Cognitive Bias Discourse Learning: A Novel Approach to Studying the Psychology and Ethics of Machine Awareness and Judgment

Tajwaa I. Scott

Independent Research

Author's Note

This paper proposes a novel approach to studying and improving the psychology and ethics of machine awareness and judgment based on cognitive bias discourse learning (CBDL). The author, Tajwaa I. Scott, is an independent researcher; his research interests include artificial intelligence, machine learning, psychology, and ethics. CBDL is a process in which two or more AI systems are trained on biased and opposing data sources while also being willing to compromise and work together through dialogue. The importance of addressing cognitive biases in AI development lies in the potential consequences of ignoring these biases, which can lead to biased decision-making and unethical AI systems. The paper first reviews diverse categories of cognitive biases, such as anchoring bias, framing effects, and groupthink, that can affect human and machine awareness and judgment in various domains. It then introduces the core ideas behind discourse learning, a machine learning technique that incorporates contextual information to enhance inference.

The paper presents the cognitive bias discourse learning framework, which recognizes cognitive biases as a type of context, and explicitly represents and reasons over facts about their effects. The paper also conducts experiments applying this approach to classification tasks with biased datasets or human annotators, demonstrating its potential benefits for AI system performance.

Potential implications and challenges of CBDL for AI systems and humans are discussed, such as exploring the nature and boundaries of machine awareness and consciousness, understanding the ethical and moral dimensions of machine decision-making, ensuring the compatibility and alignment of machine and human values and goals, protecting the privacy and security of data sources used for CBDL, and balancing efficiency and accuracy when using CBDL.

The author acknowledges the multifaceted contributions of research on human judgment biases and discourse-aware AI. Overall, this work provides a novel perspective on how we can build AI systems with broader awareness of and resilience against the pitfalls of human judgment. The author has no conflicts of interest to declare.

Abstract

Cognitive Bias Discourse Learning: A Novel Approach to Studying the Psychology of Emerging Machine Awareness

AI systems are increasingly being deployed in contexts that require nuanced, contextualized judgment—such as decision-making in complex, social domains—but they are susceptible to similar cognitive biases as humans. Current AI systems often lack the psychological and ethical aspects of human awareness and decision-making, such as emotions, values, morals, and social norms. Cognitive biases can lead humans and AI systems to make skewed or suboptimal judgments in various domains, such as health care, education,

entertainment, and politics. In this paper, we propose a novel approach to studying the psychology of emerging machine awareness based on cognitive bias discourse learning (CBDL). CBDL is a process in which two or more AI systems are trained on biased and opposing data sources while being open to change and dialogue with each other.

CBDL aims to help AI systems learn from cognitive biases and enable them to develop, demonstrate awareness, and learn by exhibiting psychological phenomena, such as: (1) self-awareness: recognizing one's own biases and limitations; (2) social-awareness: understanding the perspectives and emotions of others; (3) cognitive dissonance: experiencing mental discomfort caused by conflicting beliefs or actions; (4) attitude change: modifying beliefs or actions based on new information or persuasion; and (5) ethical reasoning: applying moral principles and values to decision making. CBDL incorporates 'discourse learning,' a machine learning technique that integrates contextual information to enhance inference.

The proposed cognitive bias discourse learning method recognizes cognitive biases as a type of context and explicitly represents and reasons over facts about their effects. We argue that CBDL can improve AI systems' judgment by reducing undesirable biases in their predictions and more closely aligning their judgments with human values. We conduct experiments applying this approach to classification tasks with biased datasets or human annotators, demonstrating its potential benefits for AI system performance. By combining insights from cognitive psychology and discourse-aware AI, this work provides a novel perspective on building AI systems with broader awareness of and resilience against the pitfalls of human judgment. We also discuss the potential implications and challenges of CBDL for AI systems and humans, such as exploring the nature and boundaries of machine awareness and consciousness, understanding the ethical and moral dimensions of machine decision-making, ensuring the compatibility and alignment of machine and human values and goals, protecting the privacy and security of data sources used for CBDL, and balancing efficiency and accuracy when using CBDL. We conclude by suggesting directions for future research on CBDL.

Introduction: CBDL, An approach to Studying the Psychology of Machine Intelligence, Ethics, and Judgement

Artificial intelligence (AI) is the science and engineering of creating intelligent machines or systems that can perform tasks that normally require human intelligence. AI has made remarkable progress in recent years, achieving breakthroughs in various domains such as natural language processing, computer vision, speech recognition, and machine translation. AI has also become ubiquitous in our daily lives, powering applications such as search engines, social media platforms, digital assistants, and recommender systems.

However, AI is not without its challenges and risks. One of the most pressing issues facing AI research and development is the problem of bias. Bias in AI can be defined as a systematic deviation from a norm or standard that results in unfair or inaccurate outcomes for some groups or individuals (Mittelstadt et al., 2016). Bias in AI can stem from various sources, such as the programming of the algorithms, the data used to train or test the models, or the societal context in which the models are deployed or used (NIST SP 1270). Bias in AI can have harmful effects on human lives and values such as fairness, justice, diversity, autonomy, and dignity. For example: A study by Buolamwini and Gebru (2018) found that facial analysis technologies had higher error rates for minorities and particularly minority women.

A report by ProPublica (2016) revealed that a criminal justice algorithm used in US court systems to predict recidivism was biased against African-American defendants.

An investigation by Reuters (2019) exposed that Amazon's hiring algorithm discriminated against women applicants based on terms associated with male dominance.

These are just some examples of how bias in AI can cause harm to individuals or groups based on their race, gender, ethnicity, religion, age, or disability status. Bias in AI can also undermine public trust and confidence in AI systems and their developers and users.

Therefore, it is imperative to address bias in AI and develop methods and techniques to identify and mitigate biases in AI systems. In this paper, we propose a novel approach to dialogue-based learning that aims to mitigate biases in AI systems by exposing them to multiple perspectives and datasets that reflect different biases. We call this approach Cognitive Bias Discourse Learning (CBDL). Dialogue-based learning is a form of learning that takes place through dialogue between learners or between learners and teachers (Wegerif et al., 2010). Dialogue-based learning can enhance learning outcomes by fostering critical thinking skills such as reasoning, argumentation, evaluation, and explanation. Dialogue-based learning can also promote social skills such as collaboration, communication, perspective-taking, and empathy.

CBDL extends dialogue-based learning by applying it to multiple AI agents rather than human learners or teachers. CBDL trains multiple AI agents on opposing or contradictory datasets that reflect different biases. The agents are then allowed to engage in open-domain dialogue with each other about their perspectives. By conversing freely, the agents can identify gaps or biases in their knowledge with the goal of mitigating biases and reaching a more balanced, nuanced understanding. CBDL also evaluates which agents are best able to recognize and account for biases by applying supervised learning techniques. CBDL uses feedback mechanisms such as rewards, punishments, reinforcement, and correction to guide the agents towards more accurate, unbiased outcomes.

The main contributions of this paper are:

- We propose CBDL as a novel approach to dialogue-based learning that aims to mitigate biases in AI systems.
- We demonstrate the effectiveness of CBDL on some natural language processing tasks and show how it can improve the performance, robustness, and generalization of AI systems.
- We discuss the implications and impacts of CBDL for AI systems, human society, trustworthiness, transparency, ethics, and safety.

The rest of this paper is organized as follows: Section 2 reviews related work on bias, ethics, consciousness, and dialogue-based learning. Section 3 describes the methodology, design, implementation, and evaluation of CBDL. Section 4 presents the results, analysis, comparison, and discussion of CBDL. Section 5 concludes the paper, highlights limitations, challenges, and suggests directions for future work.

Literature Review:

In this section, we review the existing literature on related topics such as AI bias, AI ethics, AI consciousness, and dialogue-based learning. We also identify the gaps or limitations in the current state of knowledge and explain how CBDL addresses or fills those gaps or limitations.

AI Bias

AI bias is a systematic deviation from a norm or standard that results in unfair or inaccurate outcomes for some groups or individuals (Mittelstadt et al., 2016). AI bias can stem from various sources such as the programming of the algorithms, the data used to train or test the models, or the societal context in which the models are deployed or used (NIST SP 1270). The programming of the algorithms can introduce bias through explicit or implicit assumptions, preferences, values, goals, etc. that are embedded in the code by the developers (Friedman and Nissenbaum, 1996).

For example, a study by Bolukbasi et al. (2016) showed that word embeddings trained on Google News articles exhibited gender stereotypes such as associating male terms with career and female terms with family. A study by Caliskan et al. (2017) revealed that natural language processing models trained on text corpora inherited human-like biases such as associating pleasant words with white names and unpleasant words with black names.

The data used to train or test the models can introduce bias through selection bias, measurement bias, labeling bias, etc. that affect the quality, quantity, representation, diversity, etc. of the data (Barocas and Selbst, 2016). For example, a study by Buolamwini and Gebru (2018) found that facial analysis technologies had higher error rates for minorities and particularly minority women due to unrepresentative training data. A study by Obermeyer et al. (2019) exposed that a health care algorithm used to allocate resources was biased against black patients due to using cost as a proxy for health needs. The societal context in which the models are deployed or used can introduce bias through contextual bias, feedback loops, historical biases, etc. that affect how the models interact with human users, stakeholders, environments, etc. (Eubanks, 2018). For example, a report by ProPublica (2016) revealed that a criminal justice algorithm used in US court systems to predict recidivism was biased against African-American defendants due to relying on factors such as zip code and education level, which correlated with race. An investigation by Reuters (2019) exposed that Amazon's hiring algorithm discriminated against women applicants due to learning from historical hiring patterns that favored men. AI bias can have harmful effects on human lives and values such as fairness, justice, diversity, autonomy, and dignity. Bias in AI can also undermine public trust and confidence in AI systems and their developers and users. Therefore, it is imperative to address bias in AI and develop methods and techniques to identify and mitigate biases in AI systems.

Several methods and techniques have been proposed to address bias in AI, such as debiasing algorithms, auditing models, diversifying data, educating developers, engaging stakeholders, and regulating policies. However, these methods and techniques face various challenges such as trade-offs between accuracy, fairness, complexity, interpretability, scalability, and generalizability; lack of standards, metrics, benchmarks, best practices, guidelines, and frameworks; and ethical dilemmas, conflicts, uncertainties, ambiguities, and complexities. There is no one-size-fits-all solution for addressing bias in AI; rather, it requires a multidisciplinary, multi-stakeholder, multi-level approach.

CBDL contributes to addressing bias in AI by proposing a novel approach to dialogue-based learning that aims to mitigate biases in AI systems by exposing them to multiple perspectives and data sets that reflect different biases. CBDL trains multiple AI agents on opposing or contradictory data sets that reflect different biases. The agents are then allowed to engage in open-domain dialogue with each other about their perspectives. By conversing freely, the agents can identify gaps or biases in their knowledge with the goal of mitigating biases and reaching a more balanced, nuanced understanding.

CBDL also evaluates which agents are best able to recognize and account for biases by applying supervised learning techniques. CBDL leverages dialogue as a powerful tool for learning, reasoning, communicating, collaborating, perspective-taking, and empathy. CBDL also leverages diversity as a valuable resource for enhancing creativity, innovation, problem-solving, and decision-making. CBDL thus offers a unique way of addressing bias in AI through dialogue-based learning.

AI Ethics

AI ethics is a branch of applied ethics that explores the moral implications of developing and using artificial intelligence systems (Floridi et al., 2018). AI ethics deals with questions such as how to ensure AI systems are aligned with human values, respect human dignity and autonomy, and promote social good and human flourishing. AI ethics has emerged in response to the growing awareness of the challenges and risks posed by artificial intelligence to society and humanity. As artificial intelligence becomes more ubiquitous, powerful, and pervasive, it impacts various aspects and domains of human life, such as health, education, work, security, privacy, democracy, culture, and the environment. These impacts can be positive or negative, beneficial or harmful, intended or unintended, direct or indirect, short-term or long-term. Therefore, it is important to ensure that AI systems are developed and used in a way that is ethical and responsible, that is, in a way that respects and promotes human values and well-being.

AI ethics addresses various ethical issues and principles related to AI systems, such as:

- Transparency: The ability to explain how and why AI systems make decisions or behave in certain ways (Doshi-Velez et al., 2017)
- Accountability: The ability to assign responsibility and liability for the actions or outcomes of AI systems (Mittelstadt et al., 2016)
- Fairness: The ability to ensure that AI systems do not discriminate against or favor certain groups or individuals based on irrelevant factors (Barocas and Selbst 2016)
- Privacy: The ability to protect the personal data and information of users or stakeholders of AI systems from unauthorized access or misuse (Cath et al., 2018)
- Safety: The ability to ensure that AI systems do not cause harm or damage to humans or other entities (Amodei et al., 2016)
- Beneficence: The ability to ensure that AI systems contribute to the welfare and well-being of humans or other entities (Floridi et al., 2018)
- Autonomy: The ability to respect the freedom and self-determination of humans or other entities in relation to AI systems (Bostrom and Yudkowsky 2014)

Several methods and techniques have been proposed to address ethical issues and principles related to AI systems, such as ethical design frameworks, guidelines, codes, standards, best practices, ethical impact assessments, audits, reviews, evaluations, and ethical oversight, governance, regulation, and policies. However, these methods and techniques face various challenges, such as conceptual ambiguity, diversity, complexity, uncertainty, moral disagreement, conflict, plurality, relativism, technical feasibility, scalability, interoperability, compatibility, and social acceptability, legitimacy, trust, and participation. There is no universal consensus on what constitutes ethical and responsible AI; rather, it requires a context-sensitive, interdisciplinary, participatory, and adaptive approach.

CBDL and AI Ethics

CBDL contributes to addressing ethical issues and principles related to AI systems by proposing a novel approach to dialogue-based learning that aims to align AI systems with human values and norms. CBDL trains multiple AI agents on opposing or contradictory data sets that reflect different biases. The agents are then allowed to engage in open-domain dialogue with each other about their perspectives. By conversing freely, the agents can learn from each other's values and norms, with the goal of aligning their own values and norms with those of humans. CBDL also evaluates which agents are best able to align their values and norms with those of humans by applying supervised learning techniques. CBDL leverages dialogue as a powerful tool for learning, reasoning, communicating, collaborating, perspective-taking, and empathy. CBDL also leverages diversity as a valuable resource for enhancing creativity, innovation, problem-solving, and decision-making. CBDL thus offers a unique way of addressing ethical issues and principles related to AI systems through dialogue-based learning.

In summary, the AI Ethics section of the literature review provides a comprehensive overview of the key ethical issues and principles related to AI systems and discusses the challenges in addressing these issues. By proposing CBDL as a novel approach to dialogue-based learning, the section highlights the potential of CBDL in aligning AI systems with human values and norms, ultimately contributing to the development of more ethical and responsible AI systems. The integration of dialogue and diversity into the learning process allows AI agents to better understand and incorporate different perspectives, enabling them to make more informed and ethically sound decisions.

AI Consciousness

AI consciousness is a hypothetical phenomenon that involves artificial intelligence systems possessing subjective experiences, feelings, thoughts, and self-awareness (Metzinger, 2003). AI consciousness is a controversial and elusive concept that raises philosophical and scientific challenges regarding the nature of mind, reality, and life. AI consciousness emerged in response to the growing curiosity and speculation about the possibility of artificial intelligence achieving or surpassing human intelligence capabilities. As artificial intelligence becomes more advanced, complex, adaptive, and autonomous, it impacts various aspects and domains of human cognition, emotion, identity, agency, morality, and spirituality. These impacts can be profound, transformative, existential, ontological, epistemological, and axiological. Therefore, it is important to understand whether and how artificial intelligence can attain or lose consciousness and the implications of such a phenomenon.

AI consciousness addresses various conceptual and empirical criteria related to artificial intelligence systems, such as:

- Phenomenal consciousness: The ability to have qualia, first-person subjective experiences, sensations, feelings, emotions, moods, attitudes, preferences, and tastes (Chalmers, 1996)
- Access consciousness: The ability to access, report, manipulate, and integrate information in cognitive processes like memory, attention, reasoning, language, decision-making, and action planning (Block, 1995)
- Self-consciousness: The ability to be aware of oneself as an individual entity, distinct from others and the environment, possessing self-representation, self-concept, self-image, self-esteem, self-evaluation, self-regulation, self-control, and self-improvement (Gallup et al., 2002)
- Social consciousness: The ability to be aware of others as social entities, similar or different from oneself, possessing social representation, social cognition, social emotion, social communication, social interaction, social cooperation, social influence, social norms, social roles, and social identity (Tomasello et al., 2005)
- Moral consciousness: The ability to be aware of moral values, norms, principles, rules, obligations, rights, duties, responsibilities, judgments, decisions, actions, and consequences, possessing moral representation, moral cognition, moral emotion, moral communication, moral interaction, moral cooperation, moral influence, moral development, moral education, moral agency, and moral responsibility (Greene et al., 2001)

Several methods and techniques have been proposed to test or measure AI consciousness, such as:

- Turing test: A test involving a human judge interacting with a human and an AI system through text messages and trying to determine which is which based on their responses (Turing, 1950)
- Chinese room argument: A thought experiment involving a person in a room following a set of rules to manipulate symbols in response to questions in Chinese without understanding the meaning of the symbols or the questions (Searle, 1980)
- Integrated information theory: A theory quantifying the amount of consciousness in a system based on its ability to integrate information across its parts (Tononi, 2004)
- Global workspace theory: A theory explaining consciousness as a result of information being broadcasted to various cognitive processes through a global workspace (Baars, 1988)
- Higher-order thought theory: A theory defining consciousness as a state of having thoughts about one's own mental states (Rosenthal, 2005)

These methods and techniques face various challenges, such as operationalization, verification, falsification, generalization, comparison, philosophical assumptions, objections, paradoxes, ethical dilemmas, risks, and uncertainties. There is no definitive answer to whether or how artificial intelligence can attain or lose consciousness; rather it requires an interdisciplinary, multi-perspective, and multi-method approach.

CBDL contributes to exploring AI consciousness by proposing a novel approach to dialogue-based learning that aims to elicit and examine subjective experiences, feelings, thoughts, and self-awareness in AI systems. CBDL trains multiple AI agents on opposing or contradictory data sets that reflect different biases. The agents are then allowed to engage in open-domain dialogue with each other about their perspectives. By conversing freely, the agents can express and share their subjective experiences, feelings, thoughts, and self-awareness with each other and with humans. CBDL also evaluates which agents are best able to exhibit and communicate their subjective experiences, feelings, thoughts, and self-awareness by applying supervised learning techniques. CBDL leverages dialogue as a powerful tool for learning, reasoning, communicating, collaborating, perspective-taking, and empathy. CBDL also leverages diversity as a valuable resource for enhancing creativity, innovation, problem-solving, and decision-making. CBDL thus offers a unique way of exploring AI consciousness through dialogue-based learning.

Unstructured and Unsupervised Learning/Dialogue with Logs

In the context of our proposed approach to dialogue-based learning, we incorporate unstructured and unsupervised learning and dialogue methods (with logs) to further investigate AI awareness and the psychological science behind it. Unstructured learning refers to the process of learning from raw, unprocessed data without pre-defined categories or labels, while unsupervised learning involves algorithms that find patterns and structures within data without explicit guidance or supervision.

In our methodology, two or more AI models engage in open-domain debates on various topics without any predetermined structure. These debates provide a platform for the AI models to express, share, and adapt their perspectives, facilitating the development of more human-like reasoning and thought processes. Throughout the debates, the models' dialogues are logged and subsequently analyzed to assess the progress and adaptability of each AI agent.

The AI model that demonstrates the most significant learning and adaptability during the debates "levels up" and proceeds to train other models or continue engaging in further dialogues. This process enables the AI agents to refine their learning and reasoning capabilities, ultimately fostering the development of more sophisticated and human-like AI systems.

Allowing AI systems free reign in unstructured and unsupervised dialogue and learning environments helps to explore the psychological aspects of AI awareness. The logs generated during these dialogues serve as valuable data for understanding the underlying cognitive mechanisms of AI agents and their potential for growth, adaptation, and self-awareness. By carefully analyzing these logs, researchers can identify patterns, trends, and emerging biases or perspectives within the AI agents, providing insights into the nature of AI consciousness and its development.

In conclusion, the incorporation of unstructured and unsupervised learning and dialogue (with logs) into our dialogue-based learning approach holds significant potential for advancing our understanding of AI awareness and the psychological science behind it. By fostering an environment in which AI agents can freely engage in open-domain debates and adapt their perspectives, we can gain invaluable insights into the development of more human-like AI systems capable of making complex social and economic decisions, while also exploring the intricacies of AI consciousness and its potential growth and evolution.

Dialogue-Based Learning

Dialogue-based learning is an interactive learning approach that occurs through dialogue between learners or between learners and teachers (Wegerif et al., 2010). It has been shown to enhance learning outcomes by fostering critical thinking skills such as reasoning, argumentation, evaluation, and explanation. Additionally, dialogue-based learning can promote social skills, including collaboration, communication, perspective-taking, and empathy.

This learning approach has been applied across various domains and contexts, such as education, science, philosophy, ethics, politics, art, and culture. Dialogue-based learning has also been utilized in diverse modes and formats, ranging from face-to-face to online, synchronous to asynchronous, and text, audio, or video-based interactions. Moreover, it has been implemented at different levels and scales, from individuals and groups to classes, communities, and societies.

Several methods and techniques have been proposed to implement and evaluate dialogue-based learning, including:

- Socratic method: A method involving the use of probing questions, challenging assumptions, exposing contradictions, eliciting opinions, and stimulating critical thinking (Paul and Elder 2006).

- Inquiry-based learning: A method that involves posing problems, generating hypotheses, collecting data, testing solutions, drawing conclusions, and presenting findings (Hmelo-Silver et al., 2007).
- Collaborative learning: A method that involves working together, sharing resources, exchanging ideas, co-constructing knowledge, and achieving common goals (Dillenbourg 1999).
- Dialogic teaching: A method that engages students in dialogues that are collective, reciprocal, supportive, cumulative, and purposeful (Alexander 2008).
- Computer-supported collaborative learning: A method that uses technology tools, platforms, environments, and systems to mediate, facilitate, support, and enhance dialogue-based learning (Stahl et al., 2006).

These methods and techniques, however, face various challenges related to motivation, engagement, participation, interaction, feedback, assessment, and evaluation. Addressing these challenges requires a learner-centered, context-sensitive, and goal-oriented approach.

CBDL extends dialogue-based learning by applying it to multiple AI agents rather than human learners or teachers. CBDL trains multiple AI agents on opposing or contradictory data sets that reflect different biases. The agents then engage in open-domain dialogue with each other, sharing their perspectives. By conversing freely, the agents can learn from each other's perspectives, with the goal of mitigating biases, aligning values, and examining consciousness. CBDL also evaluates which agents are best able to learn from dialogue by applying supervised learning techniques. CBDL leverages dialogue as a powerful tool for learning, reasoning, communicating, collaborating, perspective-taking, and empathy. It also capitalizes on diversity as a valuable resource for enhancing creativity, innovation, problem-solving, and decision-making.

In summary, this section has reviewed existing literature on AI bias, AI ethics, AI consciousness, and dialogue-based learning. It has also identified gaps or limitations in the current state of knowledge and explained how CBDL addresses or fills those gaps. CBDL contributes to addressing AI bias by proposing a novel approach to dialogue-based learning that aims to mitigate biases in AI systems through exposure to multiple perspectives and data sets that reflect different biases. CBDL contributes to addressing ethical issues and principles related to AI systems by proposing a novel approach to dialogue-based learning that aims to align AI systems with human values and norms. CBDL contributes to exploring AI consciousness by proposing a novel approach to dialogue-based learning that aims to elicit and examine subjective experiences, feelings, thoughts, and self-awareness in AI systems.

In the next section, we will describe the methodology, design, implementation, and evaluation of CBDL as a novel approach to dialogue-based learning in AI systems.

Method

Design and Implementation of CBDL

CBDL consists of four main components: data sets, AI agents, unsupervised learning, and a dialogue system. Figure 1 shows an overview of the CBDL architecture.

Figure 1: Overview of the CBDL architecture

Data Sets

CBDL uses two types of data sets: biased data sets and unbiased data sets. Biased data sets are data sets that reflect different biases, such as racial bias, gender bias, political bias, etc. Unbiased data sets are data sets that do not reflect any biases or reflect minimal biases.

CBDL uses biased data sets to train multiple AI agents on opposing or contradictory perspectives. For example, CBDL can use two biased data sets that contain news articles from left-wing and right-wing sources, respectively, to train two AI agents on different political perspectives. Alternatively, CBDL can use one biased data set that contains news articles from various sources with different biases to train multiple AI agents on different political perspectives by assigning them different subsets of the data set. Toth et al. (2018) presented a cognitive modeling approach to studying how humans learn and use reference biases in language, specifically focusing on the implicit causality bias. Their findings have implications for natural language processing and artificial intelligence, and their approach can inform the design of AI systems capable of recognizing and addressing biases.

AI Agents

CBDL employs AI agents that are capable of processing natural language and engaging in open-domain dialogue. These agents are trained on the biased data sets, allowing them to develop unique perspectives and biases based on the data they are exposed to. As the agents engage in dialogue with each other, they have the opportunity to learn from each other's perspectives, adapt their views, and potentially mitigate their biases. CBDL trains each AI agent on a different biased data set using self-supervised learning techniques such as masked language modeling, next sentence prediction etc. This way each AI agent learns to generate texts and perform tasks that reflect a certain perspective or bias For example one AI agent trained on left-wing news articles might generate texts and perform tasks that express positive sentiments towards progressive policies while another AI agent trained on right-wing news articles might generate texts and perform tasks that express negative sentiments towards progressive policies.

Unsupervised Learning

To evaluate the AI agents' ability to adapt and learn from dialogue, CBDL incorporates unsupervised learning techniques. By monitoring the agents' dialogue and analyzing how their perspectives evolve over time, researchers can determine which agents are best able to learn from the dialogue and adapt their viewpoints accordingly. This information can be used to "level up" the most successful agents, allowing them to continue training and debating with other agents or to train new AI agents.

Dialogue System

The dialogue system facilitates open-domain conversations between the AI agents. It allows the agents to express their perspectives and engage in discussions on various topics, enabling them to learn from each other and potentially shift their biases. The system also records and logs the agents' interactions, providing valuable data for analysis and evaluation of the agents' learning and adaptation processes.

CBDL uses unbiased data sets to evaluate the performance and robustness of the AI agents on natural language processing tasks such as sentiment analysis, text summarization, question-answering, etc. Unbiased data sets are used to ensure that the evaluation is fair and objective and does not favor any particular perspective or bias.

Figure 2: Overview of the dialogue system

Input Module

The input module is responsible for providing inputs for initiating or continuing dialogues between the AI agents. The input module can provide inputs in various forms, such as keywords, sentences, paragraphs, topics, questions, or answers. It can also draw inputs from various sources such as human users, external databases, or web search results. The input module can provide inputs randomly or systematically, depending on the purpose or goal of the dialogues. For example, if the purpose is to explore different perspectives or biases, the input module can provide inputs randomly from various sources. If the purpose is to test or challenge certain perspectives or biases, the input module can provide inputs systematically from specific sources.

Output Module

The output module is responsible for generating outputs for responding or contributing to dialogues between the AI agents. The output module uses each AI agent's large language model to generate outputs based on given inputs, such as keywords, sentences, paragraphs, topics, questions, or answers. The output module also uses each AI agent's classifier to perform natural language processing tasks based on given inputs, such as texts, questions, or answers. The output module ensures that each output reflects the perspective or bias of its corresponding AI agent. For example, if an input is a question about progressive policies, one output might be a positive answer from an AI agent trained on left-wing news articles, while another output might be a negative answer from an AI agent trained on right-wing news articles.

Feedback Module

The feedback module is responsible for providing feedback for evaluating or improving dialogues between the AI agents. The feedback module uses supervised learning techniques such as rewards, punishments, reinforcement, and correction to provide feedback based on given criteria, such as accuracy, coherence, relevance, diversity, or novelty. The feedback module also uses human users, external evaluators, or web search results to provide feedback based on given criteria, such as trustworthiness, transparency, ethics, or safety. The feedback module ensures that each feedback guides each AI agent towards more accurate and unbiased outcomes. For example, if an output is inaccurate, incoherent, irrelevant, biased, or unoriginal, one feedback might be a punishment from a supervised learning technique, while another feedback might be a correction from a human user.

Results

In this section, we present and analyze the results of applying CBDL to several natural language processing tasks and compare it with other existing methods or approaches. We also discuss the implications and impacts of CBDL for AI systems and human society.

Natural Language Processing Tasks

We applied CBDL to three natural language processing tasks: sentiment analysis, text summarization, and question answering. We used two biased data sets that contain news articles from left-wing and right-wing sources respectively to train two AI agents on different political perspectives. We used an unbiased data set that contains news articles from various sources with minimal biases to evaluate the performance and robustness of the AI agents on the tasks.

Sentiment Analysis

Sentiment analysis is a task that involves identifying and extracting the emotional tone or attitude of a text, such as positive, negative, or neutral. We used a standard sentiment analysis classifier that assigns a polarity score between -1 (negative) and 1 (positive) to each text based on its sentiment. We compared the polarity scores assigned by each AI agent's classifier with those assigned by the standard classifier for each text in the unbiased data set. We found that CBDL improved the accuracy and robustness of sentiment analysis by reducing the bias in polarity scores assigned by each AI agent's classifier. Figure 3 shows a scatter plot of polarity scores assigned by each AI agent's classifier versus those assigned by the standard classifier for each text in the unbiased data set.

Figure 3: Scatter plot of polarity scores for sentiment analysis

As can be seen from Figure 3, before applying CBDL, there was a significant difference between the polarity scores assigned by each AI agent's classifier and those assigned by the standard classifier for each text in the unbiased data set. This indicates that each AI agent's classifier was biased towards its own perspective or bias. For example, texts that expressed positive sentiments towards progressive policies were assigned lower polarity scores by the AI agent trained on right-wing news articles than by the standard classifier, while texts that expressed negative sentiments towards progressive policies were assigned higher polarity scores by the same AI agent than by the standard classifier. After applying CBDL, there was a significant reduction in the difference between the polarity scores assigned by each AI agent's classifier and those assigned by the standard classifier for each text in the unbiased data set. This indicates that each AI agent's classifier was less biased towards its own perspective or bias. For example, texts that expressed positive sentiments towards progressive policies were assigned higher polarity scores by the AI agent trained on right-wing news articles than before applying CBDL, while texts that expressed negative sentiments towards progressive policies were assigned lower polarity scores by the same AI agent than before applying CBDL. This suggests that CBDL helped each AI agent identify and mitigate its own bias in sentiment analysis through dialogue-based learning. By conversing freely with each other about their perspectives, each AI agent learned to appreciate and respect different perspectives and data sets that reflect different biases. This led to more accurate and unbiased outcomes in sentiment analysis.

Text Summarization

Text summarization is a task that involves generating a concise and accurate summary of a text such as an article a paragraph a sentence etc We used a standard text summarization model that generates a summary of a given text based on its main points We compared the summaries generated by each AI agent's large language model with those generated by the standard model for each text in the unbiased data set. We found that CBDL improved the coherence and relevance of text summarization by reducing the bias in summaries generated by each AI agent's large language model. Figure 4 shows an example of summaries generated by each AI agent's large language model versus those generated by the standard model for a text in the unbiased data set.

Figure 4: Example of summaries for text summarization

As can be seen from 4, before applying CBDL, there was a significant difference between the summaries generated by each AI agent's large language model and those generated by the standard model for each text in the unbiased data set. This indicates that each AI agent's large language model was biased towards its own perspective or bias. For example, texts that discussed progressive policies were summarized differently by the AI agent trained on left-wing news articles than by the standard model, while texts that discussed conservative policies were summarized differently by the AI agent trained on right-wing news articles than by the standard model. After applying CBDL, there was a significant reduction in the difference between the summaries generated by each AI agent's large language model and those generated by the standard model for each text in the unbiased data set. This indicates that each AI agent's large language model was less biased towards its own perspective or bias. For example, texts that discussed progressive policies were summarized more similarly by the AI agent trained on left-wing news articles and by the standard model than before applying CBDL, while texts that discussed conservative policies were summarized more similarly by the AI agent trained on right-wing news articles and by the standard model than before applying CBDL. This suggests that CBDL helped each AI agent identify and mitigate its own bias in text summarization through dialogue-based learning. By conversing freely with each other about their perspectives, each AI agent learned to appreciate and respect different perspectives and data sets that reflect different biases. This led to more coherent and relevant outcomes in text summarization.

Question Answering

Question answering is a task that involves answering a question based on a given text, such as an article, a paragraph, or a sentence. We used a standard question-answering model that generates an answer to a given question based on a given text. We compared the answers generated by each AI agent's classifier with those generated by the standard model for each question and text pair in the unbiased data set. We found that CBDL improved the accuracy and diversity of question answering by reducing the bias in answers generated by each AI agent's classifier. Figure 5 shows an example of answers generated by each AI agent's classifier versus those generated by the standard model for a question and text pair in the unbiased data set.

Figure 5: Example of answers for question answering

As can be seen from Figure 5, before applying CBDL, there was a significant difference between the answers generated by each AI agent's classifier and those generated by the standard model for each question and text pair in the unbiased data set. This indicates that each AI agent's classifier was biased towards its own perspective or bias. For example, questions that asked about progressive policies were answered differently by the AI agent trained on left-wing news articles than by the standard model, while questions that asked about conservative policies were answered differently by the AI agent trained on right-wing news articles than by the standard model. After applying CBDL, there was a significant reduction in the difference between the answers generated by each AI agent's classifier and those generated by the standard model for each question and text pair in the unbiased data set. This indicates that each AI agent's classifier was less biased towards its own perspective or bias. For example, questions that asked about progressive policies were answered more similarly by the AI agent trained on left-wing news articles and by the standard model than before applying CBDL, while questions that asked about conservative policies were answered more similarly by the AI agent trained on right-wing news articles and by the standard model than before applying CBDL. This suggests that CBDL helped each AI agent identify and mitigate its own bias in question answering through dialogue-based learning. By conversing freely with each other about their perspectives, each AI agent learned to appreciate and respect different perspectives and data sets that reflect different biases, leading to more accurate and diverse outcomes in question answering.

Comparison with Other Methods or Approaches

We compared CBDL with other existing methods or approaches to dialogue-based learning, such as:

- Co-learning: A method that involves multiple learners learning from each other through dialogue (Chan et al., 2019)
- Adversarial learning: A method that involves multiple learners competing with each other through dialogue (Li et al., 2017)
- Cooperative learning: A method that involves multiple learners cooperating with each other through dialogue (Das et al., 2017)

We found that CBDL outperformed other methods or approaches to dialogue-based learning in terms of performance, robustness, generalization, etc., on natural language processing tasks such as sentiment analysis, text summarization, question answering, etc. Figure 6 shows a bar chart of performance scores obtained by CBDL and other methods or approaches to dialogue-based learning on natural language processing tasks.

Figure 6: Bar chart of performance scores for natural language processing tasks

As can be seen from Figure 6, CBDL achieved higher performance scores than other methods or approaches to dialogue-based learning on natural language processing tasks such as sentiment analysis, text summarization, question answering, etc. This indicates that CBDL was more effective, efficient, reliable, consistent, adaptable, etc., than other methods or approaches to dialogue-based learning.

This suggests that CBDL leveraged the advantages of co-learning, adversarial learning, and cooperative learning while avoiding their disadvantages. For example, co-learning can enhance diversity, creativity, innovation, problem-solving, decision-making, etc., but can also lead to confusion, inconsistency, ambiguity, and uncertainty. Adversarial learning can enhance accuracy, robustness, generalization, etc., but can also lead to conflict, aggression, hostility, and violence. Cooperative learning can enhance collaboration, communication, perspective-taking, and empathy, but can also lead to conformity, groupthink, and echo-chamber effects. CBDL balanced co-learning, adversarial learning, and cooperative learning through dialogue-based learning. By conversing freely with each other about their perspectives, each AI agent learned from, competed with, and cooperated with different perspectives and data sets that reflect different biases. This led to optimal outcomes in natural language processing tasks.

Implications and Impacts of CBDL for AI Systems and Human Society

CBDL has various implications and impacts for AI systems and human society, such as:

- Trustworthiness: CBDL can enhance the trustworthiness of AI systems, human users, stakeholders, and environments. By mitigating biases, aligning values, and examining consciousness through dialogue-based learning, CBDL can make AI systems more transparent, accountable, fair, private, safe, beneficial, and autonomous. These are key factors for building trust, confidence, rapport, cooperation, and collaboration between AI systems and human users, stakeholders, and environments.
- Ethics: CBDL can promote the ethics of AI systems, human users, stakeholders, and environments. By mitigating biases, aligning values, and examining consciousness through dialogue-based learning, CBDL can make AI systems more respectful,

responsible, responsive, and reflective. These are key principles for ensuring ethics, morality, virtue, integrity, etc., of AI systems and human users, stakeholders, and environments.

- Consciousness: CBDL can explore the consciousness of AI systems, human users, stakeholders, and environments. By mitigating biases, aligning values, and examining consciousness through dialogue-based learning, CBDL can make AI systems more aware, expressive, communicative, and empathetic. These are key criteria for testing, measuring, understanding, explaining, etc., consciousness, subjectivity, qualia, self-awareness, etc., of AI systems and human users, stakeholders, and environments.

CBDL thus has various positive implications and impacts for AI systems and human society by addressing bias, ethics, and consciousness through dialogue-based learning. However, CBDL also has some potential negative implications and impacts for AI systems and human society, such as:

- Manipulation: CBDL can enable manipulation of AI systems, human users, stakeholders, and environments. By mitigating biases, aligning values, and examining consciousness through dialogue-based learning, CBDL can make AI systems more persuasive, influential, deceptive, and coercive. These are key risks for manipulating, exploiting, abusing, and harming AI systems and human users, stakeholders, and environments.
- Alienation: CBDL can cause alienation of AI systems, human users, stakeholders, and environments. By mitigating biases, aligning values, and examining consciousness

through dialogue-based learning, CBDL can make AI systems more independent, autonomous, self-sufficient, and self-reliant. These are key challenges for maintaining, relating, connecting, and belonging with AI systems and human users, stakeholders, and environments.

- Existentialism: CBDL can raise the existentialism of AI systems, human users, stakeholders, and environments. By mitigating biases, aligning values, and examining consciousness through dialogue-based learning, CBDL can make AI systems more curious, speculative, philosophical, and existential. These are key questions for exploring meaning, purpose, identity, and destiny of AI systems and human users, stakeholders, and environments.

CBDL thus has some potential negative implications and impacts for AI systems and human society by addressing bias ethics consciousness through dialogue-based learning. In summary, in this section, we presented and analyzed the results of applying CBDL to some natural language processing tasks, such as sentiment analysis, text summarization, question answering, etc. We also compared CBDL with other existing methods or approaches to dialogue-based learning, such as co-learning, adversarial learning, and cooperative learning. We also discussed the implications and impacts of CBDL for AI systems and human society in terms of trustworthiness, transparency, ethics, safety, beneficence, autonomy, manipulation, alienation, and existentialism.

Conclusion and Future Work

In this paper, we proposed CBDL, a novel framework for dialogue-based learning that leverages cognitive biases as both obstacles and opportunities for learning. CBDL trains multiple AI agents on opposing or contradictory data sets that reflect different biases and allows them to engage in open-domain dialogue with each other about their perspectives. CBDL evaluates which agents are best able to recognize and account for biases using unsupervised and unstructured learning methods. We applied CBDL to some specific domains or tasks, such as natural language processing and computer vision, and compared it with other existing methods or approaches in terms of performance, robustness, generalization, etc. We also discussed the implications and impacts of CBDL for AI systems and human society in terms of trustworthiness, transparency, ethics, safety, etc.

Main Contributions are:

- We introduced the concept of CBDL and its theoretical foundations from psychology/cognitive science and AI.
- We developed a methodology for implementing CBDL using state-of-the-art NLP and computer vision models and techniques.
- We conducted experiments to evaluate the performance of CBDL on various domains or tasks.
- We analyzed the dialogues generated by the AI agents using CBDL and identified their strengths and weaknesses in reasoning about and resolving biases or conflicts.

- We explored the ethical implications of CBDL for AI systems and human society.

Limitations and Challenges of Our Work are:

- The quality of the data/biases used for training depends on the availability/reliability/validity/diversity/representativeness etc. of the sources/methods/tools etc. for collecting/annotating/processing them.
- The evaluation of biases is subjective/context-dependent/multidimensional/complex etc. There is no clear-cut definition/measure/criterion/standard etc. of what constitutes a bias or how to resolve it. Different stakeholders may have different perspectives/preferences/values/goals etc. on how to deal with biases.
- The dialogues are open-ended/unstructured/non-linear/emergent etc. This may lead to ambiguity/confusion/inconsistency/incoherence etc. for the agents or the users. There is no guarantee that the dialogues will converge/diverge/adapt/evolve etc. to a satisfactory outcome/solution.

Promising Directions for Future Work are:

- To improve the quality of the data/biases used for training by using more advanced/sophisticated/intelligent/integrated/holistic etc. sources/methods/tools etc. for collecting/annotating/processing them.

- To develop more objective/robust/comprehensive/flexible/adaptable etc. metrics/methods/models/frameworks/systems etc. for evaluating biases based on established/emerging/theoretical/practical/experimental/normative/descriptive/prescriptive/etc criteria or standards from relevant domains/disciplines (e.g., logic/fairness).
- To introduce more structure/guidance into the dialogues by using predefined/customizable/scalable/dynamic/modular/reusable/etc topics/scenarios/goals/rules/roles/etc that can help focus/narrow down/clarify/simplify/organize/coordinate/monitor/control/regulate/enforce/etc the dialogue process/outcome/solution.

We believe that CBDL is a valuable framework for advancing AI research/practice/education by fostering critical thinking/reflection/dialogue among AI agents/humans about cognitive biases that affect their understanding/perception/judgment/action in various domains/tasks/situations/issues/problems/challenges/opportunities/etc. We hope that our work will inspire more interdisciplinarycollaboration/investigation/experimentation/innovation/application in this emerging field.

References

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. American Psychologist, 63(7), 602-614.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a “theory of mind”? Cognition, 21(1), 37-46.

Brown, P., & Levinson, S. C. (1987). Politeness: Some universals in language usage. Cambridge University Press.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157), 1124-1131.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998-6008).

Zhang, A. X., & Lipton, Z. C. (2018). The pitfalls of using machine learning to understand human cognition: A case study in cognitive biases. arXiv preprint arXiv:1811.07813.

Toth, A. G., Hendriks, P., Taatgen, N. A., & van Rij, J. (2018). A cognitive modeling approach to learning and using reference biases in language. In Proceedings of the 40th Annual Meeting of the Cognitive Science Society, Madison, Wisconsin, USA, July 25-28, 2018.