Socially-Aware Machine Learning

Prashanth Vijayaraghavan

Dissertation proposal for the degree of Doctor of Philosophy
in the Media Arts and Sciences
at the Massachusetts Institute of Technology

pralav@media.mit.edu

Dissertation Committee:

Deb Roy
Associate Professor, Laboratory of Social Machines
Massachusetts Institute of Technology

David Bamman
Assistant Professor, School of Information
University of California, Berkeley

Douwe Kiela
Research Scientist, Facebook AI Research

Chandra Bhagavatula
Research Scientist, Allen Institute for Artificial Intelligence

Abstract

Social intelligence refers to our ability to understand and infer the behaviour of other people in terms of their metal states, including intentions, desires, emotions and beliefs [1]. Research in cognitive studies suggests that Theory of Mind (ToM), at least in part, plays an important role in explaining social intelligence [2, 3]. This capacity to forecast and reason about others' mental states helps us negotiate our interactions and regulate our relationships in different social contexts. ToM is central to several theories of social inference and is strongly associated with increased social competence. This includes understanding social interactions such as coordinating and building stronger relationships with peers [4, 5, 6]. In addition, it also contributes to better narrative processing skills both in young children [7] and throughout life, whereas deficits in social intelligence are associated with increased risks of interpreting peer victimization and higher levels of peer rejection [8, 9, 10, 11]. While humans acquire such capabilities at a very early age [12, 13], machines still lack these qualities, often requiring extensive training data.

The purpose of this dissertation is to advance the efforts towards empowering machines with social intelligence using an approach referred to hereafter as Socially-aware machine learning (SAML). SAML will consist of two aspects. The first will address the challenges in making systems go beyond functional intelligence by equipping them with techniques to interpret and reason about human social dynamics under specific social contexts. The other aspect will focus on equipping humans to interpret the decisions of these systems that can have meaningful implications on questions of AI safety and social responsibility. Finally, we will demonstrate the transferability and social competence of systems with SAML using different social intelligence tasks that determine their ability to effectively comprehend and navigate complex social interactions.

Although building a completely sophisticated and flexible socially intelligent system may be a grand challenge, we will draw inspiration from human ToM and cognitive psychology to develop techniques that further the research in social reasoning by modeling and representing mental states of others [14]. Past research [15, 16, 17] has studied ways of representing mental states by mapping the transitions from traits and demonstrate how models trained to predict personality traits can be used to predict others momentary mental states. Other works [18, 19, 20, 21, 22] have used commonsense knowledge and reasoning to infer people's mental states based on their social situations. However, many of these models struggle to achieve good performance in domains such as the commonsense understanding of social interactions, or of other aspects of commonsense psychology [23, 24, 25, 26]. In addition, these models need to explain their decisions in a manner in which people can calibrate their trust on them.

In this dissertation, we propose to conduct SAML research that will (a) highlight deep learning approaches that develop rich representations of personality traits and mental states of people grounded in intuitive theories of human psychology and commonsense reasoning (b) produce human understandable explanations to allow for more transparency and (c) harness transferability and acquired commonsense knowledge to generalize learning to varied tasks and social situations. We will investigate whether a socially aware AI system can demonstrate improved (a) narrative processing skills, specifically, story comprehension through better inference of characters' mental states, (b) discerning skills through better detection and understanding of online hate, a form of peer victimization, and (c) pragmatic skills by generating emotion and intent-based responses in conversational systems. The success of these efforts will emphasize the value of SAML towards developing intelligent agents endowed with ToM sophistication which are more aligned with our human goals and preferences.

Contents

1 Introduction, Goals, and Research Questions 4

2 Towards Equipping AI with ToM 5

2.1 Background . . . 5

2.1.1 Modeling Human Social Behaviour . . . 6

2.1.2 Making Behavior Models Explainable & Ethical . . . 7

2.2 Research plan . . . 9

2.2.1 Using Personas to Explain Human Social Behaviour . . . 9

2.2.2 Inferring Mental States to Understand Human Social Behaviour . . . 11

2.2.3 Building Explainability into Models . . . 14

3 Towards Generalization across Social Competence Tasks 15

3.1 Narrative Processing Skills . . . 16

3.1.1 Background . . . 16

3.1.2 Research Plan: Story Comprehension . . . 17

3.2 Discerning Skills . . . 18

3.2.1 Background . . . 18

3.2.2 Research Plan: Understanding Online Hate Victimization . . . 19

3.3 Pragmatic Skills . . . 21

3.3.1 Background . . . 21

3.3.2 Research Plan: Empathetic Response Generation . . . 22

4 Contributions & Expected Results 24

5 Proposed Timeline 24

6 Resources Required 25

Committee Member Biographies

Deb Roy

Associate Professor, Laboratory of Social Machines, Massachusetts Institute of Technology

David Bamman

Assistant Professor, School of Information, University of California, Berkeley

David Bamman is an assistant professor in the School of Information at UC Berkeley (with an affiliated appointment in EECS). He works on applying natural language processing and machine learning to empirical questions in the humanities and social sciences. His research often involves adding linguistic structure (e.g., syntax, semantics, coreference) to statistical models of text. As such, He is especially interested in developing core NLP techniques for a variety of languages and domains (e.g., literary text, social media). Before Berkeley, He received his PhD at Carnegie Mellon (School of Computer Science, Language Technologies Institute) and was a senior researcher at the Perseus Project of Tufts University.

Douwe Kiela

Research Scientist, Facebook AI Research

Douwe Kiela is a research scientist at Facebook AI Research. He received his PhD (and his MPhil) from the University of Cambridge, where he was supervised by Stephen Clark. Before that, he did an undergraduate degree at Utrecht University with a double major in Cognitive Artificial Intelligence and Philosophy; and then a master's degree in Logic at the University of Amsterdam's Institute for Logic, Language & Computation. His work focuses on machine learning and natural language processing. His research interests lie in developing better models for language understanding and grounded language learning.

Chandra Bhagavatula

Research Scientist, Allen Institute for Artificial Intelligence

Chandra Bhagavatula is a research scientist at Allen Institute for Artificial Intelligence. He completed his PhD from Northwestern University in 2016, under the supervision of Prof. Doug Downey. His research interests are in the areas of Information Extraction, Natural Language Processing and he is currently exploring the application of Machine Learning algorithms to build Semantic Scholar - AI2's academic search engine.

1 Introduction, Goals, and Research Questions

Human intelligence is capable of understanding the nuances of social interactions on a daily basis without much conscious effort. In social settings, we use different social signals in the form of language, gestures and expressions to convey our beliefs, intentions and emotions. To make effective decisions, we model social situations by constantly hypothesizing and representing others' mental states from their social actions and evaluate how much we impute our own beliefs onto others [27, 28]. This aspect of human intelligence, known as social intelligence, plays a critical role not only in understanding implied but unstated meanings from narratives and social interactions but also in expressing ideas in speech or writing [29]. Neurologically, it is this cognitive capability of Theory of Mind (ToM) that helps us steer through different social interactions and as well reflect upon our own social behaviour. This is closely associated with perspective taking – the ability to imagine the world from multiple points of view [30] and empathy [31, 32]. Beyond interpreting the social world, this complex of abilities enables us to understand prosocial behavior such as to empathize, build peer relationships, form judgements, provide care, to name a few [33, 34, 35]. In fact, a deficit in ToM can impair comprehension of fictional narratives, social-pragmatic inference abilities involving violations of Gricean norms for social communication [36] and may increase the risk of discerning peer victimization in several ways [37]. It is, therefore, clear that development of ToM is closely linked to exhibiting improved narrative processing [38, 39, 40], pragmatic [41, **?**] and discerning skills [8, 37]. Such facets of social cognition rely on modeling others' mental states and perspectives, a topic that is still open to much more exploration in AI research and crucial to impart social intelligence to machines.

Although there have been significant advances in machine learning (ML) and deep learning (DL), developing a system capable of artificial general intelligence (AGI) with natural support for Theory of Mind remains a grand challenge. The goal of our research is to make significant progress in realizing this grand challenge of developing systems endowed with social intelligence. We use the term, Socially-aware machine learning (SAML), to refer to the ML and DL approaches that make systems socio-culturally adept and behaviourally intelligent. SAML will encompass: (a) modeling mentalizing – the ability to interpret the mental states that underlies human social behaviour and predict their social actions and (b) generating meaningful explanations about the decisions of the systems. In this dissertation, we look at the broader scope of what it means to have a socially-aware intelligent system by demonstrating marked improvement in narrative processing skills, discerning skills and pragmatic skills, via multiple tasks involving different kinds of data.

At the heart of our research, we draw upon existing studies on human Theory of mind and commonsense psychology where our social predictions are predicated on knowledge about other people, such as their mental states or personality traits which in turn determine the quality of social interactions. Just as we use our commonsense reasoning to intuit how traits predict mental states (e.g., highly extraverted individuals tend to feel self-assured), how mental states predict social actions (e.g., grateful people tend to cooperate) [42, 43], and how personality traits predict actions (e.g., agreeable people tend to cooperate) [44, 45], intelligent systems can benefit from commonsense understanding of people's social behaviour and the underlying influence of socio-cultural beliefs. We will explore ways of learning personality traits and integrating commonsense knowledge from related literature and develop novel techniques towards better understanding of people's behaviour and social relationships. There has been a considerable amount of attention towards building intelligent agents that learn the personas of people to generate personalized human-like responses [46, 47], understand users' online hate speech engagement [48], customize recommendations for users [49, 50], to list a few. Recent advances in incorporating commonsense knowledge to intelligent systems include studies that apply computer vision approaches to images, videos and cartoons. Similarly, commonsense knowledge is inferred from written texts such as news, stories and encyclopedias like Wikipedia [51]. However, in domains that require commonsense understanding of social interactions, or of other aspects of commonsense psychology, there is either limited work done or the models developed struggle hard to reason about others' behaviour under different social situations as shown in previous works [26, 52]. In addition to modeling the mental states of others, De Graaf et al. [53] note that people can calibrate their trust on intelligent systems only when they are capable of explaining their decisions in a comprehensible manner. A majority of the work has focused on post-hoc interpretations of deep learning models using textual, visual, local and example-based explanations. [54, 55]. Hence, our SAML research will seek to integrate commonsense knowledge of social situations with research in ML and DL to better understand and model the socio-cultural underpinnings of human behavior. Moreover, we will aim to develop intelligent systems that allow for improved transparency. This will help us evaluate their own behaviors for consistency with ethical norms about fairness. Further, we formulate social competence tasks for systems endowed with SAML based on studies that attribute the variability in children's performance in narrative processing, detecting peer victimization and pro-social communication to development of Theory of mind.

In this dissertation, we propose to broadly address the following research questions: (a) How can we develop intelligent systems that are socially-aware i.e. capable of modeling and explaining human behavior grounded in theories of human ToM and commonsense psychology? (b) Can such systems with rich transferable representations of human social behavior exhibit better performance in social competence tasks that require narrative processing skills, discerning skills and pragmatic skills? In the following sections, we will outline SAML research, comprising background literature on main research topics, followed by the research plan, proposed timeline, and required resources.

2 Towards Equipping AI with ToM

2.1 Background

The ability to anticipate, represent and reason about what other will think, feel or do in different situations is central to social cognition. Consider a scenario where one experiences difficulty in predicting social signals or implications, like agreeable people tend to be courteous and warm or exhausted people tend to show anger, it can lead to a complicated social life filled with misconceptions, faux pas and miscommunication. Fortunately, humans can predict others' probable social actions based on either their personality traits (eg. agreeableness) or mental states (eg. tiredness). Neuroimaging studies have also suggested the mentalizing, or "theory of mind" network plays a role in social cognitive processing more broadly, including reflecting on personality characteristics of one's self and others, inferring mental states including emotion processing, and intentions from actions.

A model proposed by Shamay-Tsoory et al. [56] divide ToM in two separate systems, namely cognitive ToM and affective ToM. Cognitive ToM is described as involved in processing inferences about others beliefs and intentions, whereas affective ToM is involved in processing inferences about other peoples emotions and feelings. This model describes affective and cognitive ToM involving common and different brain areas studied by Poletti, Enrici, and Adenzato. Several studies in the field of personality psychology [57] have some congruence with the above idea of cognitive behavior models. However, in these studies, personality is defined as “A dynamic and organized set of characteristics possessed by a person that uniquely influences their cognition, motivations, and behaviors in various situations” [57]. The dimensional theories like ”Big Five” model or theories that propose additional six personality dimensions are known to be tailored to understand stereotypes, mind perception and common behaviours. A work by Tamir et al. [58] studied if these low dimensional personality dimensional theories can efficiently aid social predictions. It was found that much of the richness of others minds can indeed be compressed to coordinates in a low-dimensional trait space. Similarly, there is considerable amount of literature [15, 59] that support how representation of people’s momentary mental states into lower dimension can facilitate social prediction in humans. However, it is still a huge challenge for machines to effectively navigate through complex social situations due to the lack of generalizable methods of representing people’s mental states or personalities and draw insights into people’s social behaviour.

2.1.1 Modeling Human Social Behaviour

With increasing human-machine hybrid technologies, the real-world interactions with AI systems are often stilted. It is important to acknowledge the challenges associated with understanding of explicitly unstated desires, emotional states and intentions of users from language. Misinterpretation of users’ implied intents and implicit beliefs from natural language could have dramatic real world consequences. Building AI systems that can interact with humans fluently will require machines to share common knowledge about how people will act, communicate and react under specific contexts and circumstances.

A number of AI researchers have attempted to adopt these ideas and build systems that can encode personality traits or mental states into representations and utilize them in different social contexts. Bridewell and Isaac (2011) [60] introduced a computational framework for common, complex, and under-investigated aspect of human social behavior like deception based on the capacity to reason about the goals of other agents, resting on mental state ascription. Fahlman [61] proposed a knowledge-base system, Scone, used to emulate some aspects of human mental behavior and support human-like common-sense reasoning and language understanding. Beyond domain specific knowledge, social understanding requires generic knowledge about social interactions and their ensuing effects on mental states. An early research conducted by Wilensky along these lines inferred the intentions of interacting agents while Dyer dealt with extracting morals from social scenarios. Winston’s [62] Genesis system was developed to understand and generate stories using computational models that use commonsense inference rules and concept patterns. This includes their work to support question answering, personality or mood-based interpretation and summarization of stories.

One possible direction, explored to overcome shortcomings of AI systems in navigating the social world, is to endow them with commonsense knowledge. While there have been significant efforts to create knowledge bases such as Cyc [63] and ConceptNet [18], there is a paucity of inferential knowledge related to people’s behaviour in the form of their motivations and their reactions. Using stories to define a space of acceptable behaviours, Harrison et al. [64] developed a technique to prevent autonomous agents from exhibiting anti-social or psychotic behaviours. Recently, knowledge base such as Event2Mind [22] and ATOMIC [21] are tailored to capture mental states of people linked to day-to-day events. Another line of work towards improving automatic recognition and interpretation of human social signals in AI systems relies on inferring personality traits. Considering that personality compels a tendency on a lot of aspects of human behavior, mental states and affective reactions, there is an enormous opportunity for sensing spontaneous natural user behavior to facilitate efficient interaction in social settings. [65] presented a hypothesis that users by similar personality are expected to display mutual behavioral patterns when cooperating through social networks. Imitating personal style in dialogue systems have demonstrated promising results. Some of the early efforts include modeling personas of movie characters and incorporating speaker persona in dialogue models based on speaking style characterized by natural language sentences [66, 47]. Despite recent successes, it is still incredibly challenging to build socially intelligent agents that can understand humans and engage in socially competent conversations that involve empathy, cooperation, persuasion, care-giving, to list a few.

Recent approaches have focused on reflecting upon the concepts of human ToM to attribute mental states such as intentions and beliefs to inanimate objects. Some notable approaches include those that use hierarchical Bayesian inference [67, 68, 69] or artificial neural networks [70, 71]. The former is generally cognitively-inspired and suggests the existence of a "psychology engine" in cognitive agents to process ToM computations while the later achieves imparting ToM to certain degree by characterizing different species of deep reinforcement learning agents. In addition to these methods, there also have been multi-agent models rooted in statistical machine learning theory and robotics. These approaches have generally evaluated on simulations of Theory of mind in relatively simple situations. However, there is very limited work in this area of combining theory of mind and language. It is also well known that two of the most fundamental elements of human cognitive capabilities are the ability to communicate through complex language systems and attainment of a theory of mind. Interestingly, both language and theory of mind develop relatively at the same time in a persons life. Language is a fundamental element in understanding emotions, thoughts and actions that is constitutive of both experiences and perceptions. Early ToM abilities facilitate the development of early Language abilities while more complex language abilities are precursor to complex ToM abilities. Language, ToM and social skills are all connected and interdependent. Hence, we focus on bridging this gap in research towards understanding language and development of ToM abilities.

In this dissertation, the first aspect of our SAML research will focus on learning to model personality traits and mental states of people by integrating commonsense knowledge of social behaviour with knowledge acquired from textual corpus such as stories. The representations learned by such models are more likely to yield AI systems that are generally safer and better at perceiving, understanding, and responding effectively to different social situations. With trait models and commonsense knowledge acting as the basis of Theory of Mind, the behavior of such socially-aware systems, specifically during human-machine social interactions, will be consequently more recognizable and aligned to people's expectations.

2.1.2 Making Behavior Models Explainable & Ethical

With AI systems getting smarter and complex, they are often incomprehensible by human intuition and are quite opaque. This can be a huge limitation when AI systems are entrusted to make complex decisions in situations that were previously handled by humans. Given that machines are nowhere close to having human-like intelligence, they are likely to fail, violate conditions or make wrong decisions. These condition raises several ethical questions around usage of such systems in sensitive domains. For example, Angwin et al. [72] at ProPublica found that the decisions of a criminal risk assessment software, COMPAS, were racially biased and unreliable. In such scenarios, it is essential for AI systems to explain their decisions. With the success of deep learning approaches, there are significant challenges in interpreting the deep layers of neural networks, raising serious concerns about trust, safety and transparency of these AI systems. It is to be noted that a number of attack methods [73, 74, 75, 76] have exposed the vulnerabilities of deep neural models to imperceptible perturbations of the input data. Despite the emergence of defense techniques, the potential of adversarial examples to fool these models and prompt unexpected behaviours has all the more emphasized the need for explainability of these black-box models. Thus, these make a strong case for understanding the ethical questions around problems we are dealing with and follow the guidelines necessary to reduce the harms of misuse that any AI system is likely to cause.

Though our goal is not to shift our focus completely on handling ethical questions around building socially intelligent AI systems, we attempt to reflect upon certain aspects of EU governance framework for algorithmic accountability and transparency [1] in our models. This includes analyzing any existing bias in the data we use in our models and addressing that. Additionally, we also aim to allow for meaningful transparency by providing explanations as to why a particular result was obtained. Towards this direction, we explore and draw ideas from various studies in this field of explainable AI.

The ubiquity of ML and DL algorithms in areas of scientific research, such as social sciences, medicine, biology, establishes the importance for explanations not only for calibrating trust on the research outcomes but also for overall progress of research. The problem of explainable AI systems has been studied by various scientific communities, with most of the literature primarily coming from the machine learning and data mining communities. However, every community approaches this problem through different lens without consensus on a core set of standards for what makes a good explanation of the AI systems' behaviour. Efforts to open up the black-boxes [54, 77] have focused on describing either the inner workings of the black-boxes or merely the decisions made by them .

Initial studies rely on readily interpretable rule-based shallow models such as decision lists, decision trees, classifiers based on association rules and Bayesian Rule Lists [78, 79, 80]. In deep learning architectures, attention mechanisms assign weights to feature representations that are known to improve task performance in text and image domains [81, 82] and as well appear semantically appropriate to humans. However, more analysis is required to assess and validate the claims of attention based interpretability [83, 84]. Several model-agnostic methods based on the relationship between features and predicted outcomes of the learned model have also been explored. The influence of features on the outcomes of the model are explained through PDP, ICE or ALE plots [85, 86, 87, 88]. Extending this area of investigation, more model-agnostic models such as global/local surrogates, shapely values were adopted [89, 90]. Surrogate models are those trained to approximate the predictions of the underlying black-box model. A majority of the work is on posthoc descriptions of how a system arrived at a decision. LIME [91] is one of the notable local surrogate models that works by training a weighted, interpretable model that

[1] https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf

traces the effects of variations to the input data on output performance. The interest in identifying parts of input data impacting the outcomes paved way for several approaches of deriving saliency maps [92]. They commonly use visualizations of representations, saliency maps or attention scores learned by neural networks as explanations [93, 82].

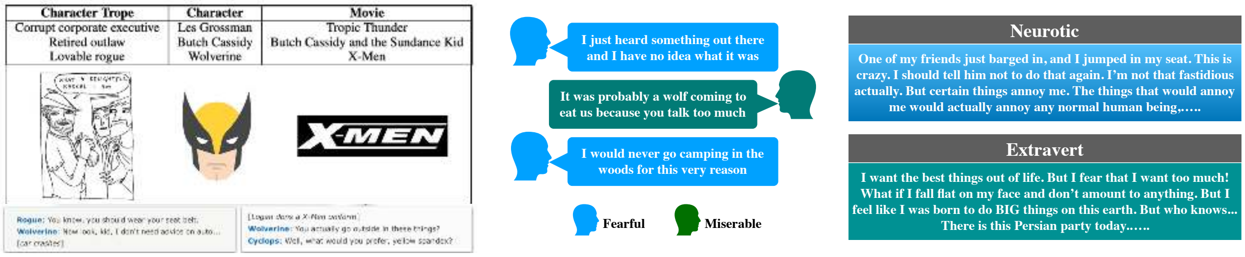
Few studies have proposed to explain and interpret deep neural networks by leveraging adversarial examples [94, 95, 96, 97] that can help detect weaknesses of the model in the form of problematic decision boundaries. These adversarial examples are used as non-linear explanation mechanisms produce, termed adversarial explanations [98]. Moreover, inspecting adversarial examples not only improves interpretability by shedding light on some of the semantic inner levels of the neural models but as well enhances model robustness. There has been some recent work on generation of natural language explanations. It is not always easy to generate a description of how neural models process the input data. Inspired from how humans provide rationales behind their actions, a neural decoder is trained to translate models internal states into meaningful natural language explanations collected from humans performing a similar task. While Chen et al. [99] generated natural language reasons behind image classification decisions similar to the way ornithologists or geologists would explain to people, a work by Ehsanet al. [100] enabled agents to produce human-like rationales reflecting the internal workings of the black-box systems and studied how these generated rationales promote feelings of confidence, human-likeness, and comfort in non-experts operating autonomous agents. Drawing ideas from the literature, our SAML research will explore different techniques for producing meaningful insights on models' decisions. While most of these previous studies were conducted in visual domain, we will adapt specific methods such as adversarial or natural language explanations [98, 100] to the textual domain in an effort to foster trust in SAML-based AI systems. By this, we not only provide aspects of the input data and model components that contributed to a particular decision but also expose some of the weaknesses of our model for responsible and ethical usage.

2.2 Research plan

This section will outline approaches that will drive us towards developing techniques to model human behaviour and and building aspects of transparency into the decision making of the models. As explained in Section **??**, we learn to model (a) personality traits of people based on the hypothesis that personality compels a tendency on various aspects of human thought, behavior, motivation and emotion [65] (b) mental states of people supported by social commonsense knowledge. We then discuss about combining adversarial examples based approaches with post-hoc techniques to explain our models' decision.

2.2.1 Using Personas to Explain Human Social Behaviour

Individual personality plays a deep and pervasive role in shaping social life. Personality governs how the character responds to experiences, situations and other people. Research indicates that it can relate to the professional and personal relationships we develop [101, 102], the technological interfaces we prefer [103], the behavior we exhibit on social media networks [104], and the political stances we take [105]. Prior studies in the field of psychology [106] have established the relationship between personality and natural languages. For example, a narcissistic person might make frequent use of first-person expressions (I, me, myself, for me, etc.). Motivated by such



Character Trope	Character	Movie
Corrupt corporate executive	Les Grossman	Tropic Thunder
Retired outlaw	Butch Cassidy	Butch Cassidy and the Sundance Kid
Lovable rogue	Wolverine	X-Men

Rogue: You know, you should wear your seat belt.
Wolverine: Now look, kid, I don't need advice on auto...
[car crashes]

[Logan dons a X-Men uniform]
Wolverine: You actually go outside in these things?
Cyclops: Well, what would you prefer, yellow spandex?

I just heard something out there and I have no idea what it was

It was probably a wolf coming to eat us because you talk too much

I would never go camping in the woods for this very reason

Neurotic

One of my friends just barged in, and I jumped in my seat. This is crazy. I should tell him not to do that again. I'm not that fastidious actually. But certain things annoy me. The things that would annoy me would actually annoy any normal human being,.....

Extravert

I want the best things out of life. But I fear that I want too much! What if I fall flat on my face and don't amount to anything. But I feel like I was born to do BIG things on this earth. But who knows... There is this Persian party today......

(a)

(b)

(c)

Figure 1: Sample data from (a) Movies Corpus, (b) Image-Chat Corpus and (c) Personal Stories Corpus

works, we focus on a language understanding task involving automatic recognition of personality traits from text. There has been considerable amount of interest in the past that used NLP tools to conduct traits analysis of fictional characters in literary texts [107, 108]. Knowing the personality information can effectively guide us towards developing psychologically plausible intelligent agents. In this work, we take a step towards deriving persona representations that explain human social behaviours and roles categorized based on their influences on language, conversations and actions in different social contexts. Towards this goal, we learn to model personas from annotated data in dialogue and storytelling domains.

Dataset We use following datasets to understand the relationship between personas and language and eventually learn persona representations from text.

1. **Dialogue Datasets:**

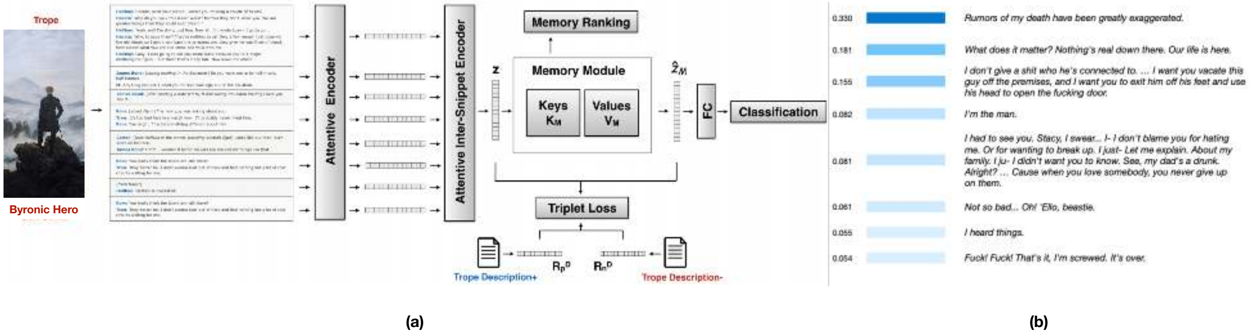
   (a) *Movies Corpus*: Using CMU movie summary corpus [66], we collect characters dialogue utterances from IMDB quotes page. As much as 70.3% of the quotes are multi-turn exchanges between multiple characters. Each of these characters are classified into one of the 72 trope categories.

   (b) *Image-Chat Corpus*: A large collection of dialogue utterances collected through image grounded human-human conversations [109] is used in our work. Each person involved in a conversation is associated with a personality type. Following the work of Chandu et al. [110], we utilize the 5 well-formed and meaningful personality clusters associated with utterances data in our work.

2. **Personal Stories Corpus**: Personal stories or reflections explain important parts of one's personality including their goals and values [111]. For a labelled dataset, we make use of personal essays, obtained originally from Pennebaker et al. [112]. This corpus consists of large stream of consciousness text collected between 1997 and 2004 and labelled with 5 personality categories [113]. The psychology students who wrote these texts were assessed based on Big Five questionnaires. Though this dataset might not be in complete alignment with other forms of data in storytelling domain, it is one of the few datasets containing gold labels suitable for our purpose.

Figure 6 displays few samples from the above datasets and the dataset stats are given in Table 1 (left). Due to the lack of large scale persona annotated personal stories dataset, we leverage on the models built using the dialogue dataset to produce persona-based representations and transfer the learnt knowledge on personal stories.

Modeling & Evaluation In order to learn rich latent persona embeddings, we begin by training a neural network architecture (See Figure 2b), referred to as attentive memory network, on dialogues dataset. This model operates not only on a character's own utterances but also the context and other interacting characters' utterances existing in those multi-turn exchanges. We implement other baseline models and incrementally add various components that intuitively increase model capacity and provide additional knowledge for this task. The best-performing models use a multi-level attention mechanism over a set of utterances. We also utilize prior knowledge in the form of textual descriptions of the different personas described as character tropes in the movie corpus dataset. The persona embeddings obtained from dialogues encapsulate the human behavioural information in their social situation. Figure 2b shows the attention scores on a batch of dialogues that represents the relevance of each dialogue towards classification of persona categories. Further details of the models, datasets and evaluations are explained in [114] [2].



(a)

0.330 Rumors of my death have been greatly exaggerated.

0.181 What does it matter? Nothing's real down there. Our life is here.

0.155 I don't give a shit who he's connected to. ... I want you vacate this guy off the premises, and I want you to exit him off his feet and use his head to open the fucking door.

0.082 I'm the man.

0.081 I had to see you. Stacy, I swear... I- I don't blame you for hating me. Or for wanting to break up. I just- Let me explain. About my family. I ju- I didn't want you to know. See, my dad's a drunk. Alright? ... Cause when you love somebody, you never give up on them.

0.061 Not so bad... Oh! 'Ello, beastie.

0.055 I heard things.

0.054 Fuck! Fuck! That's it, I'm screwed. It's over.

(b)

Figure 2: (a) Illustration of Attentive Memory Network (b) Attention scores for a batch of dialogues for "bryonic hero" persona category.

We extend our model to further fine-tune the learnt persona embeddings for Image-Chat corpus and Personal Stories corpus. We sharpen our representations using such datasets and evaluate on the test data associated with their corresponding corpus. Our best-performing model outperforms other baseline models on these datasets. We are able to perform simple narrative analysis such as identifying similar characters, clustering movies/personal stories using our latent persona embeddings. The purpose of such representations doesn't end with narrative analysis. The results of our preliminary evaluations suggest that these methods could be applied to different social competence tasks explained in Section 3.

2.2.2 Inferring Mental States to Understand Human Social Behaviour

Endowing machines with the ability to infer mental state representations plays an important role in developing socially intelligent systems. Stories are one of the most common yet powerful means

[2] https://arxiv.org/pdf/1810.08717.pdf

Dataset	Total Size	Dataset	Total Size
Movie Corpus	$\sim 16k$	ROCStories	$\sim 100k$
Image-Chat Corpus	$\sim 35k$	Personal SM Stories	$\sim 450k$
Personal Stories Corpus	$\sim 2.5k$	Simple Strange Stories	~ 600

Table 1: Statistics of datasets used for modeling personas (left) and mental states (right)

of communication used to enhance engagement with complex issues and understanding of the social world. People share and consume stories in a variety of ways to convey and make sense of their experiences. Understanding a story not only requires keeping track of sequence of events happening in the story but also inferring and interpreting the mental states of characters and interactions between them. It is thus natural to consider the usage of stories towards building a model for inferring aspects of people's mental states based on their actions in social situations.

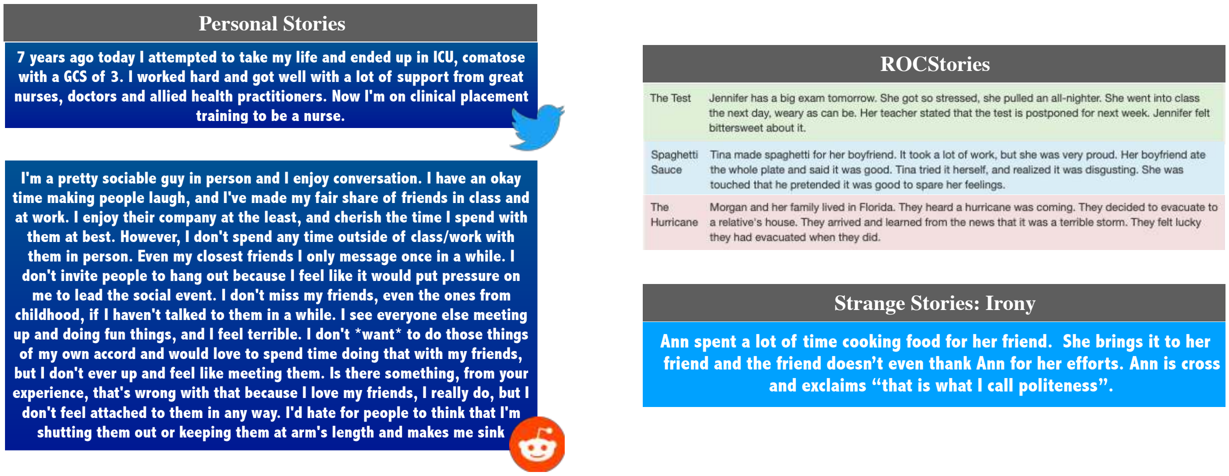
In our work, we propose the task of learning a representation of others' mental states, specifically their goals and emotional reactions. This can be really important for a wide variety of tasks that can benefit from perspective-taking associated with events. We leverage the social commonsense knowledge bases ATOMIC [21] containing stereotypical intents and reactions of people on day-to-day events. Recent research in psychology suggests that readers of personal narratives and fiction score higher on measures of empathy and theory of mind (ToM)-the ability to think about others' thoughts and feelings than non-readers, even after controlling for age, gender, intelligence and personality factors [115]. Therefore, it is clear how stories can act as a source of modeling our mental state inference model. Schank and Abelson [116] argued that one's own collection of personal stories was the knowledge base itself. Our efforts are also directed towards utilizing personal stories from different social media platforms for enhancing our social commonsense reasoning model.

Dataset In addition to publicly available stories dataset, we also collect different forms of stories to train and evaluate our models. Sample data from our dataset is given in Figure 3. Table 1 (right) shows the stats of our dataset. A brief description of the datasets are as follows:

1. **ROCStories:** This high quality collection of non-fictional daily short life stories captures a rich set of commonsense causal and temporal relations between daily life events [117]. It consists of five-sentence commonsense stories containing sufficient context and avoids unnecessary tracks of irrelevant information in the story targeted towards deeper story understanding.

2. **Personal SM Stories:** We collect personal stories from social media (SM) platforms – Twitter & Reddit. By curating hashtags for Twitter and subreddits for Reddit, we obtain a large corpus of personal stories containing different categories and post metadata.

3. **Simple Strange Stories:** Inspired from ToM Tests, such as Happé's Strange Stories test [118], we collect mentalistic stories consisting of short vignettes. Each story falls under one of the five story types (eg. Satire, Irony, etc.) referring to a subset of categories enlisted in the strange story test [119]. The mentalistic stories are also annotated with the character's intent or motive. This dataset is limited in size primarily intended to demonstrate our model's ToM capabilities.



Personal Stories

7 years ago today I attempted to take my life and ended up in ICU, comatose with a GCS of 3. I worked hard and got well with a lot of support from great nurses, doctors and allied health practitioners. Now I'm on clinical placement training to be a nurse.

I'm a pretty sociable guy in person and I enjoy conversation. I have an okay time making people laugh, and I've made my fair share of friends in class and at work. I enjoy their company at the least, and cherish the time I spend with them at best. However, I don't spend any time outside of class/work with them in person. Even my closest friends I only message once in a while. I don't invite people to hang out because I feel like it would put pressure on me to lead the social event. I don't miss my friends, even the ones from childhood, if I haven't talked to them in a while. I see everyone else meeting up and doing fun things, and I feel terrible. I don't *want* to do those things of my own accord and would love to spend time doing that with my friends, but I don't ever up and feel like meeting them. Is there something, from your experience, that's wrong with that because I love my friends, I really do, but I don't feel attached to them in any way. I'd hate for people to think that I'm shutting them out or keeping them at arm's length and makes me sink

ROCStories

The Test	Jennifer has a big exam tomorrow. She got so stressed, she pulled an all-nighter. She went into class the next day, weary as can be. Her teacher stated that the test is postponed for next week. Jennifer felt bittersweet about it.
Spaghetti Sauce	Tina made spaghetti for her boyfriend. It took a lot of work, but she was very proud. Her boyfriend ate the whole plate and said it was good. Tina tried it herself, and realized it was disgusting. She was touched that he pretended it was good to spare her feelings.
The Hurricane	Morgan and her family lived in Florida. They heard a hurricane was coming. They decided to evacuate to a relative's house. They arrived and learned from the news that it was a terrible storm. They felt lucky they had evacuated when they did.

Strange Stories: Irony

Ann spent a lot of time cooking food for her friend. She brings it to her friend and the friend doesn't even thank Ann for her efforts. Ann is cross and exclaims "that is what I call politeness".

Figure 3: Sample data from personal Stories, ROCStories and Simple Strange stories Corpus

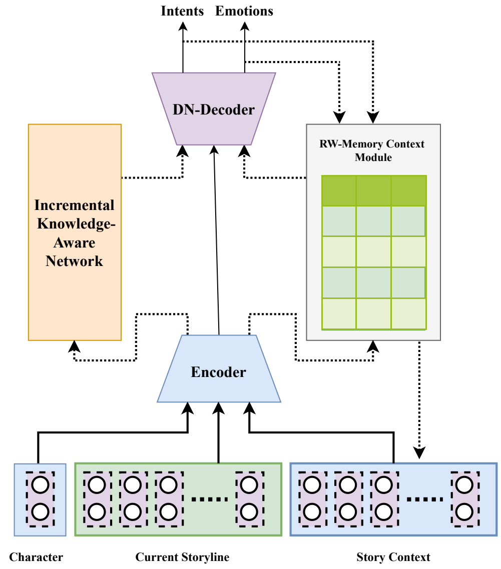
Modeling & Evaluation Instead of predicting characters' final mental state for the entire narrative, our model will focus on inferring interpretable trajectories of characters' mental states. involves the following challenges:

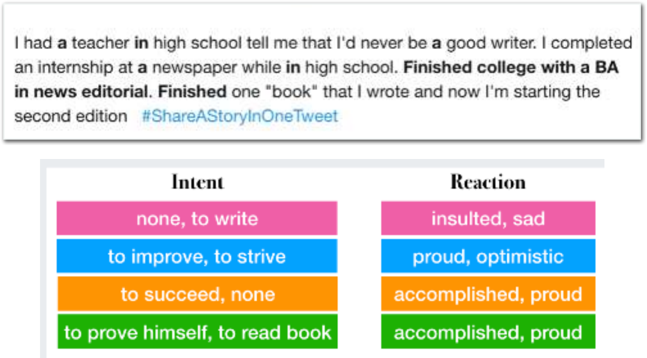
- The growing need to equip models with social commonsense reasoning capabilities in order to analyze and construe aspects that are not always explicitly mentioned. This includes implicit details about the person involved in the social situation that can be trivially inferred.
- The ability to continually learn and update the commonsense knowledge as new facts are discovered in various domains without forgetting the accumulated past knowledge.
- The importance of accounting for interdependencies between current story context based on processed story sentences and their corresponding mental state representations to infer the new mental states.

Drawing ideas from the literature in text understanding, commonsense reasoning and lifelong learning learning, we implement a deep sequence generation model that addresses the challenges above. The encoder-decoder architecture is augmented with the following components:

- *Incremental Knowledge-Aware Network* that is able to infuse social commonsense knowledge to our model and as well continually learn new domain knowledge without catastrophic forgetting.
- *DN-Decoder* that involves a two-step decoding process based on deliberation networks [120]. This is motivated by human cognition process, where people make a first draft based on existing information and then polish the response with the help of background knowledge.
- *Read-Write Context Memory Module* that handles the interdependencies between prior and current story and mental state context information [121].

We experiment with different deep neural models as encoder and decoder and conduct ablation studies that ascertain the importance of various components of the model. We find that our model is capable of learning coherent mental representations from characters' based on their actions and their affect states explained in the story. Figure 4 shows the illustration of our model and sample mental state trajectories from social media stories. We conduct experiments on the incremental knowledge-aware network so that it can act as a suitable knob for future updates to the model. We evaluate our work on held-out test set in our story collections datasets explained above. Additionally, we also measure our models' performance on a publicly available Character psychology dataset released by Rashkin et al. [122]. It consists of $300k$ low-level annotations for character motivations and emotional reactions. Our best-performing model outperforms the baseline models in different settings (unsupervised, semi-supervised or supervised) suitable for the dataset under consideration.




I had a **teacher** in high school tell me that I'd never be a good writer. I completed an internship at a newspaper while **in** high school. **Finished college with a BA in news editorial. Finished** one "book" that I wrote and now I'm starting the second edition #ShareAStoryInOneTweet

Intent	Reaction
none, to write	insulted, sad
to improve, to strive	proud, optimistic
to succeed, none	accomplished, proud
to prove himself, to read book	accomplished, proud

Figure 4: (left) Illustration of our model, (right) Sample mental state trajectories on a short personal story

2.2.3 Building Explainability into Models

There are several methods to approach explainability in deep learning models. Given that there is an extensive literature in this field tied closely to visual domain, we will adapt specific techniques to textual domain and improve the explanations based on task-specific requirements. Having discussed many of these approaches in Section 2.1.2, we are specifically interested in exploring the possibility of using adversarial examples to not just identify the vulnerabilities of the model but also to better understand the inner workings of the model.

Although interpretability and adversarial examples are not often considered together, few papers [123] have identified the conceptual association between them and explained how research on one of them can provide insights on the other. One of the common methods in cognitive neuroscience to understand how the brain works involves identification of examples that are usually mistaken by humans. In an effort to replicate such methods in ML research, Ritter et al. [124] chose an analysis that explains how children learn word labels for objects and applied it to DNNs.

They found that DNNs demonstrated a bias to categorizing objects by shape rather than by colour, similar to what was observed in humans. This work and provides a case for using the study of adversarial examples in human behaviour to broaden how we study adversarial examples and interpretability in DNNs. Adversarial examples expose DNNs' deficiencies and highlight the challenges related to learning input mappings and class boundaries that align with our intentions as model builders. Adversarial examples, therefore, offer an interesting opportunity for us to better perceive the model's decision spaces and feature representations. Enhancing the explainability of the model through adversarial examples will allow us to find means to identify sensitivities of the model, make the models more robust to minor perturbations and as well develop-cum-align our models with human expectations of the models' functionalities.

Towards this objective of introducing explainability into models using adversarial examples, we first develop an Adversarial Examples Generator (AEG), a model capable of generating adversarial text examples to fool the black-box text classification models [125]. We use black-box attacks as they do not utilize model parameters or gradient information. This is specifically advantageous as they can be applied to different kinds of target models based on querying the model without any dependency on the architecture or parameters of the model. By adopting a self-critical sequence training approach for hybrid character-word encoder and decoder, adversarial example with word and character based perturbations are generated. Based on evaluation on IMDB reviews and AG's news dataset, we find that our model is capable of generating semantics-preserving perturbations that leads to steep decrease in accuracy of those target models. It is important to note that our model not only exposes the weaknesses of the target model irrespective of the nature of the model (either word or character level) but it also identifies the most important words that contribute to particular categorization of the target model. This can offer certain degree of interpretability. Therefore, such indicators can be leveraged to explain the various facets of the target model. Depending on the nature of the tasks, we will seek to provide different kinds of explanations using either adversarial examples or existing post-hoc techniques to instill trust in AI systems using SAML.

3 Towards Generalization across Social Competence Tasks

In the light of our work, it is crucial to demonstrate the social competence abilities of the models that endow knowledge about human social behaviour to machines. In this section, we will review the literature on the relations between ToM and social competence and further investigate the background work on a set of social or academic competence tasks to establish the value of our models, in the context of building socially intelligent systems.

Social competence is defined as, "forming and maintaining positive social interactions with others while reaching personal goals" [126]. Several studies [127, 128] have indicated the general decrease in children's aggressive behaviours and an increase in pro-social acts throughout their preschool years. This is usually attributed to improvements in several domains such as social cognition, language and self-regulation [129]. Despite some inconsistencies in the literature, many findings highlight the association between ToM and social competence. Construed broadly, ToM covers a range of capabilities such as perspective taking, simulating mental states, identifying character traits, social and emotional reasoning. Linked closely to acquisition of social vocabulary and processing of social information, ToM is critical for facilitating active engagement in social and academic activities.

Kimhi [130] discussed ToM development across the life span in persons focusing on the social and academic manifestations of ToM that are critical for everyday life skills. Considering the social manifestations of ToM in symbolic play, conversation, and autobiographical memory and academic manifestations of ToM in reading comprehensions, narrative skills, and writing abilities, many related works including the literature with mixed evidence on the significance [131] and direction of association [132, 133] were discussed. However, there is consensus that children's ability to understand others mental states, though may not be sufficient by itself, appears to be necessary for engaging in adaptive and positive behavior [134] and these abilities are reflected in their social interactions [135]. This is further supported by results from assessment of individuals with autism. Even their high verbal and intellectual levels do not aid in navigating effectively in social and academic settings exacerbated by the diminished attention to social cues and difficulty in social adaptive behavior [136, 137]. Consequently, interventions have been proposed and developed to enhance ToM in children and young people with autism spectrum conditions. There are different categories of ToM interventions such as groupsspecific ToM socio-cognitive training that focuses on improving specific ToM skills, and more general social skills interventions that incorporate ToM training among other social skills. Specific ToM socio-cognitive training [138, 139] (eg. Thought Bubble Training) has been found to enhance the targeted skills; yet, generalization to other skills generalization to the natural environment has been minimal for the most part. More sophisticated interventions (eg. dyadic & group social interventions) involve training strategies [140, 141, 142] that integrate social interaction training in child's natural settings with the main social interactive agents (teachers, peers and parents) involved along with specific sociocognitive abilities. Such studies aim at reinforcing more holistic social functioning instead of teaching only particular ToM skills and hence are not limited to conditions under which specific ToM skill training is provided.

Taking a leaf out of the various experiments with children and young adults with autism spectrum conditions, we are interested in developing sophisticated learning techniques directed at enhancing the social skills generalized to real-life settings. Therefore, we incrementally augment social commonsense knowledge in the form of intents, emotions and personality characteristics based on understanding of the social world and fine-tune to specific social competence skills not under controlled experimental conditions but by exposing them to innate social interactions of uniquely distinct social contexts. For our dissertation, we use past augmented knowledge and also adapt quickly to the different social contexts by harnessing the transferability and ideas from the practice of continual lifelong learning. We apply the models and representations obtained using SAML in three different social competence tasks, each focused on improving different social or academic manifestations of ToM. The tasks include (a) story comprehension from the perspective of characters' mental states (b) understanding and detection of online abuse and hate, a form of peer victimization and (c) emotion and intent-based pro-social response generation in conversational systems. These tasks demonstrate various sociocognitive skills such as the narrative processing, discerning and pragmatic skills respectively.

3.1 Narrative Processing Skills

3.1.1 Background

Narratives, real or fictional, contain nuanced information about the social world with rich instances of social interactions where the reader must impute the intentions and emotions of characters.

Since these social interactions do not explicitly describe the mental states of characters, it puts the onus on the reader to actively construct mental models of people based on the evolving social situations in the story. Further, the capacity to reason about the mental states of another person is crucial for alternating between the perspectives of different interacting characters and eventually understanding the complete story [143]. This capability necessarily draws on ToM [144, 145].

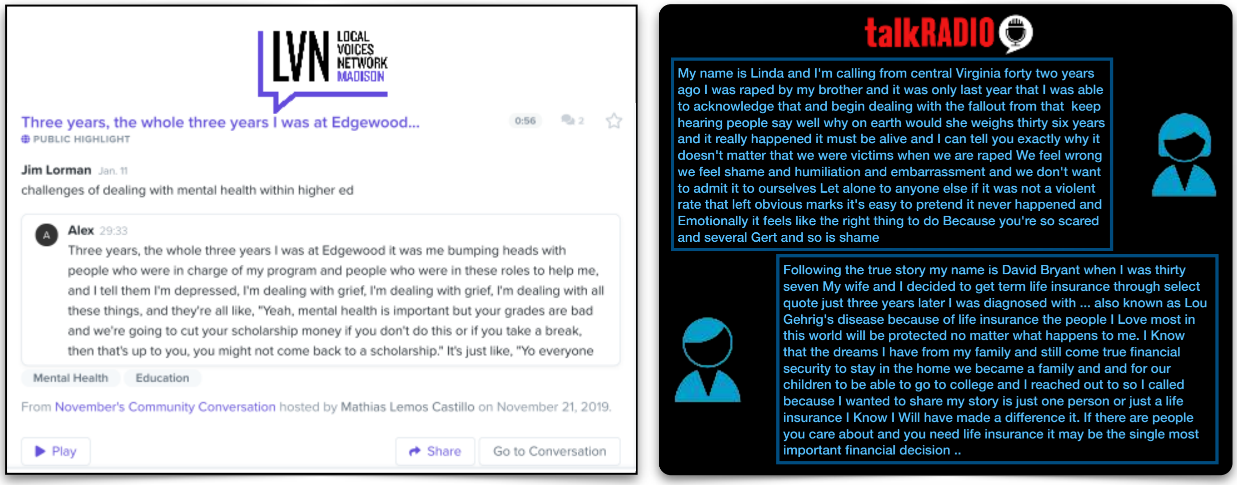
A growing body of work has developed concerning the relations between ToM and narratives [38, 39, 40]. Previous researches on ToM and its link with narratives have clearly established their interdependence; A study by Garcia et al. [146] found that children with autism spectrum conditions had difficulty adjusting to perspectives of characters within the narratives. Thus, narratives can be used as an instrument in the evaluation of children's ToM skills [147, 148]. At the same time, Guajardo and Watson [149] established how reading storybooks combined with discussion about characters' mental states improved young children's ToM performance. It is, therefore, clear that narrative experiences may better the ability of individuals with autism to "mentalize" [150] and as well act as a means to measure the ToM skills. This interdependence is linked to the acquisition of an articulated mental language, comprising of elaborate social and emotional vocabulary. As a result, story narrative comprehension can be a more challenging task for machines requiring them to develop commonsense understanding of human social behaviour and perspective-taking of various interacting characters in the story [151, 152, 153, 154].

In the past, different aspects of story understanding have been studied. This includes modeling inter-personal relationships, narrative and open domain question answering, understanding narrative structures, learning scripts, event schemas and narrative chains, story plot generation and creative or artistic story telling, to list a few [155, 156, 157, 158]. In addition, there have been recent works focusing on predicting what happens next or a choosing a coherent story ending from options [159]. Of particular interest in this dissertation is the collection of dataset of stories from different domains. We intend to understand, learn and explain the chains of characters' social behaviour in the story. This offers immense opportunity to analyze, comprehend and study stories and personal narratives of people from these domains using our models that embed human social behaviour through personality or mental states. Thus, we extend our work to accommodate these different domains of data and advance the research towards building better mentalizing models.

3.1.2 Research Plan: Story Comprehension

Given the relationship between ToM and narratives, we transfer the knowledge obtained from training our models explained in Sections 2.2.2 and 2.2.1 to other forms of stories and personal narrative dataset. Experimental studies in psychology have discovered multiple factors that play a role in evoking empathy. Specifically, the language of a storyteller can influence what emotions individual readers feel. With the growing interest in real life stories and experiences of people, we deem it important to identify such personal stories and analyze patterns that emerge from them. Real life stories can be gathered from social media and other publicly available platforms where people share their personal narratives. This includes the following:

- RadioTalk [160], a corpus of speech recognition transcripts sampled from talk radio broadcasts in the United States between October of 2018 and March of 2019. The corpus encompasses approximately 2.8 billion words of automatically transcribed speech from 284,000 hours of radio, together with metadata about the speech, such as geographical location, speaker turn boundaries, gender, and radio program information.



Three years, the whole three years I was at Edgewood...

Jim Lorman Jan. 11

challenges of dealing with mental health within higher ed

Alex 29:33

Three years, the whole three years I was at Edgewood it was me bumping heads with people who were in charge of my program and people who were in these roles to help me, and I tell them I'm depressed, I'm dealing with grief, I'm dealing with grief, I'm dealing with all these things, and they're all like, "Yeah, mental health is important but your grades are bad and we're going to cut your scholarship money if you don't do this or if you take a break, then that's up to you, you might not come back to a scholarship." It's just like, "Yo everyone

From November's Community Conversation hosted by Mathias Lemos Castillo on November 21, 2019.

My name is Linda and I'm calling from central Virginia forty two years ago I was raped by my brother and it was only last year that I was able to acknowledge that and begin dealing with the fallout from that keep hearing people say well why on earth would she weighs thirty six years and it really happened it must be alive and I can tell you exactly why it doesn't matter that we were victims when we are raped We feel wrong we feel shame and humiliation and embarrassment and we don't want to admit it to ourselves Let alone to anyone else if it was not a violent rate that left obvious marks it's easy to pretend it never happened and Emotionally it feels like the right thing to do Because you're so scared and several Gert and so is shame

Following the true story my name is David Bryant when I was thirty seven My wife and I decided to get term life insurance through select quote just three years later I was diagnosed with ... also known as Lou Gehrig's disease because of life insurance the people I Love most in this world will be protected no matter what happens to me. I Know that the dreams I have from my family and still come true financial security to stay in the home we became a family and and for our children to be able to go to college and I reached out to so I called because I wanted to share my story is just one person or just a life insurance I Know I Will have made a difference it. If there are people you care about and you need life insurance it may be the single most important financial decision ..

Figure 5: Sample stories from LVN & Talk Radio broadcasts

- LVN Corpus, a dataset of transcripts of facilitated in-person conversations designed to bring under-heard community voices, perspectives and stories to the center of a healthier public dialogue. Spanning over four different US cities, we use LVN [3] data specifically from Madison, WI, containing 102 conversations and 485 voices.

We filter stories from these transcript datasets using various template matching techniques. Figure 5 shows sample stories extracted from the above datasets. We apply our models to these stories and analyze the results of our model. Possible questions that can be answered using our modeling approach include: Can we infer patterns of intent and emotion trajectories across stories? Is there a pattern in the emotional trajectories of people with similar personas? Using such patterns, can we distinguish between stories and non-stories in text blocks? Is it possible to detect events or capture high points in those narratives? Besides using standard evaluation metrics on annotated gold set data, we use human judgements to ascertain our models' ability to transfer knowledge to these domains of stories.

3.2 Discerning Skills

3.2.1 Background

Discernment is the ability to decipher what is true and false, what is good or bad, or what is helpful or not helpful. Children with autism may be poor at discerning deceit, insincerity, manipulation and victimization by peers due to the difficulties in intuiting other humans' mental states and feelings. It is by now clear that strong ToM is associated with better discerning capabilities in social situations. While it is difficult to map and explore the complete space of discerning skills, we will focus on imparting the ability to discern anti-social attitudes and behaviours in this dissertation. A weak ToM is associated with greater levels of peer rejection [8] and increased risk of aggression and peer victimization [37]. This high risk of bullying victimization is demonstrated in children and

[3] https://lvn.org/about

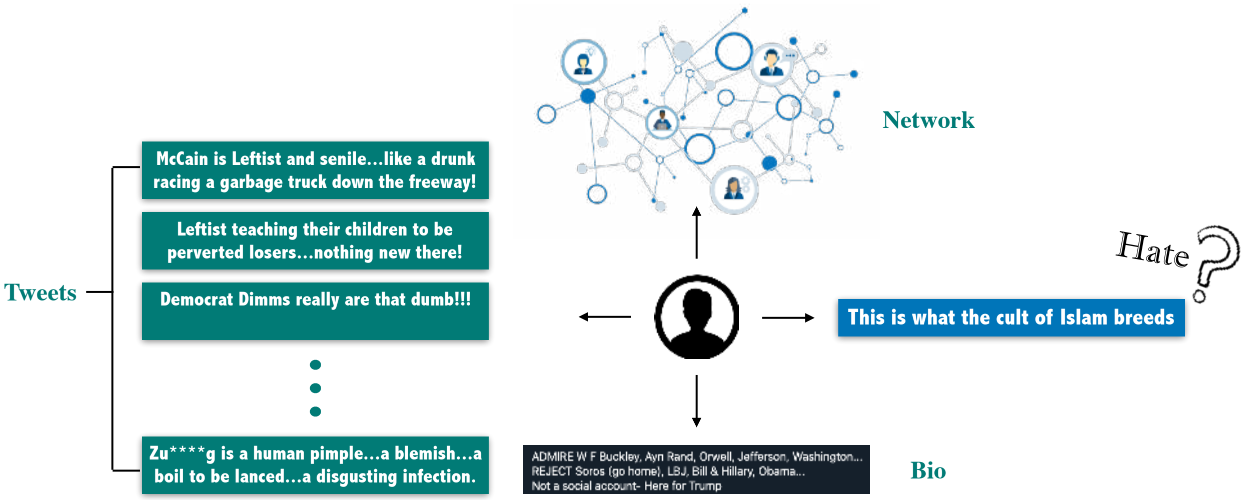
adolescents with autism. Several possible etiologies including aggressive behaviours, fewer friendships and communication problems [161] have been proposed. Consistent with several previous researches on the relationship between low ToM skills and victimization, the lack of discernment in others' social cues, intentions and emotions fueled by negative social behaviour, communication challenges, limited friendships and incorrect interpretation of what counts as bullying have been attributed as few of the crucial elements correlated to victimization [162, 163, 164, 165]. Similarly, machines still lack this ability to capture the motivations and emotions related to online social media content thereby crippling them from effectively understanding and detecting online hate, which is a unique form of peer victimization. Microsoft's chatbot, Tay, is an infamous example of how machines without social intelligence can turn into a racist troll within 24 hours of interaction with people on social media [166]. The failure of Tay chatbot underlines the importance of competent discernment and its relationship with ToM skills, particularly in the context of hate and victimization. In our dissertation, we introduce our human behaviour models to recognize and distinguish hateful or abusive content. This will help prevent the creation of potentially evil AI agents.

Past research has focused on identifying and analyzing impacts of different forms of online peer victimization. This includes use of hurtful comments accounting to hate speech or cyberbullying, dissemination of false or malicious information and misuse of private user details [167, 168, 169, 170, 170, 171, 172, 173]. A study by Gini et al. [174] emphasized the adverse consequences of such online victimization where the victims may be impelled to extremely destructive behaviours like suicide, overtly or covertly by the perpetrators. Efficient detection of hateful comments can be considered as a crucial step towards mitigating victimization. Majority of the work in this area of research [175] primarily use linguistic features from textual content. This limits the scope of interpreting online hate which is reflected in the results of their evaluations. Statistical techniques [176] have been applied to tackle bullying and hate using commonsense knowledge bases. In our research, we develop a hate speech detection model that incorporates socio-cultural cues along with representations obtained from human behaviour models. We demonstrate the advantages of integrating this additional knowledge and further discuss the potential of such models in the context of building socially-aware and responsible chatbots or conversational agents.

3.2.2 Research Plan: Understanding Online Hate Victimization

As explained in Section 3.2, children with limited ToM skills tend to have insufficient discernment abilities. Specifically, we focus on challenges associated with comprehending hurtful comments as it can lead to higher risk of becoming victims of bullying and harassment. Errors in understanding social cues in ambivalent social situations can be reduced by improving ToM capabilities. In our work, we focus on the influence of enhanced ToM in social media. In social media interactions, systems sans social intelligence perpetuate spread of hateful comments. Therefore, there is a growing need to identify and counter the problem of hateful content on social media. Online hate, a form of peer victimization, can have serious manifestations, including mental health issues, social polarization and hate crimes.

While prior works have proposed automated techniques to detect hate speech online, these techniques primarily fail to look beyond the textual content. Few attempts [48] have been made to study the personality traits of targets and instigators of hate. Some of the findings indicate similarities in personality traits such as low emotional awareness, and high anger and immoderation, which differ from personality traits of the general Twitter user population. Hence, interpreting



McCain is Leftist and senile...like a drunk racing a garbage truck down the freeway!

Leftist teaching their children to be perverted losers...nothing new there!

Democrat Dimms really are that dumb!!!

Zu****g is a human pimple...a blemish...a boil to be lanced...a disgusting infection.

This is what the cult of Islam breeds

ADMIRE W F Buckley, Ayn Rand, Orwell, Jefferson, Washington...
REJECT Soros (go home), LBJ, Bill & Hillary, Obama...
Not a social account- Here for Trump

Figure 6: Sample hate data along with the other features used for measuring hate.

socio-cultural cues and inferring personality characteristics or mental states of the user can help discern hateful expressions better. In this work, we propose a deep neural multi-modal model that can jointly learn from text semantics and representations of persona or mental states obtained from our behaviour models. [177] gives details about our earlier model. We extend this work by using users' tweets along with the text under consideration to obtain their social behavioural representations. We further use provide provide interpretable insights into decisions of our model. We integrate both the aspects of SAML into our hate detection model.

By performing ablation studies and thorough evaluation of different modeling techniques, we will demonstrate the advantage of applying socially aware models towards understanding online hate. We will also compare the models based on how the representations produced by these models can delineate different categories of hate. The challenge is to conduct experiments that establish the connection between mental states of hatred. Our preliminary results using our persona and mental state representations are given in Table 3. Character-based model seems to have an advantage as far as tweets are concerned due to its idiosyncratic nature. Figure 2 shows the multi-modal data used in our model. We leverage on network features and additional data for socio-cultural (SC) features and then use specific user tweets or context information to compute persona or mental states representing human behavioural (HB) features. Persona-based features provide a significant jump to our model compared to mental states due to the gap between the trained story-based behavior models and hate speech models using non-contextual tweet information or the incoherent list of tweets.

We would like to further run a thorough evaluation of our model by introducing some modifications to our model and better explain the decisions made by the model. This is a crucial aspect towards enhancing their robustness and building user trust on such models. To this end, we will employ adversarial example generation model as explained in Section 2.2.3 for our online hate classifier. This will be used in addition to other post-hoc methods intended to identify relevant features necessary for a particular classification outcome.

Datasets	Details
Founta et al. [178]	None: 53.8%; Hate: 4.96%; Abusive: 27.15%; Spam: 14%; Tweets: $\sim 100k$;
Davidson et al. [179]	None: 16.8%; Hate: 5.8%; Offensive: 77.4%; Tweets: $\sim 25k$;
Park & Fung [180]	None: 68%; Sexism: 20%; Racism: 11%; Tweets: $\sim 18k$;
Goelbeck et al. [181]	None: 74%; Harassment: 26%; Tweets: $\sim 21k$;
Our dataset	None: 58.1%; Hate: 16.6%; Abusive: 25.3% Tweets: $\sim 258k$;

Table 2: Summary of different datasets.

Model	F1 (Hate)	F1 (Overall)
DL Models: Text Only		
BiGRU+Attn	0.683	0.801
BiLSTM-2DCNN	0.661	0.795
BiGRU+Char+Attn	**0.744**	**0.864**
BERT	0.725	0.838
DL Models: Text+SC		
BERT+FF	0.766	0.876
BiGRU+Char+Attn+FF	**0.784**	**0.90**
DL Models: Text+SC+HB		
BERT+FF+Persona	0.803	0.909
BiGRU+Char+Attn+FF+Persona	**0.824**	**0.935**
BERT+FF+MS	0.784	0.881
BiGRU+Char+Attn+FF+MS	0.806	0.902

Table 3: Results of our preliminary experiments. (SC: Socio-cultural Features, HB: Human behaviour Features)

3.3 Pragmatic Skills

3.3.1 Background

Another major aspect of social development is effective communication with one's peers. For an effective communication, one should be able to understand the social context, beliefs, intentions and emotions of all the people involved. Prior works have emphasized the need for a well-developed ToM to engage meaningfully in conversations with capabilities to comprehend humor, metaphors, deception, irony and affective tone [41, 182]. Pragmatic functioning facilitates the appropriate use of language in different social contexts. Social language skills that are often used in our day-to-day social interactions are commonly referred to as Pragmatic language skills. In addition to social, cognitive and linguistic abilities, pragmatic competences skills also require awareness of the mental states of the other interacting agents and the capacity to manipulate complex representations of the communicative interaction, and answer differentially depending on the type of question asked [39, 182, 183, 184, 185]. Though the literature suggests robust association between language understanding and ToM development [186], findings regarding the contribution of ToM towards pragmatic competence is still inconclusive. For example, works by Norbury [41, 187, 188] and Happe [189] suggest conflicting conclusions regarding the role of ToM in pragmatic competence. Despite these findings, it is important to understand that the ability to understand others' mental states and their relation with behavior is an undisputed requirement of human communication [190]. Specifically, social-pragmatics is often alluded to circumstances that demand the ability of people to represent others' intentions, beliefs and emotions. Therefore, the relationship between social pragmatics and ToM can hardly be denied.

Many of the studies involving human-human interactions have clearly established the advantages of engaging in caring and empathetic conversations by reciprocating appropriately to the social cues and emotional state of their social partners [191, 192, 193, 194, 195]. The recent advances in AI have accelerated the interest in automated conversational agents in many areas. Such agents have become prevalent in customer services [196, 197, 198] and more recently in mental health care services (Woebot[4] or Tess [5]). Despite significant progress in syntactic [199, 200] and semantic processing [201, 202] of utterances, they remain far from pragmatic processing [203], a key skill displayed by humans. Drawing ideas from neuroscience psychology and linguistics, it is necessary to expand beyond syntactic and semantic understanding of utterances and process pragmatic clues to infer their complete meaning. Usually, the conversation partners understand the unstated (implicit) meaning of an utterance by comprehending the intentions of its speaker [204, 205]. Several language architectures trained on barely curated social media conversations or independent books, though improved the syntactic correctness of generated sentences, the risk of generating potentially insensitive responses bordering on bigotry and hate can't be discounted. [206, 207, 208, 209, 210, 166]. Since social pragmatic skills necessitate the understanding of the human expectations and behaviour, most automated agents either lack such skills [211, 212] or dodge this requirement by modifying the tone of responses laced with wit or humour (Cleverbot [6], Zo [7], Watson [8]). Though this makes for interesting and amusing conversational agents, lack of pragmatic reasoning diminishes their relevance in serious social contexts [196].

In our dissertation, we do not intend to solve all the challenges related to conversational agents. Instead, we propose to advance the research towards treating conversational agents as social actors that can gauge the emotions and motivations of people. By introducing our models that represent human social behaviour, we seek to address some of the questions that arise about the pragmatic social skills of such agents, extent to which such agents can engage emotionally with humans in natural social settings and the degree to which humans can calibrate trust on these agents.

3.3.2 Research Plan: Empathetic Response Generation

Endowing machines with pragmatic skills is one the fundamental challenges in AI research. There have been several efforts towards mimicking human behaviour through text based conversations. With the advent of personal intelligent assistants like Siri, Alexa, Google Assistant, the daunting challenge for such conversational agents is to be natural and believable by rising beyond being mere

[4] https://woebot.io

[5] http://x2ai.com/

[6] https://www.cleverbot.com

[7] https://twitter.com/zochats

[8] https://www.ibm.com/watson/how-to-build-a-chatbot

conveyors of information and eventually be able to build a bond with human users. By understanding human social behaviour and establishing an emotional connection with people, conversational agents can turn into an efficient social actor which can appropriately communicate to people knowing what, how and when to say certain things in different social situations. Since our behaviour models are primarily intended to learn, understand and represent human personality traits or intents and emotion, we leverage these models in different conversational systems. Some of the prior works have emphasized the need for empathetic conversation modeling, knowledge and memory modeling, interpretable machine intelligence, deep reasoning, cross-media and continuous streaming artificial intelligence, and modeling and calibration of emotional or intrinsic rewards reflected in human needs [213] for an open domain social conversational system. Though our behaviour models do not comprise of all the mentioned features, we definitely lay our attention on the importance of inferring personas, intent and emotion based on human behaviour, assimilating social commonsense knowledge in a memory and finally, giving insights into the decisions of the model. We believe that our approach can expedite the research towards building an efficient human-like conversational systems.

In this work, we investigate different techniques to take advantage of our human social behaviour models and validate its relevance in improving pragmatic skills for conversational systems. This will involve dynamic recognition of personas, intents and emotions of the conversation partner and generate appropriate interpersonal responses. We do not claim to build a conversational agent from scratch that can impart all the pragmatic skills required for an efficient social interaction. However, we will demonstrate the advantage of applying the knowledge from the trained behaviour models in the form of a more refined and empathetic responses. Several language architectures are trained on text scraped from social media conversations without much curation [208]. It is likely that these models trained on such data could throw aggressive responses. Therefore, in addition to the behaviour models, we will use our online hate detection model to ensure no such hurtful responses are generated.

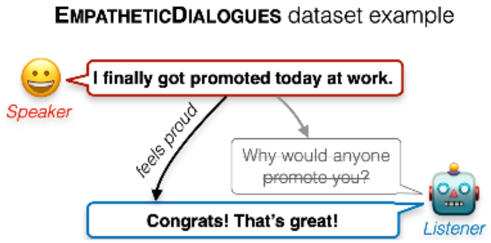

EMPATHETICDIALOGUES dataset example
I finally got promoted today at work.
Speaker
feels proud
Why would anyone ~~promote you?~~
Congrats! That's great!
Listener

Figure 7: Sample from *EmpatheticDialogues* dataset as illustrated in [214].

The key idea is to encourage the generator to produce responses with augmented knowledge based on persona or intent/emotion representations inferred from past utterances of the conversation partner. These representations can be used as an additional input to the model to encode the current mental state of the interacting user. The most likely response can be computed after assessing the agent's own behaviour by rewarding them using our models for hate detection and intent/emotion inference. By encoding interacting users' persona and intent/emotions and rewarding responses that are non-hurtful and driven by positive and intent/emotions, we hope to improve the pragmatic abilities of the conversational agent. Besides the behavioural representations, we are also interested in utilizing the memory models explained in Section 3.3 as read-only knowledge store for generation of empathetic responses. Evaluating conversational agents is still an open research problem. The need for empathy in responses adds another layer of complexity in evaluating such models. We use *EmpatheticDialogues* dataset [214] as one of the benchmark datasets to evaluate our models' ability to generate empathetic responses. In addition to automated metrics for evaluating these responses, we also compare their social appropriateness using human evaluations.

4 Contributions & Expected Results

The overall goal of this dissertation is to advance the research towards building socially intelligent AI systems which in turn can help us make progress towards the development of AGI. We will approach this idea through different projects, that will make the following contributions:

- Developing models towards understanding human social behaviour with suitable knobs to continually learn and improve.
- Demonstrate the power of our representations obtained from behaviour models in different ToM-critical social competence tasks requiring narrative processing skills, discerning skills and pragmatic skills.

Thee projects will be supported by publication-quality research papers and open source code. Given these efforts are successful,it could potentially lead to the development of sophisticated AI systems that is well-aligned with human behavior, goals, and preferences.

5 Proposed Timeline

March 2020 - April 2020

- Finalize mental state models and conduct experiments.
- Run thorough evaluation of these models.
- Write up the results of the experiments and submit paper for publication.

April 2020 - June 2020

- Train/Fine-tune our online hate and empathetic response generation model using the representations obtained from personas/mental-state models.
- Conduct experiments comparing the state-of-the-art models for online hate and empathetic dialogue systems and evaluate the results.
- Write and submit the empathetic response generation paper for publication.

June 2020 - August 2020

- Write and revise the dissertation.
- Dissertation Defense.

6 Resources Required

The following resources will be required or useful for the completion of the proposed research:

- Computational resources, especially extensive GPU hours and data storage to train models.
- Funds to hire undergraduate assistants.
- Funds to hire mechanical turks, crucial for evaluations.
- Funds for conference travel to present the research.
- Assistance in disseminating links to data collection websites to encourage participation.

References

[1] S. Baron-Cohen, H. A. Ring, S. Wheelwright, E. T. Bullmore, M. J. Brammer, A. Simmons, and S. C. Williams, "Social intelligence in the normal and autistic brain: an fmri study," *European journal of neuroscience*, vol. 11, no. 6, pp. 1891–1898, 1999.

[2] H. Wimmer and J. Perner, "Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception," *Cognition*, vol. 13, no. 1, pp. 103–128, 1983.

[3] Z.-T. Yeh, "Role of theory of mind and executive function in explaining social intelligence: A structural equation modeling approach," *Aging & Mental Health*, vol. 17, no. 5, pp. 527–534, 2013.

[4] S. Grueneisen, E. Wyman, and M. Tomasello, "i know you don't know i know children use second-order false-belief reasoning for peer coordination," *Child development*, vol. 86, no. 1, pp. 287–293, 2015.

[5] M. Caputi, S. Lecce, A. Pagnin, and R. Banerjee, "Longitudinal effects of theory of mind on later peer relations: the role of prosocial behavior.," *Developmental psychology*, vol. 48, no. 1, p. 257, 2012.

[6] V. Slaughter, K. Imuta, C. C. Peterson, and J. D. Henry, "Meta-analysis of theory of mind and peer popularity in the preschool and early school years," *Child development*, vol. 86, no. 4, pp. 1159–1174, 2015.

[7] L. Atkinson, L. Slade, D. Powell, and J. P. Levy, "Theory of mind in emerging reading comprehension: A longitudinal study of early indirect and direct effects," *Journal of Experimental Child Psychology*, vol. 164, pp. 225–238, 2017.

[8] R. Banerjee, D. Watling, and M. Caputi, "Peer relations and the understanding of faux pas: Longitudinal evidence for bidirectional associations," *Child development*, vol. 82, no. 6, pp. 1887–1905, 2011.

[9] S. Berggren, "Emotion recognition and expression in autism spectrum disorder: Significance, complexity, and effect of training," 2017.

[10] C. Maiano, C. L. Normand, M.-C. Salvas, G. Moullec, and A. Aimé, "Prevalence of school bullying among youth with autism spectrum disorders: A systematic review and meta-analysis," *Autism research*, vol. 9, no. 6, pp. 601–615, 2016.

[11] A. M. Leslie and U. Frith, "Autistic children's understanding of seeing, knowing and believing," *British Journal of Developmental Psychology*, vol. 6, no. 4, pp. 315–324, 1988.

[12] J. M. Lucariello, T. M. Durand, and L. Yarnell, "Social versus intrapersonal tom: Social tom is a cognitive strength for low-and middle-ses children," *Journal of Applied Developmental Psychology*, vol. 28, no. 4, pp. 285–297, 2007.

[13] C. Moore, *The development of commonsense psychology*. Psychology Press, 2013.

[14] A. Gopnik and H. M. Wellman, "Why the child's theory of mind really is a theory," *Mind & Language*, vol. 7, no. 1-2, pp. 145–171, 1992.

[15] D. I. Tamir, M. A. Thornton, J. M. Contreras, and J. P. Mitchell, "Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence," *Proceedings of the National Academy of Sciences*, vol. 113, no. 1, pp. 194–199, 2016.

[16] M. A. Thornton and J. P. Mitchell, "Theories of person perception predict patterns of neural activity during mentalizing," *Cerebral cortex*, vol. 28, no. 10, pp. 3505–3520, 2017.

[17] M. A. Thornton and J. P. Mitchell, "Consistent neural activity patterns represent personally familiar people," *Journal of cognitive neuroscience*, vol. 29, no. 9, pp. 1583–1594, 2017.

[18] H. Liu and P. Singh, "Conceptnet–a practical commonsense reasoning tool-kit," *BT technology journal*, vol. 22, no. 4, pp. 211–226, 2004.

[19] A. Gordon, A. Kazemzadeh, A. Nair, and M. Petrova, "Recognizing expressions of commonsense psychology in english text," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pp. 208–215, Association for Computational Linguistics, 2003.

[20] M. Sap, H. Rashkin, D. Chen, R. LeBras, and Y. Choi, “Socialiqa: Commonsense reasoning about social interactions,” *arXiv preprint arXiv:1904.09728*, 2019.

[21] M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi, “Atomic: an atlas of machine commonsense for if-then reasoning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 3027–3035, 2019.

[22] H. Rashkin, M. Sap, E. Allaway, N. A. Smith, and Y. Choi, “Event2mind: Commonsense inference on events, intents, and reactions,” *arXiv preprint arXiv:1805.06939*, 2018.

[23] A. S. Gordon, “Commonsense interpretation of triangle behavior,” in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[24] E. Grant, A. Nematzadeh, and T. L. Griffiths, “How can memory-augmented neural networks pass a false-belief task?,”

[25] A. Talmor, J. Herzig, N. Lourie, and J. Berant, “Commonsenseqa: A question answering challenge targeting commonsense knowledge,” *arXiv preprint arXiv:1811.00937*, 2018.

[26] A. Nematzadeh, K. Burns, E. Grant, A. Gopnik, and T. L. Griffiths, “Evaluating theory of mind in question answering,” *arXiv preprint arXiv:1808.09352*, 2018.

[27] R. Saxe and N. Kanwisher, “People thinking about thinking people: the role of the temporo-parietal junction in theory of mind,” *Neuroimage*, vol. 19, no. 4, pp. 1835–1842, 2003.

[28] C. L. Baker, R. Saxe, and J. B. Tenenbaum, “Action understanding as inverse planning,” *Cognition*, vol. 113, no. 3, pp. 329–349, 2009.

[29] K. Oatley, “Why fiction may be twice as true as fact: Fiction as cognitive and emotional simulation,” *Review of general psychology*, vol. 3, no. 2, pp. 101–117, 1999.

[30] N. Epley, B. Keysar, L. Van Boven, and T. Gilovich, “Perspective taking as egocentric anchoring and adjustment.,” *Journal of personality and social psychology*, vol. 87, no. 3, p. 327, 2004.

[31] J. Zaki, “Empathy: a motivated account.,” *Psychological bulletin*, vol. 140, no. 6, p. 1608, 2014.

[32] A. Mehrabian and N. Epstein, “A measure of emotional empathy 1,” *Journal of personality*, vol. 40, no. 4, pp. 525–543, 1972.

[33] N. Eisenberg, *Altruistic emotion, cognition, and behavior (PLE: Emotion)*. Psychology Press, 2014.

[34] H. M. Wellman and J. G. Miller, “Including deontic reasoning as fundamental to theory of mind,” *Human Development*, vol. 51, no. 2, pp. 105–135, 2008.

[35] K. Imuta, J. D. Henry, V. Slaughter, B. Selcuk, and T. Ruffman, “Theory of mind and prosocial behavior in childhood: A meta-analytic review.,” *Developmental psychology*, vol. 52, no. 8, p. 1192, 2016.

[36] L. Surian, “Are children with autism deaf to gricean maxims?,” *Cognitive neuropsychiatry*, vol. 1, no. 1, pp. 55–72, 1996.

[37] S. Shakoor, S. R. Jaffee, L. Bowes, I. Ouellet-Morin, P. Andreou, F. Happé, T. E. Moffitt, and L. Arseneault, “A prospective longitudinal study of childrens theory of mind and adolescent involvement in bullying,” *Journal of Child Psychology and Psychiatry*, vol. 53, no. 3, pp. 254–261, 2012.

[38] J. W. Astington and J. Pelletier, “Theory of mind, language, and learning in the early years: Developmental origins of school readiness,” *The development of social cognition and communication*, pp. 205–230.

[39] C. Fernández, “Mindful storytellers: Emerging pragmatics and theory of mind development,” *First Language*, vol. 33, no. 1, pp. 20–46, 2013.

[40] A. McKeough, “A neo-structural analysis of childrens narrative and its development,” *The minds staircase: Exploring the conceptual underpinnings of childrens thought and knowledge*, pp. 171–188, 1992.

[41] C. F. Norbury, “Barking up the wrong tree? lexical ambiguity resolution in children with language impairments and autistic spectrum disorders,” *Journal of experimental child psychology*, vol. 90, no. 2, pp. 142–171, 2005.

[42] J. Wessberg, C. R. Stambaugh, J. D. Kralik, P. D. Beck, M. Laubach, J. K. Chapin, J. Kim, S. J. Biggs, M. A. Srinivasan, and M. A. Nicolelis, "Real-time prediction of hand trajectory by ensembles of cortical neurons in primates," *Nature*, vol. 408, no. 6810, p. 361, 2000.

[43] A. J. Elliot and T. M. Thrash, "Approach-avoidance motivation in personality: approach and avoidance temperaments and goals.," *Journal of personality and social psychology*, vol. 82, no. 5, p. 804, 2002.

[44] M. A. Thornton and D. I. Tamir, "Mental models accurately predict emotion transitions," *Proceedings of the National Academy of Sciences*, vol. 114, no. 23, pp. 5982–5987, 2017.

[45] H. von Helmholtz, *Treatise on physiological optics*, vol. 3. Courier Corporation, 2013.

[46] Y. Luan, C. Brockett, B. Dolan, J. Gao, and M. Galley, "Multi-task learning for speaker-role adaptation in neural conversation models," *arXiv preprint arXiv:1710.07388*, 2017.

[47] J. Li, M. Galley, C. Brockett, G. P. Spithourakis, J. Gao, and B. Dolan, "A persona-based neural conversation model," *arXiv preprint arXiv:1603.06155*, 2016.

[48] M. ElSherief, S. Nilizadeh, D. Nguyen, G. Vigna, and E. Belding, "Peer to peer hate: Hate speech instigators and their targets," in *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[49] F. Xia, N. Y. Asabere, H. Liu, Z. Chen, and W. Wang, "Socially aware conference participant recommendation with personality traits," *IEEE Systems Journal*, vol. 11, no. 4, pp. 2255–2266, 2014.

[50] J. Banerjee, G. Raravi, M. Gupta, S. K. Ernala, S. Kunde, and K. Dasgupta, "Capres: context aware persona based recommendation for shoppers," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[51] T. H. Trinh and Q. V. Le, "A simple method for commonsense reasoning," *arXiv preprint arXiv:1806.02847*, 2018.

[52] E. Davis and G. Marcus, "Commonsense reasoning and commonsense knowledge in artificial intelligence.," *Commun. ACM*, vol. 58, no. 9, pp. 92–103, 2015.

[53] M. M. De Graaf and B. F. Malle, "How people explain action (and autonomous intelligent systems should too)," in *2017 AAAI Fall Symposium Series*, 2017.

[54] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, p. 93, 2019.

[55] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, *et al.*, "Interpretability of deep learning models: a survey of results," in *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pp. 1–6, IEEE, 2017.

[56] S. G. Shamay-Tsoory and J. Aharon-Peretz, "Dissociable prefrontal networks for cognitive and affective theory of mind: a lesion study," *Neuropsychologia*, vol. 45, no. 13, pp. 3054–3067, 2007.

[57] R. M. Ryckman, "Theory of personality," *USA. Michele Sordi*, 2004.

[58] D. I. Tamir and M. A. Thornton, "Modeling the predictive social mind," *Trends in cognitive sciences*, vol. 22, no. 3, pp. 201–212, 2018.

[59] H. M. Gray, K. Gray, and D. M. Wegner, "Dimensions of mind perception," *science*, vol. 315, no. 5812, pp. 619–619, 2007.

[60] W. Bridewell and A. Isaac, "Recognizing deception: A model of dynamic belief attribution," in *2011 AAAI Fall Symposium Series*, 2011.

[61] S. E. Fahlman, "Using scone's multiple-context mechanism to emulate human-like reasoning," in *2011 AAAI Fall Symposium Series*, 2011.

[62] P. H. Winston, "The genesis story understanding and story telling system a 21st century step toward artificial intelligence," tech. rep., Center for Brains, Minds and Machines (CBMM), 2014.

[63] D. B. Lenat, R. V. Guha, K. Pittman, D. Pratt, and M. Shepherd, “Cyc: toward programs with common sense,” *Communications of the ACM*, vol. 33, no. 8, pp. 30–49, 1990.

[64] B. Harrison and M. O. Riedl, “Towards learning from stories: An approach to interactive machine learning,” in *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[65] A. Souri, S. Hosseinpour, and A. M. Rahmani, “Personality classification based on profiles of social networks users and the five-factor model of personality,” *Human-centric Computing and Information Sciences*, vol. 8, no. 1, p. 24, 2018.

[66] D. Bamman, B. OConnor, and N. A. Smith, “Learning latent personas of film characters,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 352–361, 2013.

[67] C. L. Baker, J. Jara-Ettinger, R. Saxe, and J. B. Tenenbaum, “Rational quantitative attribution of beliefs, desires and percepts in human mentalizing,” *Nature Human Behaviour*, vol. 1, no. 4, pp. 1–10, 2017.

[68] W. Yoshida, R. J. Dolan, and K. J. Friston, “Game theory of mind,” *PLoS computational biology*, vol. 4, no. 12, 2008.

[69] C. Baker, R. Saxe, and J. Tenenbaum, “Bayesian theory of mind: Modeling joint belief-desire attribution,” in *Proceedings of the annual meeting of the cognitive science society*, vol. 33, 2011.

[70] A. Lerer and A. Peysakhovich, “Learning social conventions in markov games,” *arXiv preprint arXiv:1806.10071*, 2018.

[71] M. Lanctot, V. Zambaldi, A. Gruslys, A. Lazaridou, K. Tuyls, J. Pérolat, D. Silver, and T. Graepel, “A unified game-theoretic approach to multiagent reinforcement learning,” in *Advances in Neural Information Processing Systems*, pp. 4190–4203, 2017.

[72] S. M. J. Angwin, J. Larson and L. Kirchner, “Machine bias.” https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016.

[73] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.

[74] R. Jia and P. Liang, “Adversarial examples for evaluating reading comprehension systems,” *arXiv preprint arXiv:1707.07328*, 2017.

[75] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1765–1773, 2017.

[76] A. Nguyen, J. Yosinski, and J. Clune, “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 427–436, 2015.

[77] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, “Explaining explanations: An overview of interpretability of machine learning,” in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pp. 80–89, IEEE, 2018.

[78] R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases,” in *Acm sigmod record*, vol. 22, pp. 207–216, ACM, 1993.

[79] C. Rudin, B. Letham, and D. Madigan, “Learning theory analysis for association rules and sequential event prediction,” *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 3441–3492, 2013.

[80] B. Letham, C. Rudin, T. H. McCormick, D. Madigan, *et al.*, “Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model,” *The Annals of Applied Statistics*, vol. 9, no. 3, pp. 1350–1371, 2015.

[81] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.

[82] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *International conference on machine learning*, pp. 2048–2057, 2015.

[83] S. Vashishth, S. Upadhyay, G. S. Tomar, and M. Faruqui, “Attention interpretability across nlp tasks,” *arXiv preprint arXiv:1909.11218*, 2019.

[84] S. Serrano and N. A. Smith, “Is attention interpretable?,” *arXiv preprint arXiv:1906.03731*, 2019.

[85] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Annals of statistics*, pp. 1189–1232, 2001.

[86] Q. Zhao and T. Hastie, “Causal interpretations of black-box models,” *Journal of Business & Economic Statistics*, no. just-accepted, pp. 1–19, 2019.

[87] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, “Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation,” *Journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65, 2015.

[88] D. W. Apley, “Visualizing the effects of predictor variables in black box supervised learning models,” *arXiv preprint arXiv:1612.08468*, 2016.

[89] E. Štrumbelj and I. Kononenko, “Explaining prediction models and individual predictions with feature contributions,” *Knowledge and information systems*, vol. 41, no. 3, pp. 647–665, 2014.

[90] M. Staniak and P. Biecek, “Explanations of model predictions with live and breakdown packages,” *arXiv preprint arXiv:1804.01955*, 2018.

[91] M. T. Ribeiro, S. Singh, and C. Guestrin, “Model-agnostic interpretability of machine learning,” *arXiv preprint arXiv:1606.05386*, 2016.

[92] N. Morch, U. Kjems, L. K. Hansen, C. Svarer, I. Law, B. Lautrup, S. Strother, and K. Rehm, “Visualization of neural networks using saliency maps,” in *Proceedings of ICNN’95-International Conference on Neural Networks*, vol. 4, pp. 2085–2090, IEEE, 1995.

[93] Q.-s. Zhang and S.-C. Zhu, “Visual interpretability for deep learning: a survey,” *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 27–39, 2018.

[94] K. Xu, S. Liu, P. Zhao, P.-Y. Chen, H. Zhang, Q. Fan, D. Erdogmus, Y. Wang, and X. Lin, “Structured adversarial attack: Towards general implementation and better interpretability,” *arXiv preprint arXiv:1808.01664*, 2018.

[95] Y. Shoshan and V. Ratner, “Regularized adversarial examples for model interpretability,” *arXiv preprint arXiv:1811.07311*, 2018.

[96] Y. Dong, H. Su, J. Zhu, and F. Bao, “Towards interpretable deep neural networks by leveraging adversarial examples,” *arXiv preprint arXiv:1708.05493*, 2017.

[97] X. Yuan, P. He, Q. Zhu, and X. Li, “Adversarial examples: Attacks and defenses for deep learning,” *IEEE transactions on neural networks and learning systems*, 2019.

[98] W. Woods, J. Chen, and C. Teuscher, “Reliable classification explanations via adversarial attacks on robust networks,” *arXiv preprint arXiv:1906.02896*, 2019.

[99] C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin, “This looks like that: deep learning for interpretable image recognition,” *arXiv preprint arXiv:1806.10574*, 2018.

[100] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison, and M. O. Riedl, “Automated rationale generation: a technique for explainable ai and its effects on human perceptions,” in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pp. 263–274, ACM, 2019.

[101] M. R. Barrick and M. K. Mount, “Autonomy as a moderator of the relationships between the big five personality dimensions and job performance.,” *Journal of applied Psychology*, vol. 78, no. 1, p. 111, 1993.

[102] P. R. Shaver and K. A. Brennan, “Attachment styles and the” big five” personality traits: Their connections with each other and with romantic relationship outcomes,” *Personality and Social Psychology Bulletin*, vol. 18, no. 5, pp. 536–545, 1992.

[103] C. Nass and K. M. Lee, “Does computer-generated speech manifest personality? an experimental test of similarity-attraction,” in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 329–336, 2000.

[104] M. Selfhout, W. Burk, S. Branje, J. Denissen, M. Van Aken, and W. Meeus, “Emerging late adolescent friendship networks and big five personality traits: A social network approach,” *Journal of personality*, vol. 78, no. 2, pp. 509–538, 2010.

[105] J. T. Jost, C. M. Federico, and J. L. Napier, “Political ideology: Its structure, functions, and elective affinities,” *Annual review of psychology*, vol. 60, pp. 307–337, 2009.

[106] L. R. Goldberg, “An alternative” description of personality”: the big-five factor structure.,” *Journal of personality and social psychology*, vol. 59, no. 6, p. 1216, 1990.

[107] M. Liu, Y. Wu, D. Jiao, M. S. Wu, and T. Zhu, “Literary intelligence analysis of novel protagonists personality traits and development,” *Digital Scholarship in the Humanities*, vol. 34, no. 1, pp. 221–229, 2019.

[108] L. Flekova and I. Gurevych, “Personality profiling of fictional characters using sense-level links between lexical resources,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1805–1816, 2015.

[109] K. Shuster, S. Humeau, A. Bordes, and J. Weston, “Engaging image chat: Modeling personality in grounded dialogue,” *arXiv preprint arXiv:1811.00945*, 2018.

[110] K. Chandu, S. Prabhumoye, R. Salakhutdinov, and A. W. Black, “my way of telling a story: Persona based grounded story generation,” in *Proceedings of the Second Workshop on Storytelling*, pp. 11–21, 2019.

[111] D. P. McAdams and E. Manczak, “Personality and the life story.,” 2015.

[112] J. W. Pennebaker and L. A. King, “Linguistic styles: Language use as an individual difference.,” *Journal of personality and social psychology*, vol. 77, no. 6, p. 1296, 1999.

[113] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, “Using linguistic cues for the automatic recognition of personality in conversation and text,” *Journal of artificial intelligence research*, vol. 30, pp. 457–500, 2007.

[114] E. Chu, P. Vijayaraghavan, and D. Roy, “Learning personas from dialogue with attentive memory networks,” *arXiv preprint arXiv:1810.08717*, 2018.

[115] R. A. Mar, K. Oatley, J. Hirsh, J. Dela Paz, and J. B. Peterson, “Bookworms versus nerds: Exposure to fiction versus non-fiction, divergent associations with social ability, and the simulation of fictional social worlds,” *Journal of Research in Personality*, vol. 40, no. 5, pp. 694–712, 2006.

[116] R. S. Wyer Jr, *Knowledge and Memory: The Real Story: Advances in Social Cognition, Volume VIII*. Psychology Press, 2014.

[117] N. Mostafazadeh, N. Chambers, X. He, D. Parikh, D. Batra, L. Vanderwende, P. Kohli, and J. Allen, “A corpus and cloze evaluation for deeper understanding of commonsense stories,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (San Diego, California), pp. 839–849, Association for Computational Linguistics, June 2016.

[118] F. G. Happé, “An advanced test of theory of mind: Understanding of story characters’ thoughts and feelings by able autistic, mentally handicapped, and normal children and adults,” *Journal of autism and Developmental disorders*, vol. 24, no. 2, pp. 129–154, 1994.

[119] T. Jolliffe and S. Baron-Cohen, “The strange stories test: A replication with high-functioning adults with autism or asperger syndrome,” *Journal of autism and developmental disorders*, vol. 29, no. 5, pp. 395–406, 1999.

[120] Y. Xia, F. Tian, L. Wu, J. Lin, T. Qin, N. Yu, and T.-Y. Liu, “Deliberation networks: Sequence generation beyond one-pass decoding,” in *Advances in Neural Information Processing Systems*, pp. 1784–1794, 2017.

[121] J. Weston, S. Chopra, and A. Bordes, "Memory networks," *arXiv preprint arXiv:1410.3916*, 2014.

[122] H. Rashkin, A. Bosselut, M. Sap, K. Knight, and Y. Choi, "Modeling naive psychology of characters in simple commonsense stories," *arXiv preprint arXiv:1805.06533*, 2018.

[123] G. Tao, S. Ma, Y. Liu, and X. Zhang, "Attacks meet interpretability: Attribute-steered detection of adversarial samples," in *Advances in Neural Information Processing Systems*, pp. 7717–7728, 2018.

[124] S. Ritter, D. G. Barrett, A. Santoro, and M. M. Botvinick, "Cognitive psychology for deep neural networks: A shape bias case study," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2940–2949, JMLR. org, 2017.

[125] P. Vijayaraghavan and D. Roy, "Generating black-box adversarial examples for text classifiers using a deep reinforced model," *arXiv preprint arXiv:1909.07873*, 2019.

[126] K. H. Rubin and L. Rose-Krasnor, "Interpersonal problem solving and social competence in children," in *Handbook of social development*, pp. 283–323, Springer, 1992.

[127] L. K. Friedrich and A. H. Stein, "Aggressive and prosocial television programs and the natural behavior of preschool children," *Monographs of the Society for Research in Child Development*, pp. 1–64, 1973.

[128] S. Perren, S. Stadelmann, A. Von Wyl, and K. Von Klitzing, "Pathways of behavioural and emotional symptoms in kindergarten children: What is the role of pro-social behaviour?," *European Child & Adolescent Psychiatry*, vol. 16, no. 4, pp. 209–214, 2007.

[129] L. J. Lengua, "Associations among emotionality, self-regulation, adjustment problems, and positive adjustment in middle childhood," *Journal of Applied Developmental Psychology*, vol. 24, no. 5, pp. 595–618, 2003.

[130] Y. Kimhi, "Theory of mind abilities and deficits in autism spectrum disorders," *Topics in Language Disorders*, vol. 34, no. 4, pp. 329–343, 2014.

[131] L. V. Badenes, R. A. Clemente Estevan, and F. J. García Bacete, "Theory of mind and peer rejection at school," *Social Development*, vol. 9, no. 3, pp. 271–283, 2000.

[132] J. M. Jenkins and J. W. Astington, "Theory of mind and social behavior: Causal models tested in a longitudinal study," *Merrill-Palmer Quarterly (1982-)*, pp. 203–220, 2000.

[133] J. G. Suway, K. A. Degnan, A. L. Sussman, and N. A. Fox, "The relations among theory of mind, behavioral inhibition, and peer interactions in early childhood," *Social Development*, vol. 21, no. 2, pp. 331–342, 2012.

[134] J. W. Astington, "Sometimes necessary, never sufficient: False-belief understanding and social competence," in *Individual differences in theory of mind*, pp. 24–49, Psychology Press, 2004.

[135] C. Hughes, "Finding your marbles: Does preschoolers' strategic behavior predict later understanding of mind?," *Developmental psychology*, vol. 34, no. 6, p. 1326, 1998.

[136] A. Klin, W. Jones, R. Schultz, F. Volkmar, and D. Cohen, "Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism," *Archives of general psychiatry*, vol. 59, no. 9, pp. 809–816, 2002.

[137] N. Bauminger-Zviely, *Social and academic abilities in children with high-functioning autism spectrum disorders*. 2013.

[138] E. Gould, J. Tarbox, D. O'Hora, S. Noone, and R. Bergstrom, "Teaching children with autism a basic component skill of perspective-taking," *Behavioral Interventions*, vol. 26, no. 1, pp. 50–66, 2011.

[139] J. Paynter and C. C. Peterson, "Further evidence of benefits of thought-bubble training for theory of mind development in children with autism spectrum disorders," *Research in Autism Spectrum Disorders*, vol. 7, no. 2, pp. 344–348, 2013.

[140] T. MacKay, F. Knott, and A.-W. Dunlop, "Developing social interaction and understanding in individuals with autism spectrum disorder: A groupwork intervention," *Journal of Intellectual and Developmental Disability*, vol. 32, no. 4, pp. 279–290, 2007.

[141] N. Bauminger, "Brief report: Group social-multimodal intervention for hfasd," *Journal of autism and developmental disorders*, vol. 37, no. 8, pp. 1605–1615, 2007.

[142] N. Bauminger-Zviely, S. Eden, M. Zancanaro, P. L. Weiss, and E. Gal, "Increasing social engagement in children with high-functioning autism spectrum disorder using collaborative technologies in the school environment," *Autism*, vol. 17, no. 3, pp. 317–339, 2013.

[143] F. Ziegler, P. Mitchell, and G. Currie, "How does narrative cue children's perspective taking?," *Developmental Psychology*, vol. 41, no. 1, p. 115, 2005.

[144] R. A. Mar, "The neuropsychology of narrative: Story comprehension, story production and their interrelation," *Neuropsychologia*, vol. 42, no. 10, pp. 1414–1434, 2004.

[145] R. A. Mar, "The neural bases of social cognition and story comprehension," *Annual review of psychology*, vol. 62, pp. 103–134, 2011.

[146] R. M. García-Pérez, R. P. Hobson, and A. Lee, "Narrative role-taking in autism," *Journal of Autism and Developmental Disorders*, vol. 38, no. 1, pp. 156–168, 2008.

[147] T. Charman and Y. Shmueli-Goetz, "The relationship between theory of mind, language and narrative discourse: an experimental study.," *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, 1998.

[148] C. R. Carnahan, P. S. Williamson, and J. Christman, "Linking cognition and literacy in students with autism spectrum disorder," *Teaching Exceptional Children*, vol. 43, no. 6, pp. 54–62, 2011.

[149] N. R. Guajardo and A. C. Watson, "Narrative discourse and theory of mind development," *The Journal of Genetic Psychology*, vol. 163, no. 3, pp. 305–325, 2002.

[150] K. F. Navona Calarco, M. Rain, and R. A. Mar, "Absorption in narrative fiction and its possible impact on social abilities," *Narrative Absorption*, vol. 27, p. 293, 2017.

[151] T. Winograd, "Understanding natural language," *Cognitive psychology*, vol. 3, no. 1, pp. 1–191, 1972.

[152] S. R. Turner, *The creative process: A computer model of storytelling and creativity*. Psychology Press, 2014.

[153] L. K. Schubert and C. H. Hwang, "Episodic logic meets little red riding hood: A comprehensive, natural representation for language understanding,"

[154] E. Charniak, *Toward a model of children's story comprehension*. PhD thesis, Massachusetts Institute of Technology, 1972.

[155] J. Valls-Vargas, J. Zhu, and S. Ontañón, "Narrative hermeneutic circle: Improving character role identification from natural language text via feedback loops," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[156] S. Srivastava, S. Chaturvedi, and T. Mitchell, "Inferring interpersonal relations in narrative summaries," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[157] M. L. Jockers, *Macroanalysis: Digital methods and literary history*. University of Illinois Press, 2013.

[158] M. M. A. Finlayson, *Learning narrative structure from annotated folktales*. PhD thesis, Massachusetts Institute of Technology, 2012.

[159] N. Chambers and D. Jurafsky, "Unsupervised learning of narrative event chains," in *Proceedings of ACL-08: HLT*, pp. 789–797, 2008.

[160] D. Beeferman, W. Brannon, and D. Roy, "Radiotalk: a large-scale corpus of talk radio transcripts," *arXiv preprint arXiv:1907.07073*, 2019.

[161] A. Y. Mikami, D. E. Szwedo, J. P. Allen, M. A. Evans, and A. L. Hare, "Adolescent peer relationships and behavior problems predict young adults communication on social networking websites.," *Developmental psychology*, vol. 46, no. 1, p. 46, 2010.

[162] C. Hughes, A. L. Cutting, and J. Dunn, “Acting nasty in the face of failure? longitudinal observations of hard-to-manage children playing a rigged competitive game with a friend,” *Journal of Abnormal Child Psychology*, vol. 29, no. 5, pp. 403–416, 2001.

[163] L. Gasser and M. Keller, “Are the competent the morally good? perspective taking and moral motivation of children involved in bullying,” *Social Development*, vol. 18, no. 4, pp. 798–816, 2009.

[164] J. Sutton, P. K. Smith, and J. Swettenham, “Social cognition and bullying: Social inadequacy or skilled manipulation?,” *British Journal of Developmental Psychology*, vol. 17, no. 3, pp. 435–450, 1999.

[165] N. Humphrey and W. Symes, “Perceptions of social support and experience of bullying among pupils with autistic spectrum disorders in mainstream secondary schools,” *European Journal of Special Needs Education*, vol. 25, no. 1, pp. 77–91, 2010.

[166] J. West, “Microsofts disastrous tay experiment shows the hidden dangers of ai.” https://qz.com/653084/microsofts-disastrous-tay-experiment-shows-the-hidden-dangers-of-ai/, 2016.

[167] C. Yang and P. Srinivasan, “Translating surveys to surveillance on social media: methodological challenges & solutions,” in *Proceedings of the 2014 ACM conference on Web science*, pp. 4–12, ACM, 2014.

[168] G. Kontaxis, I. Polakis, S. Ioannidis, and E. P. Markatos, “Detecting social network profile cloning,” in *2011 IEEE international conference on pervasive computing and communications workshops (PERCOM Workshops)*, pp. 295–300, IEEE, 2011.

[169] J. Aro, “The cyberspace war: propaganda and trolling as warfare tools,” *European View*, vol. 15, no. 1, pp. 121–132, 2016.

[170] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, “Anyone can become a troll: Causes of trolling behavior in online discussions,” in *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pp. 1217–1230, ACM, 2017.

[171] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, “Hate speech detection with comment embeddings,” in *Proceedings of the 24th international conference on world wide web*, pp. 29–30, ACM, 2015.

[172] S. Sood, J. Antin, and E. Churchill, “Profanity use in online communities,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1481–1490, ACM, 2012.

[173] V. K. Singh, M. L. Radford, Q. Huang, and S. Furrer, “They basically like destroyed the school one day: On newer app features and cyberbullying in schools,” in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 1210–1216, ACM, 2017.

[174] G. Gini and D. L. Espelage, “Peer victimization, cyberbullying, and suicide risk in children and adolescents,” *Jama*, vol. 312, no. 5, pp. 545–546, 2014.

[175] Z. Waseem and D. Hovy, “Hateful symbols or hateful people? predictive features for hate speech detection on twitter,” in *Proceedings of the NAACL student research workshop*, pp. 88–93, 2016.

[176] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, “Common sense reasoning for detection, prevention, and mitigation of cyberbullying,” *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 2, no. 3, p. 18, 2012.

[177] P. Vijayaraghavan, H. Larochelle, and D. Roy, “Interpretable multi-modal hate speech detection,”

[178] A.-M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, “Large scale crowdsourcing and characterization of twitter abusive behavior,” in *11th International Conference on Web and Social Media, ICWSM 2018*, AAAI Press, 2018.

[179] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated hate speech detection and the problem of offensive language,” *arXiv preprint arXiv:1703.04009*, 2017.

[180] J. H. Park and P. Fung, “One-step and two-step classification for abusive language detection on twitter,” *arXiv preprint arXiv:1706.01206*, 2017.

[181] J. Golbeck, Z. Ashktorab, R. O. Banjo, A. Berlinger, S. Bhagwan, C. Buntain, P. Cheakalos, A. A. Geller, Q. Gergory, R. K. Gnanasekaran, *et al.*, "A large labeled corpus for online harassment research," in *Proceedings of the 2017 ACM on Web Science Conference*, pp. 229–233, ACM, 2017.

[182] B. G. Bara, F. M. Bosco, and M. Bucciarelli, "Developmental pragmatics in normal and abnormal children," *Brain and language*, vol. 68, no. 3, pp. 507–528, 1999.

[183] P. De Villiers, "Assessing pragmatic skills in elicited production," in *Seminars in Speech and Language*, vol. 25, pp. 57–71, Copyright© 2004 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New , 2004.

[184] S. Loukusa, E. Leinonen, and N. Ryder, "Development of pragmatic language comprehension in finnish-speaking children," *First Language*, vol. 27, no. 3, pp. 279–296, 2007.

[185] T. M. Gallagher, "Language skill and the development of social competence in school-age children," *Language, Speech, and Hearing Services in Schools*, vol. 24, no. 4, pp. 199–205, 1993.

[186] J. W. Astington and J. M. Jenkins, "A longitudinal study of the relation between language and theory-of-mind development.," *Developmental psychology*, vol. 35, no. 5, p. 1311, 1999.

[187] C. F. Norbury, "Factors supporting idiom comprehension in children with communication disorders," *Journal of Speech, Language, and Hearing Research*, 2004.

[188] C. F. Norbury, "Practitioner review: Social (pragmatic) communication disorder conceptualization, evidence and clinical implications," *Journal of Child Psychology and Psychiatry*, vol. 55, no. 3, pp. 204–216, 2014.

[189] F. G. Happé, "Communicative competence and theory of mind in autism: A test of relevance theory," *Cognition*, vol. 48, no. 2, pp. 101–119, 1993.

[190] F. M. Bosco, M. Tirassa, and I. Gabbatore, "Why pragmatics and theory of mind do not (completely) overlap," *Frontiers in psychology*, vol. 9, p. 1453, 2018.

[191] K. R. Wentzel, "Student motivation in middle school: The role of perceived pedagogical caring.," *Journal of educational psychology*, vol. 89, no. 3, p. 411, 1997.

[192] W. Levinson, R. Gorawara-Bhat, and J. Lamb, "A study of patient clues and physician responses in primary care and surgical settings," *Jama*, vol. 284, no. 8, pp. 1021–1027, 2000.

[193] T. Bickmore and J. Cassell, "Relational agents: a model and implementation of building user trust," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 396–403, ACM, 2001.

[194] S. S. Kim, S. Kaplowitz, and M. V. Johnston, "The effects of physician empathy on patient satisfaction and compliance," *Evaluation & the health professions*, vol. 27, no. 3, pp. 237–251, 2004.

[195] J. Fraser, I. Papaioannou, and O. Lemon, "Spoken conversational ai in video games: Emotional dialogue management increases user engagement.," in *IVA*, pp. 179–184, 2018.

[196] C. Chakrabarti and G. F. Luger, "Artificial conversations for customer service chatter bots: Architecture, algorithms, and evaluation metrics," *Expert Systems with Applications*, vol. 42, no. 20, pp. 6878–6897, 2015.

[197] L. Cui, S. Huang, F. Wei, C. Tan, C. Duan, and M. Zhou, "Superagent: A customer service chatbot for e-commerce websites," in *Proceedings of ACL 2017, System Demonstrations*, pp. 97–102, 2017.

[198] A. Xu, Z. Liu, Y. Guo, V. Sinha, and R. Akkiraju, "A new chatbot for customer service on social media," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 3506–3510, ACM, 2017.

[199] R. Socher, C. D. Manning, and A. Y. Ng, "Learning continuous phrase representations and syntactic parsing with recursive neural networks,"

[200] D. Chen and C. Manning, "A fast and accurate dependency parser using neural networks," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 740–750, 2014.

[201] J. Berant and P. Liang, "Semantic parsing via paraphrasing," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1415–1425, 2014.

[202] P. Pasupat and P. Liang, "Compositional semantic parsing on semi-structured tables," *arXiv preprint arXiv:1508.00305*, 2015.

[203] B. Jacquet, O. Masson, F. Jamet, and J. Baratgin, "On the lack of pragmatic processing in artificial conversational agents," in *International Conference on Human Systems Engineering and Design: Future Trends and Applications*, pp. 394–399, Springer, 2018.

[204] H. P. Grice, "Logic and conversation, syntax and semantics," *Speech Acts*, vol. 3, pp. 41–58, 1975.

[205] D. Sperber and D. Wilson, *Relevance: Communication and cognition*, vol. 142. Harvard University Press Cambridge, MA, 1986.

[206] A. Ritter, C. Cherry, and B. Dolan, "Unsupervised modeling of twitter conversations," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 172–180, Association for Computational Linguistics, 2010.

[207] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing dialogue agents: I have a dog, do you have pets too?," *arXiv preprint arXiv:1801.07243*, 2018.

[208] P.-E. Mazaré, S. Humeau, M. Raison, and A. Bordes, "Training millions of personalized dialogue agents," *arXiv preprint arXiv:1809.01984*, 2018.

[209] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[210] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," *arXiv preprint arXiv:1901.11504*, 2019.

[211] J. Weizenbaum *et al.*, "Eliza—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[212] R. Wallace, "The elements of aiml style," *Alice AI Foundation*, vol. 139, 2003.

[213] H.-Y. Shum, X.-d. He, and D. Li, "From eliza to xiaoice: challenges and opportunities with social chatbots," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 10–26, 2018.

[214] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," *arXiv preprint arXiv:1811.00207*, 2018.