Wearable Reasoner: Towards Enhanced Human Rationality Through A Wearable Device With An Explainable AI Assistant

Valdemar Danry*
MIT Media Lab,
Massachusetts Institute of
Technology (MIT)
Cambridge, MA
valdemar.muda@gmail.com

Pat Pataranutaporn*
MIT Media Lab,
Massachusetts Institute of
Technology (MIT)
Cambridge, MA
patpat@media.mit.edu

Yaoli Mao*
Columbia University
New York City, NY
ym2429@tc.columbia.edu

Pattie Maes
MIT Media Lab,
Massachusetts Institute of
Technology (MIT)
Cambridge, MA
pattie@media.mit.edu

Figure 1: Left: Overview of Wearable Reasoner system. Right: Envisioned use cases of Wearable Reasoner for detecting empty claims (arguments presented without evidence) in political speech, advertisements, or high stake meetings.

ABSTRACT

Human judgments and decisions are prone to errors in reasoning caused by factors such as personal biases and external misinformation. We explore the possibility of enhanced reasoning by implementing a wearable AI system as a human symbiotic counterpart. We present “Wearable Reasoner”, a proof-of-concept wearable system capable of analyzing if an argument is stated with supporting evidence or not. We explore the impact of argumentation mining and explainability of the AI feedback on the user through an experimental study of verbal statement evaluation tasks. The results demonstrate that the device with explainable feedback is effective in enhancing rationality by helping users differentiate between statements supported by evidence and without. When assisted by an AI system with explainable feedback, users significantly consider claims supported by evidence more reasonable and agree more with them compared to those without. Qualitative interviews demonstrate users’ internal processes of reflection and integration

of the new information in their judgment and decision making, emphasizing improved evaluation of presented arguments.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI); Empirical studies in ubiquitous and mobile computing.**

KEYWORDS

Wearable Computing, Human-Machine Symbiosis, Assisted Reasoning, Judgment, Decision Making, Cognitive Enhancement, Augmented Humans, Argumentation Mining, Explainable Artificial Intelligence, Informal Logic

ACM Reference Format:

Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2020. Wearable Reasoner: Towards Enhanced Human Rationality Through A Wearable Device With An Explainable AI Assistant. In *AHs ’20: Augmented Humans International Conference (AHs ’20)*, March 16–17, 2020, Kaiserslautern, Germany. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3384657.3384799>

* Authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

AHs ’20, March 16–17, 2020, Kaiserslautern, Germany

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-7603-7/20/03...\$15.00

<https://doi.org/10.1145/3384657.3384799>

1 INTRODUCTION

We live in a world accelerated by information, overloading us with tremendous volumes of data, exceeding our biological brain’s processing capability. Researchers have shown that people frequently use “cognitive shortcuts” for processing information that do not involve rational and cautious thinking. [33]. These shortcuts act as efficient rules of thumbs allowing people to make decisions quickly.

However, those decisions are prone to errors including fallacious (illogical) reasoning and irrational behavior from personal biases and external misinformation, a phenomenon known as “bounded rationality” [19, 74].

People’s bounded rationality has consequences ranging from personal decision making to nation-wide politics. Politicians and marketing companies exploit these loopholes in our cognitive processes. A typical example is to present persuasive claims without evidence to support them[69]. During the 2019 US presidential election campaign, one rally had 67 percent of statements with false evidence or lacking evidence altogether[67]. Such campaigns and advertisements are devised to manipulate and mislead public opinion and decisions to achieve favorable outcomes or desired profits, often regardless of true public interests or concerns. Thus, it is evermore crucial to support and empower people to make use of rational thinking about external information in critical judgments and decisions.

In the broader psychology literature, rational thinking is defined as the capability to critically evaluate the quality of new information (e.g. whether and how well a given claim is supported by evidence) and effectively integrate it with one’s own judgment and decision making[79, 80]. Specifically, it can be considered as probing people to think about their own thinking in terms of “How do I know that?”, “What is the evidence?”, and “If this is true, then how does this follow?” [20].

Following Licklider’s vision of human-machine symbiosis which introduced the idea of using machines in augmenting human intelligence [44], we believe that technology could play a significant role in augmenting our rational thinking, judgment and decision making. Instead of providing individuals with excessive information (we are already overloaded), we advocate for enhancing and training cognitive capabilities of information processing, that is, supporting users to think about thinking [60].

Based on recent advances in artificial intelligence (AI), argument mining, and computational linguistics, we envision the possibility of having an AI assistant as a symbiotic counterpart to the biological human brain. As a “second brain”, the AI serves as an extended, rational reasoning organ that assists the individual and can teach them to become more rational over time by making them aware of biased and fallacious information through just-in-time feedback. To ensure the transparency of the AI system, and prevent it from becoming an AI “black box”, it is important for the AI to be able to explain how it generates its classifications. This Explainable AI additionally allows the person to speculate, internalize and learn from the AI system, and prevents an over-reliance on the technology.

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 is stated with evidence or not. 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.

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?**

Our findings demonstrate that our prototype with Explainable AI feedback has a significant effect on users’ level of agreement as well as on how reasonable they find the presented arguments to be. When assisted with such feedback, users tend to agree more with claims supported by evidence and consider them more reasonable compared to those without. In qualitative interviews users report their internal processes of speculating and integrating the information from the AI system in their own judgment and decision making, resulting in improved evaluation of presented arguments.

2 RELATED WORK

2.1 Human Judgment and Decision Making

Cognitive scientists have introduced the dual-system theory to describe the nature of human thought in reasoning [75], judgment and decision making [34]. System 1 is fast, automatic, unconscious, emotional, and error-prone while system 2 is slow, controlled, conscious, rational, but unfortunately requires a lot of mental effort.

Researchers have systematically found that people are rarely rational and cautious thinkers in judgment and decision making. Instead, they use cognitive shortcuts to make inferences that form judgments and decisions. That is to say, system 1 is mostly working as the default as it is fast and requires little cognitive resources, with system 2 sometimes intervening and overriding system 1 as it is slow and cognitively demanding. We are so-called “cognitive misers” [18, 78] due to the fact that the human mind is rather limited in time, knowledge, attention, and cognitive resources[73].

These shortcuts are also known as heuristics, well exemplified in the book “Thinking, Fast and Slow” authored by the Nobel laureate Daniel Kahneman and Amos Tversky [33]. They refer to a set of schemas, scripts, and stereotypes from prior knowledge or experience such as trial and error, a rule of thumb, an intuitive judgment, and common sense, in contrast with careful analytical thinking.

Despite equipping us with a smart repertoire of fast and frugal rules for decision making under uncertainty, heuristics simultaneously make our actions prone to errors from personal biases, fallacious reasoning, and external misinformation [19]. As a result, people’s judgments and decision making systematically deviate from rationality and objectivity, and such deviations are known as “biases” [35, 73].

People suffer from these cognitive biases by making less optimal judgments and decisions. Once they have formed these judgments, they sometimes even regard them as their own ‘truths’ or ‘facts’, tuning out ideologically incongruent news, curating echo chambers

on social media, and moving to ideologically like-minded communities [87]. Ever since the 2016 U.S. election, “fake news” and fallacious argumentation on social media has been notoriously blamed for manipulating, magnifying and taking advantage of these cognitive biases [41, 61].

Despite journalistic fact-checking being the primary means for countering such influences on public beliefs, fact-checking does not address the issue of fallacious argumentation (such as correlation/causation mistakes, personal attacks etc.) which are errors in logical thinking and independent from facts[5]. Furthermore, research has shown that mere correction such as fact-checking often has minimal effects on presidential candidate evaluations or voter choices [58]. Thus, merely correcting beliefs with fact-checking or calling out fallacies might affect someone’s attitudes but does not change their actual judgments or decisions. Instead, *explaining* the underlying reasoning from facts to opinions has shown to reduce biases significantly, e.g. by stating why something is or is not the case [42].

Following this thread, this study seeks to explore interventions that support people in reasoning about critical social topics, i.e. explaining whether a certain statement is presented with or without evidence (the latter known also as a bare-assertion fallacy[88]), to understand how such interventions would affect people’s judgment and decision making.

2.2 Argumentation Mining

The emerging field of argumentation mining presents interesting possibilities for the domain of assisted reasoning. As a rising subject in computational linguistics and AI, Argumentation Mining focuses on extracting structured arguments from natural and often unstructured text [14, 49]. 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”[49].

To do this, most argumentation mining approaches rely on a claim-premise model of argument 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 [37, 45, 85].

For the claim-premise model, a claim is certified as sound by examining the soundness of its provided reasons or evidence, e.g. a claim can be certified as sound because of an expert statement or study saying so (as we generally assume that experts and studies are sound) [27]. Naturally, then, if no evidence is given to support or attack a certain claim, then it is hard to justify holding the claim as either true or false - and in such case, we are no longer faced with a substantial argument but merely an empty claim¹. Following the preceding section, argumentation mining then allows us to identify these empty claims through inherent argument structures and notify the user that a claim was given without evidence.

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*[49].

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 [46, 62, 85]. 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 [9, 26, 39, 45, 55]. However, even though this approach identifies components as either premises or claims, this is mostly useful for designing argument search engines, where you can search for evidence that supports or attacks a certain claim, e.g. to use in debates as with IBM’s ‘Project Debater’ [76].

The downside of component identification, however, is that it gives us no clue as to what the relationship between the identified components are [49]. Relations identification on the other hand attempts to do this by identifying the supporting, attacking, or lack of relationship between specific units [11, 37, 49, 56], thereby helping to understand the precise reasoning (the ‘why’) behind an argument.

Another application of argumentation mining is that of argument evaluation. Argument evaluation attempts to identify faulty reasoning to prevent it from interfering with our judgments and decisions. Within the argument evaluation domain of argumentation mining, prior work has focused on evaluating the validity of an argument by identifying the use of fallacious argumentation schemes such as personal attacks [25, 43] and convincing vocabulary and references [24]. Additionally, work has also been focused on evaluating the soundness of an argument through fact-checking [70] and classifying its type of evidence (if any is given) [63, 66].

Even though argumentation mining is still in its infancy [50], there have been several attempts of practical use of the technology, especially within legal, medical, and governmental argumentation [3], whereas work on argument evaluation has focused on improving automatic scoring of persuasive essays [57]. To our best knowledge, not much work has been done within the domain of argumentation mining for assisted reasoning. We hypothesize that this is because the contemporary work on argumentation mining has been primarily focused on ‘mining’ and evaluating arguments from *online discourses* rather than ‘real life’. Going into ‘real life’ applications of the technology thus suggests interesting potential use-cases, such as real-time ‘in the wild’ argument mining and evaluation.

2.3 Wearable Augmentation

As research in AI continues to progress, one solution for enhancing our cognitive abilities [51] could be to have a symbiotic relationship with an intelligent system [44]. Doug Englebart in the 1960s further explored this idea in his proposal “A Conceptual Framework for Augmenting Human Intellect” focusing on technology for rapid access to useful information and increasing users’ comprehension in a complex situation [16]. He also developed one of the first intelligent user interfaces called “Intelligence Amplifying System of Tools” to demonstrate his vision. Since then, researchers have explored various technological interfaces for enhancing human cognitive abilities [31, 81].

More specifically, “wearable technology” has enabled on-body monitoring of human physiological and behavioral information for

¹There are some cases in which it is argued that a claim is justified without evidence. We will return to this point in the discussion.

providing just-in-time feedback and assistance to the user. Example areas of enhancement include: memory and recall [40, 64, 68, 90], attention [8, 38], emotion regulation [13, 59], sleep [2, 23], creativity [4, 23], and more.

In terms of wearable technology for enhancing judgment and decision making, most researchers focus on enhancing users' access to useful information by creating more seamless human-machine interfaces through brain-computer interfaces [36, 86], or deploying novel augmented reality displays [32, 82]. To our knowledge, the research community has not yet explored the intersection of argument mining and wearable computing, where the device in needed contexts can assist the user in real-time to be more rational and less prone to heuristic influences in judgment and decision making.

2.4 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 [7, 65]. Prior research has reported a series of biased examples of AI use in smart homes [52, 91], decision making support for disease diagnosis [30], and autonomous vehicles [83].

To address these issues, DARPA recently launched the Explainable AI program that aims to have AI systems explain their own behaviors to achieve the much needed trust and transparency [22]. 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 [12]. 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" [29, 48]. Similar results have also been found in recent psychological research on explanation-seeking [47].

Thus, in our study, we utilize argumentation mining capabilities for generating explainable feedback that brings user awareness on the quality of argumentation, in our case whether presented claims are supported by evidence or not.

3 MOTIVATION AND CONTRIBUTIONS

Our aim is to leverage state of the art wearables, argumentation mining, and Explainable AI systems to create technology that can assist and train users' reasoning abilities by inspecting and evaluating real-life argument structures in real time.

In our case, as research has shown that arguments presented as self-evidently true are easily believed (despite lacking supporting evidence) [69], we want to encourage people to be more skeptical and seek or demand evidence before jumping to conclusions. In our study we do this by making the user of our prototype aware of empty claims and non-empty claims (arguments presented without or with evidence) in real time. Our hope is that by doing so, we can help users avoid mistakenly regarding something as true or

false without knowing 'why', and thus can avoid such influences on their judgments and decision making.

We consider the explainability an essential design aspect of the Wearable Reasoner prototype, which can be leveraged by the device's real-time voice feedback and interaction capability within the context of diverse socially critical topics. Thus, we designed two types of interventions to examine the effects of explainability in a controlled experiment: 1) an Explainable AI mode that provides causal reasoning of why the machine learning algorithm made a certain classification and 2) a Non-Explainable AI mode that solely delivers a one-word answer without providing an explanation.

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 evidence is present or not) through an Explainable- or Non-Explainable AI mode, and 3) exploration and evaluation of the novel area of enhanced reasoning computation through a closed environment controlled user study.

4 METHODOLOGY

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

4.1 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.

4.1.1 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 [89]. Further, research shows that humans can perceive multiple audio streams at the same time [10, 54]. This highlights the opportunity of using wearable auditory interfaces for minimally disruptive enhancement.

For our implementation we used the commercially available Bose Frames². We chose a glasses form factor because of its social acceptability, comfort for 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.

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

Figure 2: System Architecture of Wearable Reasoner

4.1.2 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 [66], 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 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

Figure 3: Confusion Matrix of the Ensemble Model for the Evidence Type Classification Model

Model	15 trains average score			
	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

for the binary ‘with/without evidence’-classification we reached an accuracy of 90.4%.

4.1.3 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 Section 2.4. As an example, when an argument consists of both claim and evidence, the device in the Explainable AI mode would provide the user with the following feedback:

“The claim was ‘A majority of Americans support a ban on flag-burning’ because of the ‘study’ based evidence: ‘A poll conducted by CNN in June 2006 found that 56% of Americans supported a flag desecration amendment.’”

And when the argument for instance consists of only claims, the device would instead tell the user that:

“The claims were ‘Social and political issues surround the issue of immigration.’ and ‘Failing to stop the illegal

Figure 4: Intervention Types

immigration waves at an early stage will only lead to much larger waves of illegal immigration in the future.’ but there were given no evidence to support them. Since there were given no evidence to support them, the claims could be either true or false”

4.2 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.

4.2.1 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.

4.2.2 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.

4.2.3 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[84]. 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.

4.2.4 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.

4.2.5 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 5: Experimental Setup

4.3 Experimental 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 5) 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 6).

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.

4.3.1 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 univariate 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 6. 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%.

4.3.2 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.

4.3.3 Environmental Setting. The experiment took place in a lab room free from external noise, with the following setup (see Figure 5). 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.

Figure 6: Distribution of 'Word Count' and 'Grade Level' for the Experiment Dataset

4.4 Experiment 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 7). Additionally, participants 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 [80].

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

5 RESULTS

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.

Figure 7: Experimental Sequences

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.

5.1 Machine Learning 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 8. 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.

5.2 Augmenting Judgment and Decision Making

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

Figure 8: The performance comparison between computer-based evaluation with real-world evaluation

aware of which statements are presented with evidence and without (see Figure 9). 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 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.

5.3 User Experience and Qualitative Comments

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

5.3.1 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.

Figure 9: Users' Judgment and Decision Making Results

"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."

5.3.2 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 [44]. The concept of symbiosis originated in the field of biology, referring to "the living together of unlike organisms" [15], "co-actions" [28], "interactions" [6], as well as potential "co-evolution" [53].

In their comments, participants seem 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.

5.3.3 Cognitive Dissonance and Cognitive Load. In addition, some participants also reported phenomena 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 9 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" [17]. 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.

5.3.4 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."

5.3.5 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."

6 DISCUSSION

6.1 Towards Enhanced Rationality

In this study, we explored a first step in augmenting human rational thinking 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” [60] by exposing and reminding users of the varying quality of presented information.

6.2 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 [72]. 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?

6.3 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 [71]. 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) [27].

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” [77]. 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[21].

Finally, in this paper we proposed one type of enhanced reasoning device which identifies arguments as being supported or lacking evidence. However, there is still much more work to be done. In future reasoning devices, a broader implementation of different types of logic (deductive, inductive, abductive) could be leveraged for greater real-life use. Smaller steps could be to incorporate detection of additional fallacies besides claims without evidence (bare-assertion fallacies).

7 CONCLUSION

In their Nobel Prize winning research, Daniel Kahneman and Amos Tversky highlighted one of the most significant challenges people face in the age of tremendous volumes of information. They showed how people use “cognitive shortcuts”, instead of reasoning in judgment and decision making [33]. As a result, people are bounded in rationality and often suffer from personal biases and external misinformation. This can have major consequences from the personal to the national level.

The vision of human-machine symbiosis has provided a promise for solving this problem. It proposes the augmentation of human cognition with an intelligent system, serving as an “extended rational reasoning organ”.

In this paper, we presented a first prototype and an experimental study of a “Wearable Reasoner”, a wearable system for enhancing rational judgment and decision making, with a comparison of Explainable and Non-Explainable feedback modes.

Our results showed how a wearable device with an Explainable AI assistant can enhance users' cognitive capabilities through real-time audio feedback. This indicated that people are less likely to simply follow the answers from the AI unless it can explain its process. This highlights the importance of explainability of AI systems for people to develop a trusted relationship with technology.

With more advancements in AI to come, this work is paving the way towards a future of human-machine symbiosis, where an AI “second brain” helps us thrive in the age of accelerated information.

REFERENCES

- [1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanahalli. 2018. 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*. ACM, 582.
- [2] Judith Amores, Mae Dotan, and Pattie Maes. 2019. An Exploration of Form Factors for Sleep-Olfactory Interfaces. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 1456–1460.
- [3] Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Simari, Matthias Thimm, and Serena Villata. 2017. Towards artificial argumentation. *AI magazine* 38, 3 (2017), 25–36.
- [4] Guillermo Bernal, Lily Zhou, Haripriya Mehta, and Pattie Maes. 2019. Paper Dreams: An Interactive Interface for Generative Visual Expression. (2019).
- [5] Maarten Boudry, Fabio Paglieri, and Massimo Pigliucci. 2015. The fake, the flimsy, and the fallacious: Demarcating arguments in real life. *Argumentation* 29, 4 (2015), 431–456.
- [6] Judith L Bronstein. 2015. *Mutualism*. Oxford University Press, USA.
- [7] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.
- [8] Daniel Chen and Roel Vertegaal. 2004. Using mental load for managing interruptions in physiologically attentive user interfaces. In *CHI'04 extended abstracts on Human factors in computing systems*. ACM, 1513–1516.
- [9] Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. Targer: Neural argument mining at your fingertips. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 195–200.
- [10] E Colin Cherry. 1953. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America* 25, 5 (1953), 975–979.
- [11] Oana Cocarascu and Francesca Toni. 2017. Identifying attack and support argumentative relations using deep learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1374–1379.
- [12] Roberto Confalonieri, Tarek R Besold, Tillman Weyde, Kathleen Creel, Tania Lombrozo, Shane Mueller, and Patrick Shafto. 2019. What makes a good explanation? Cognitive dimensions of explaining intelligent machines. *CogSci 2019: Creativity+ Cognition+ Computation* (2019).
- [13] Jean Costa, Alexander T Adams, Malte F Jung, François Guimbretière, and Tanzeem Choudhury. 2017. EmotionCheck: A Wearable Device to Regulate Anxiety through False Heart Rate Feedback. *GetMobile: Mobile Computing and Communications* 21, 2 (2017), 22–25.
- [14] Mamoru Deguchi and Kazunori Yamaguchi. 2019. Argument Component Classification by Relation Identification by Neural Network and TextRank. In *Proceedings of the 6th Workshop on Argument Mining*. 83–91.
- [15] Angela E Douglas. 1994. *Symbiotic interactions*. Number 577.85 D733s. Oxon, GB: Oxford University Press, 1994.
- [16] Douglas C Engelbart. 1962. Augmenting human intellect: A conceptual framework. *Menlo Park, CA* (1962).
- [17] Leon Festinger. 1962. Cognitive dissonance. *Scientific American* 207, 4 (1962), 93–106.
- [18] Susan T Fiske and Shelley E Taylor. 2013. *Social cognition: From brains to culture*. Sage.
- [19] Gerd Gigerenzer and Reinhard Selten. 2002. *Bounded rationality: The adaptive toolbox*. MIT press.
- [20] Edward M Glaser. 1985. Educating for responsible citizenship in a democracy. In *National Forum*, Vol. 65. Honor Society of Phi Kappa Phi, 24.
- [21] Google. 2020. People + AI Guidebook: Explainability + Trust. (2020). <https://pair.withgoogle.com/chapter/explainability-trust/>
- [22] David Gunning. 2016. *Broad Agency Announcement Explainable Artificial Intelligence (XAI)*. Technical Report. Technical report.
- [23] Adam Haar Horowitz, Ishaan Grover, Pedro Reynolds-Cuellar, Cynthia Breazeal, and Pattie Maes. 2018. Dormio: Interfacing with Dreams. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, alt10.
- [24] Ivan Habernal and Iryna Gurevych. 2016. 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)*. 1589–1599.
- [25] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. *arXiv preprint arXiv:1802.06613* (2018).
- [26] Shohreh Haddadan, Serena Villata, and Elena Cabrio. 2019. Yes, we can! mining arguments in 50 years of us presidential campaign debates. (2019).
- [27] Ali Hasan and Richard Fumerton. 2018. Foundationalist Theories of Epistemic Justification. In *The Stanford Encyclopedia of Philosophy* (fall 2018 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.
- [28] Edward F Haskell. 1949. A clarification of social science. *Main Currents in Modern Thought* 7, 2 (1949), 45–51.
- [29] Robert R Hoffman, Gary Klein, and Shane T Mueller. 2018. Explaining Explanation For “Explainable AI”. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 62. SAGE Publications Sage CA: Los Angeles, CA, 197–201.
- [30] Andreas Holzinger, Bernd Malle, Peter Kieseberg, Peter M Roth, Heimo Müller, Robert Reihs, and Kurt Zatloukal. 2017. Towards the augmented pathologist: Challenges of explainable-ai in digital pathology. *arXiv preprint arXiv:1712.06657* (2017).
- [31] Jochen Huber, Roy Shilkrot, Pattie Maes, and Suranga Nanayakkara. 2018. *Assistive Augmentation*. Springer.
- [32] Fatih Incekara, Marion Smits, Clemens Dirven, and Arnaud Vincent. 2018. Clinical feasibility of a wearable mixed-reality device in neurosurgery. *World neurosurgery* 118 (2018), e422–e427.
- [33] Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.
- [34] Daniel Kahneman and Shane Frederick. 2002. Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment* 49 (2002), 81.
- [35] Daniel Kahneman, Stewart Paul Slovic, Paul Slovic, and Amos Tversky. 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.
- [36] Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018. Alterego: A personalized wearable silent speech interface. In *23rd International Conference on Intelligent User Interfaces*. ACM, 43–53.
- [37] Jonathan Kobbe, Juri Opitz, Maria Becker, Ioana Hulpus, Heiner Stuckenschmidt, and Anette Frank. 2019. Exploiting background knowledge for argumentative relation classification. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
- [38] Nataliya Kosmyna, Utkarsh Sarawgi, and Pattie Maes. 2018. AttentivU: Evaluating the Feasibility of Biofeedback Glasses to Monitor and Improve Attention. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. ACM, 999–1005.
- [39] HA Kusmantini, I Asror, and MA Bijaksana. 2019. 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*, Vol. 1192. IOP Publishing, 012009.
- [40] Mik Lamming and Mike Flynn. 1994. Forget-me-not: Intimate computing in support of human memory. In *Proc. FRIEND'21, 1994 Int. Symp. on Next Generation Human Interface*. 4.
- [41] 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, and others. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.
- [42] Jennifer S Lerner and Philip E Tetlock. 1999. Accounting for the effects of accountability. *Psychological bulletin* 125, 2 (1999), 255.
- [43] John Licato, Mark Boger, and Zhitian Zhang. 2018. Developing a Dataset for Personal Attacks and Other Indicators of Biases. In *2018 AAAI Spring Symposium Series*.
- [44] Joseph Carl Robnett Licklider. 1960. Man-computer symbiosis. *IRE transactions on human factors in electronics* 1 (1960), 4–11.
- [45] Matthias Liebeck, Katharina Esau, and Stefan Conrad. 2016. 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)*. 144–153.
- [46] Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)* 16, 2 (2016), 10.
- [47] Emily G Liquin and Tania Lombrozo. 2020. A functional approach to explanation-seeking curiosity. *Cognitive Psychology* 119 (2020), 101276.
- [48] Tania Lombrozo and Susan Carey. 2006. Functional explanation and the function of explanation. *Cognition* 99, 2 (2006), 167–204.
- [49] Anastasios Lytos, Thomas Lagkas, Panagiotis Sarigiannidis, and Kalina Bontcheva. 2019a. The evolution of argumentation mining: From models to social media and emerging tools. *Information Processing & Management* 56, 6 (2019), 102055.
- [50] Anastasios Lytos, Thomas Lagkas, Panagiotis Sarigiannidis, and Kalina Bontcheva. 2019b. The evolution of argumentation mining: From models to social media and emerging tools. *Information Processing & Management* 56, 6 (2019), 102055.
- [51] Pattie Maes. 1995. Agents that reduce work and information overload. In *Readings in human-computer interaction*. Elsevier, 811–821.
- [52] Sarah Mennicken, Jo Vermeulen, and Elaine M Huang. 2014. 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*. ACM, 105–115.
- [53] Nancy A Moran. 2007. Symbiosis as an adaptive process and source of phenotypic complexity. *Proceedings of the National Academy of Sciences* 104, suppl 1 (2007), 8627–8633.
- [54] Neville Moray. 1959. Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly journal of experimental psychology* 11, 1 (1959), 56–60.
- [55] Gaku Morio and Katsuhide Fujita. 2018. End-to-End Argument Mining for Discussion Threads Based on Parallel Constrained Pointer Architecture. *arXiv*

- preprint *arXiv:1809.00563* (2018).
- [56] Huy Nguyen and Diane Litman. 2016. Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1127–1137.
 - [57] Huy V Nguyen and Diane J Litman. 2018. Argument mining for improving the automated scoring of persuasive essays. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
 - [58] Brendan Nyhan, Ethan Porter, Jason Reifler, and Thomas J Wood. 2019. Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior* (2019), 1–22.
 - [59] Athina Papadopoulou, Jaclyn Berry, Terry Knight, and Rosalind Picard. 2019. Affective Sleeve: Wearable Materials with Haptic Action for Promoting Calmness. In *International Conference on Human-Computer Interaction*. Springer, 304–319.
 - [60] Seymour Papert. 2005. You can't think about thinking without thinking about thinking about something. *Contemporary Issues in Technology and Teacher Education* 5, 3 (2005), 366–367.
 - [61] Gordon Pennycook and David G Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188 (2019), 39–50.
 - [62] Georgios Petasis. 2019. Segmentation of Argumentative Texts with Contextualised Word Representations. In *Proceedings of the 6th Workshop on Argument Mining*. 1–10.
 - [63] Kashyap Kiritbhai Popat. 2019. Credibility analysis of textual claims with explainable evidence. (2019).
 - [64] Bradley J Rhodes. 1997. The wearable remembrance agent: A system for augmented memory. *Personal Technologies* 1, 4 (1997), 218–224.
 - [65] Alexander S Rich and Todd M Gureckis. 2019. Lessons for artificial intelligence from the study of natural stupidity. *Nature Machine Intelligence* 1, 4 (2019), 174.
 - [66] Rutu Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. 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*. Association for Computational Linguistics, Lisbon, Portugal, 440–450. DOI : <http://dx.doi.org/10.18653/v1/D15-1050>
 - [67] Salvador Rizzo. 2020. Analysis | Anatomy of a Trump rally: 67 percent of claims are false or lacking evidence. (Jan 2020). <https://www.washingtonpost.com/politics/2020/01/07/anatomy-trump-rally-percent-claims-are-false-or-lacking-evidence/>
 - [68] Oscar Rosello, Marc Exposito, and Pattie Maes. 2016. NeverMind: using augmented reality for memorization. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 215–216.
 - [69] Nancy L Rosenblum and Russell Muirhead. 2020. *A lot of people are saying: the new conspiracism and the assault on democracy*. Princeton University Press.
 - [70] Ishan Sahu. 2016. *A study on detecting fact vs non-fact in news articles*. Ph.D. Dissertation. Indian Statistical Institute, Kolkata.
 - [71] John R Searle and John Rogers Searle. 1969. *Speech acts: An essay in the philosophy of language*. Vol. 626. Cambridge university press.
 - [72] Upendra Shardanand and Pattie Maes. 1995. Social information filtering: algorithms for automating" word of mouth". In *Chi*, Vol. 95. Citeseer, 210–217.
 - [73] Herbert A Simon. 1956. Rational choice and the structure of the environment. *Psychological review* 63, 2 (1956), 129.
 - [74] Herbert Alexander Simon. 1997. *Models of bounded rationality: Empirically grounded economic reason*. Vol. 3. MIT press.
 - [75] Steven A Sloman. 1996. The empirical case for two systems of reasoning. *Psychological bulletin* 119, 1 (1996), 3.
 - [76] Noam Slonim. 2018. Project Debater. In *COMMA*. 4.
 - [77] Kacper Sokol and Peter Flach. 2020. One Explanation Does Not Fit All. *KI-Künstliche Intelligenz* (2020), 1–16.
 - [78] Keith E Stanovich. 2009. *What intelligence tests miss: The psychology of rational thought*. Yale University Press.
 - [79] Keith E Stanovich and Richard F West. 2008. On the failure of cognitive ability to predict myside and one-sided thinking biases. *Thinking & Reasoning* 14, 2 (2008), 129–167.
 - [80] Keith E Stanovich, Richard F West, and Maggie E Toplak. 2016. *The rationality quotient: Toward a test of rational thinking*. MIT press.
 - [81] Thad E Starner. 2002. Attention, memory, and wearable interfaces. *IEEE pervasive computing* 1, 4 (2002), 88–91.
 - [82] Oren M Tepper, Hayeem L Rudy, Aaron Lefkowitz, Katie A Weimer, Shelby M Marks, Carrie S Stern, and Evan S Garfein. 2017. Mixed reality with HoloLens: where virtual reality meets augmented reality in the operating room. *Plastic and reconstructive surgery* 140, 5 (2017), 1066–1070.
 - [83] Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. 2017. Toward a geographic understanding of the sharing economy: Systemic biases in UberX and TaskRabbit. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 3 (2017), 21.
 - [84] Ottokar Tilk and Tanel Alumäe. 2016. Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration. In *Interspeech 2016*.
 - [85] Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2019. Robust Argument Unit Recognition and Classification. *arXiv preprint arXiv:1904.09688* (2019).
 - [86] Davide Valeriani, Caterina Cinel, and Riccardo Poli. 2019. Hybrid Collaborative Brain-Computer Interfaces to Augment Group Decision-Making. In *Neuroergonomics*. Elsevier, 187–190.
 - [87] Jay J Van Bavel and Andrea Pereira. 2018. The partisan brain: An Identity-based model of political belief. *Trends in cognitive sciences* 22, 3 (2018), 213–224.
 - [88] Randall VanderMey. 2012. *Comp 2*. Wadsworth.
 - [89] Tomás Alfonso Vega Gálvez. 2019. *uJawstures: Jaw-teeth microgestures for discreet hands-and-eyes-free mobile device interaction*. Ph.D. Dissertation. Massachusetts Institute of Technology.
 - [90] Sunil Vemuri, Chris Schmandt, and Walter Bender. 2006. iRemember: a personal, long-term memory prosthesis. In *Proceedings of the 3rd ACM workshop on Continuous archival and retrieval of personal experiences*. ACM, 65–74.
 - [91] Rayoung Yang and Mark W Newman. 2013. 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*. ACM, 93–102.