverse AI-enhanced reasoning systems and reevaluating our approach to AI interactions, we can begin to form a more comprehensive picture of the future of AI and its implications for human reasoning.

There are, however, limitations to this work, such as its focus on specific forms of reasoning and its restriction to controlled experimental environments. As such, it is crucial to extend this research into real-world contexts and integrate additional aspects of human cognition to develop a more holistic understanding of AI-enhanced reasoning. This may involve exploring the role of social factors, assessing long-term learning effects, and investigating novel combinations of AI interaction methods and human cognitive capacities.

Ultimately, the pursuit of AI-enhanced reasoning is not only about creating effective AI systems but also about considering the broader implications of these systems on our society and fostering a symbiotic relationship between humans and AI where human capabilities and agency are first priority. Such a relationship should promote human growth and flourishing, rather than leading to dependence and passivity. As we continue to explore the potential of AI-enhanced reasoning, it is essential to prioritize human agency and autonomy and ensure that these systems empower humans to think more critically, make better informed decisions, and navigate the increasingly complex world around them.

This thesis has taken concrete steps towards understanding the possibilities and challenges of AI-enhanced reasoning, but there is still much work to be done to unlock the full potential of these systems. I hope that this work

inspires further research in this area and encourages developers, researchers, and policymakers to consider the ethical and societal impact of advanced AI systems that assist in decision making tasks. By navigating the complexities of AI-driven support for critical thinking and decision-making, we can work together to build a more rational, discerning, and informed society for the betterment of humanity.

BIBLIOGRAPHY

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, page 582. ACM, 2018.

[2] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. Where is your evidence: improving fact-checking by justification modeling. In *Proceedings of the first workshop on fact extraction and verification (FEVER)*, pages 85–90, 2018.

[3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[4] Richard E Boyatzis. *Transforming qualitative information: Thematic analysis and code development*. sage, 1998.

[5] Judith L Bronstein. *Mutualism*. Oxford University Press, USA, 2015.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

[8] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

[9] Vladimíra Čavojová. When beliefs and logic contradict: issues of values, religion and culture. *Advances in Culturally-Aware Intelligent Systems and in Cross-Cultural Psychological Studies*, pages 367–390, 2018.

[10] Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. Targer: Neural argument mining at your fingertips. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 195–200, 2019.

[11] E Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979, 1953.

[12] Oana Cocarascu and Francesca Toni. Identifying attack and support argumentative relations using deep learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1374–1379, 2017.

[13] Geoffrey L Cohen, David K Sherman, Anthony Bastardi, Lillian Hsu, Michelle McGoey, and Lee Ross. Bridging the partisan divide: Self-affirmation reduces ideological closed-mindedness and inflexibility in negotiation. *Journal of personality and social psychology*, 93(3):415, 2007.

[14] Roberto Confalonieri, Tarek R Besold, Tillman Weyde, Kathleen Creel, Tania Lombrozo, Shane Mueller, and Patrick Shafto. What makes a good explanation? cognitive dimensions of explaining intelligent machines. *CogSci 2019: Creativity+ Cognition+ Computation*, 2019.

[15] John W Creswell and J David Creswell. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2017.

[16] Valdemar Danry, Pat Pataranutaporn, Adam Haar Horowitz, Paul Strohmeier, Josh Andres, Rakesh Patibanda, Zhuying Li, Takuto Nakamura, Jun Nishida, Pedro Lopes, et al. Do cyborgs dream of electric limbs? experiential factors in human-computer integration design and evaluation. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2021.

[17] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. Wearable reasoner: towards enhanced human rationality through a wearable device with an explainable ai assistant. In *Proceedings of the Augmented Humans International Conference*, pages 1–12, 2020.

[18] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. Don't just tell me, ask me: Ai systems that intelligently frame explanations as questions improve human logical discernment accuracy over causal ai

explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2023.

[19] Valdemar Danry, Pat Pataranutaporn, Florian Mueller, Pattie Maes, and Sang-won Leigh. On eliciting a sense of self when integrating with computers. In *Augmented Humans 2022*, pages 68–81. 2022.

[20] Mamoru Deguchi and Kazunori Yamaguchi. Argument component classification by relation identification by neural network and textrank. In *Proceedings of the 6th Workshop on Argument Mining*, pages 83–91, 2019.

[21] Angela E Douglas. *Symbiotic interactions*. Number 577.85 D733s. Oxon, GB: Oxford University Press, 1994, 1994.

[22] Yanqing Duan, John S Edwards, and Yogesh K Dwivedi. Artificial intelligence for decision making in the era of big data–evolution, challenges and research agenda. *International journal of information management*, 48:63–71, 2019.

[23] Ullrich KH Ecker, Stephan Lewandowsky, Olivia Fenton, and Kelsey Martin. Do people keep believing because they want to? preexisting attitudes and the continued influence of misinformation. *Memory & cognition*, 42:292–304, 2014.

[24] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. The impact of placebic explanations on trust in intelligent systems. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*, pages 1–6, 2019.

[25] Ziv Epstein, Nicolo Foppiani, Sophie Hilgard, Sanjana Sharma, Elena Glassman, and David Rand. Do explanations increase the effectiveness of ai-crowd generated fake news warnings? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 183–193, 2022.

[26] Matthew Feinberg and Robb Willer. The moral roots of environmental attitudes. *Psychological science*, 24(1):56–62, 2013.

[27] Leon Festinger. Cognitive dissonance. *Scientific American*, 207(4):93–106, 1962.

[28] Valerie S Folkes. Mindlessness or mindfulness: A partial replication and extension of langer, blank, and chanowitz. 1985.

[29] Shane Frederick. Cognitive reflection and decision making. *Journal of Economic perspectives*, 19(4):25–42, 2005.

[30] Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. *arXiv preprint arXiv:2104.08790*, 2021.

[31] Dongfang Gaozhao. Flagging fake news on social media: An experimental study of media consumers' identification of fake news. *Government Information Quarterly*, 38(3):101591, 2021.

[32] Gerd Gigerenzer and Reinhard Selten. *Bounded rationality: The adaptive toolbox*. MIT press, 2002.

[33] Google. People + ai guidebook: Explainability + trust, 2020.

[34] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1):e2110013119, 2022.

[35] Ivan Habernal and Iryna Gurevych. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, 2016.

[36] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. *arXiv preprint arXiv:1802.06613*, 2018.

[37] Shohreh Haddadan, Serena Villata, and Elena Cabrio. Yes, we can! mining arguments in 50 years of us presidential campaign debates. 2019.

[38] Jonathan Haidt. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, 108(4):814, 2001.

[39] Jonathan Haidt. *The righteous mind: Why good people are divided by politics and religion*. Vintage, 2012.

[40] Ali Hasan and Richard Fumerton. Foundationalist theories of epistemic justification. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2018 edition, 2018.

[41] Edward F Haskell. A clarification of social science. *Main Currents in Modern Thought*, 7(2):45–51, 1949.

[42] Naeemul Hassan, Bill Adair, James T Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. The quest to automate fact-checking. In *Proceedings of the 2015 computation+ journalism symposium*. Citeseer, 2015.

[43] Robert R Hoffman, Gary Klein, and Shane T Mueller. Explaining explanation for "explainable ai". In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 62, pages 197–201. SAGE Publications Sage CA: Los Angeles, CA, 2018.

[44] Andreas Holzinger, Bernd Malle, Peter Kieseberg, Peter M Roth, Heimo Müller, Robert Reihs, and Kurt Zatloukal. Towards the augmented pathologist: Challenges of explainable-ai in digital pathology. *arXiv preprint arXiv:1712.06657*, 2017.

[45] Jonathan Howard. Hasty generalization, survival bias, special pleading, and burden of proof. In *Cognitive Errors and Diagnostic Mistakes*, pages 211–246. Springer, 2019.

[46] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. Interacting with opinionated language models changes users' views. *Arxiv Open Access*, 2022.

[47] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[48] Daniel Kahneman, Stewart Paul Slovic, Paul Slovic, and Amos Tversky. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press, 1982.

[49] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey Hancock. Working with ai to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Stanford Preprint*, 2023.

[50] Katarina Kertysova. Artificial intelligence and disinformation: How ai changes the way disinformation is produced, disseminated, and can be countered. *Security and Human Rights*, 29(1-4):55–81, 2018.

[51] Jonathan Kobbe, Juri Opitz, Maria Becker, Ioana Hulpus, Heiner Stuckenschmidt, and Anette Frank. Exploiting background knowledge for argumentative relation classification. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2019.

[52] Deanna Kuhn. *The skills of argument*. Cambridge University Press, 1991.

[53] HA Kusmantini, I Asror, and MA Bijaksana. Argumentation mining: classifying argumentation components with partial tree kernel and support vector machine for constituent trees on imbalanced persuasive essay. In *Journal of Physics: Conference Series*, volume 1192, page 012009. IOP Publishing, 2019.

[54] Emily R Lai. Critical thinking: A literature review. *Pearson's Research Reports*, 6(1):40–41, 2011.

[55] Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 29–38, 2019.

[56] Ellen J Langer, Arthur Blank, and Benzion Chanowitz. The mindlessness of ostensibly thoughtful action: The role of" placebic" information in interpersonal interaction. *Journal of personality and social psychology*, 36(6):635, 1978.

[57] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[58] Nguyen-Thinh Le and Laura Wartschinski. A cognitive assistant for improving human reasoning skills. *International Journal of Human-Computer Studies*, 117:45–54, 2018.

[59] Stephan Lewandowsky, Ullrich KH Ecker, and John Cook. Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of applied research in memory and cognition*, 6(4):353–369, 2017.

[60] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, 13(3):106–131, 2012.

[61] John Licato, Mark Boger, and Zhitian Zhang. Developing a dataset for personal attacks and other indicators of biases. In *2018 AAAI Spring Symposium Series*, 2018.

[62] Joseph Carl Robnett Licklider. Man-computer symbiosis. *IRE transactions on human factors in electronics*, (1):4–11, 1960.

[63] Matthias Liebeck, Katharina Esau, and Stefan Conrad. What to do with an airport? mining arguments in the german online participation project tempelhofer feld. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 144–153, 2016.

[64] Matthew D Lieberman. Intuition: a social cognitive neuroscience approach. *Psychological bulletin*, 126(1):109, 2000.

[65] Matthew D Lieberman, Ruth Gaunt, Daniel T Gilbert, and Yaacov Trope. Reflexion and reflection: a social cognitive neuroscience approach to attributional inference. 2002.

[66] Marco Lippi and Paolo Torroni. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):10, 2016.

[67] Emily G Liquin and Tania Lombrozo. A functional approach to explanation-seeking curiosity. *Cognitive Psychology*, 119:101276, 2020.

[68] Tania Lombrozo. Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, 20(10):748–759, 2016.

[69] Tania Lombrozo and Susan Carey. Functional explanation and the function of explanation. *Cognition*, 99(2):167–204, 2006.

[70] Anastasios Lytos, Thomas Lagkas, Panagiotis Sarigiannidis, and Kalina Bontcheva. The evolution of argumentation mining: From models to social media and emerging tools. *Information Processing & Management*, 56(6):102055, 2019.

[71] Scott E Maxwell, Harold D Delaney, and Ken Kelley. *Designing experiments and analyzing data: A model comparison perspective*. Routledge, 2017.

[72] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.

[73] Sarah Mennicken, Jo Vermeulen, and Elaine M Huang. From today's augmented houses to tomorrow's smart homes: new directions for home automation research. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 105–115. ACM, 2014.

[74] Hugo Mercier and Dan Sperber. *The enigma of reason*. Harvard University Press, 2017.

[75] Nancy A Moran. Symbiosis as an adaptive process and source of phenotypic complexity. *Proceedings of the National Academy of Sciences*, 104(suppl 1):8627–8633, 2007.

[76] Neville Moray. Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly journal of experimental psychology*, 11(1):56–60, 1959.

[77] Gaku Morio and Katsuhide Fujita. End-to-end argument mining for discussion threads based on parallel constrained pointer architecture. *arXiv preprint arXiv:1809.00563*, 2018.

[78] Caitlin Morris, Valdemar Danry, and Pattie Maes. Ember: A system for transfer of interoceptive sensations to improve social perception. In *Designing Interactive Systems Conference*, pages 277–287, 2022.

[79] Shane T Mueller, Robert R Hoffman, William Clancey, Abigail Emrey, and Gary Klein. Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai. *arXiv preprint arXiv:1902.01876*, 2019.

[80] Huy Nguyen and Diane Litman. Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1127–1137, 2016.

[81] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

[82] Brendan Nyhan and Jason Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010.

[83] Seymour Papert. You can't think about thinking without thinking about thinking about something. *Contemporary Issues in Technology and Teacher Education*, 5(3):366–367, 2005.

[84] Demetris Paschalides, Alexandros Kornilakis, Chrysovalantis Christodoulou, Rafael Andreou, George Pallis, Marios Dikaiakos, and Evangelos Markatos. Check-it: A plugin for detecting and reducing the spread of fake news and misinformation on the web. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 298–302, 2019.

[85] Pat Pataranutaporn, Valdemar Danry, Lancelot Blanchard, Lavanay Thakral, Naoki Ohsugi, Pattie Maes, and Misha Sra. Living memories:

Ai-generated characters as digital mementos. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 889–901, 2023.

[86] Pat Pataranutaporn, Valdemar Danry, Joanne Leong, Parinya Punpongsanon, Dan Novy, Pattie Maes, and Misha Sra. Ai-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, 3(12):1013–1022, 2021.

[87] Pat Pataranutaporn, Valdemar Danry, and Pattie Maes. Machinoia, machine of multiple me: Integrating with past, future and alternative selves. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.

[88] Pat Pataranutaporn, Joanne Leong, Valdemar Danry, Alyssa P Lawson, Pattie Maes, and Misha Sra. Ai-generated virtual instructors based on liked or admired people can improve motivation and foster positive emotions for learning. In *2022 IEEE Frontiers in Education Conference (FIE)*, pages 1–9. IEEE, 2022.

[89] R Paul and L Elder. The art of socratic questioning. dillon beach, ca: foundation for critical thinking. *Psychology, Monograph Series II*, 3:107–127, 2006.

[90] Richard Paul and Linda Elder. Critical thinking: The art of socratic questioning. *Journal of developmental education*, 31(1):36, 2007.

[91] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science*, 31(7):770–780, 2020.

[92] Gordon Pennycook and David G Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50, 2019.

[93] Gordon Pennycook and David G Rand. Who falls for fake news? the roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of personality*, 88(2):185–200, 2020.

[94] Georgios Petasis. Segmentation of argumentative texts with contextualised word representations. In *Proceedings of the 6th Workshop on Argument Mining*, pages 1–10, 2019.

[95] Alexander S Rich and Todd M Gureckis. Lessons for artificial intelligence from the study of natural stupidity. *Nature Machine Intelligence*, 1(4):174, 2019.

[96] Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[97] John R Searle and John Rogers Searle. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press, 1969.

[98] Upendra Shardanand and Pattie Maes. Social information filtering: algorithms for automating" word of mouth". In *Chi*, volume 95, pages 210–217. Citeseer, 1995.

[99] Herbert Alexander Simon. *Models of bounded rationality: Empirically grounded economic reason*, volume 3. MIT press, 1997.

[100] Kacper Sokol and Peter Flach. One explanation does not fit all. *KI-Künstliche Intelligenz*, pages 1–16, 2020.

[101] Misha Sra, Valdemar Danry, and Pattie Maes. Situated vr: Toward a congruent hybrid reality without experiential artifacts. *IEEE Computer Graphics and Applications*, 42(3):7–18, 2022.

[102] Keith E Stanovich, Richard F West, and Maggie E Toplak. *The rationality quotient: Toward a test of rational thinking*. MIT press, 2016.

[103] Manfred Stede and Jodi Schneider. *Argumentation mining*. Morgan & Claypool, 2019.

[104] Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. Toward a geographic understanding of the sharing economy: Systemic biases in uberx and taskrabbit. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(3):21, 2017.

[105] Ottokar Tilk and Tanel Alumäe. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *Interspeech 2016*, 2016.

[106] Maggie E Toplak, Richard F West, and Keith E Stanovich. Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning*, 20(2):147–168, 2014.

[107] Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. Robust argument unit recognition and classification. *arXiv preprint arXiv:1904.09688*, 2019.

[108] Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. Fake news detection in social networks via crowd signals. In *Companion proceedings of the the web conference 2018*, pages 517–524, 2018.

[109] Amos Tversky and Daniel Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232, 1973.

[110] Jay J Van Bavel and Andrea Pereira. The partisan brain: An identity-based model of political belief. *Trends in cognitive sciences*, 22(3):213–224, 2018.

[111] Tomás Alfonso Vega Gálvez. *uJawstures: Jaw-teeth microgestures for discreet hands-and-eyes-free mobile device interaction*. PhD thesis, Massachusetts Institute of Technology, 2019.

[112] Jan G Voelkel, Robb Willer, et al. Artificial intelligence can persuade humans on political issues. *OSF Preprints*.

[113] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.

[114] Rayoung Yang and Mark W Newman. Learning from a learning thermostat: lessons for intelligent systems for the home. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 93–102. ACM, 2013.

[115] John Zarocostas. How to fight an infodemic. *The lancet*, 395(10225):676, 2020.

[116] Qiaoning Zhang, Matthew L Lee, and Scott Carter. You complete me: Human-ai teams and complementary expertise. In *CHI Conference on Human Factors in Computing Systems*, pages 1–28, 2022.