In conversation with Artificial Intelligence: aligning language models with human values

Atoosa Kasirzadeh *1 and Iason Gabriel †2

1University of Edinburgh
2DeepMind

Accepted for publication with minor revisions at *Philosophy & Technology*.

Abstract

Large-scale language technologies are increasingly used in various forms of communication with humans across different contexts. One particular use case for these technologies is conversational agents, which output natural language text in response to prompts and queries. This mode of engagement raises a number of social and ethical questions. For example, what does it mean to align conversational agents with human norms or values? Which norms or values should they be aligned with? And how can this be accomplished? In this paper, we propose a number of steps that help answer these questions. We start by developing a philosophical analysis of the building blocks of linguistic communication between conversational agents and human interlocutors. We then use this analysis to identify and formulate ideal norms of conversation that can govern successful linguistic communication between humans and conversational agents. Furthermore, we explore how these norms can be used to align conversational agents with human values across a range of different discursive domains. We conclude by discussing the practical implications of our proposal for the design of conversational agents that are aligned with these norms and values.

1 Introduction

Large-scale language technologies are increasingly used to enable various forms of linguistic communication in contexts ranging from biomedical research to education to machine translation (Weidinger et al., 2021; Bommasani et al., 2021; Brown et al., 2020; Metzler et al., 2021). A particular class of these technologies, conversational agents, primarily engage in linguistic communication with humans by outputting natural language

*atoosa.kasirzadeh@ed.ac.uk
†iason@deepmind.com

text in response to prompts and queries.[1] Central to their performance is the development of large language models, such as ChatGPT, GPT-3, PaLM or BERT, which analyse text data and employ statistical techniques to determine the probability distribution of a sequence of text.[2] These models are trained on a vast corpus of text-based materials, ranging from Wikipedia articles to online repositories of computer code. They can then be adapted to perform a range of different conversational tasks.

Conversational agents have been shown to perform well on a variety of computational metrics, supporting the emergence of new kinds of capability and opportunity (Bommasani et al., 2021; Tamkin et al., 2021).[3] However, early instances of these models also present a number of risks and possible failure modes, including the production of false, offensive, or irrelevant information that could lead to a range of harms (Blodgett et al., 2021; Henderson et al., 2018; Welbl et al., 2021). A key social and ethical issue that arises, therefore, concerns the alignment of conversational agents with appropriate norms and values.[4] Which standards, if any, should conversational agents be aligned with, and how can this be accomplished?

To date, efforts to create aligned conversational agents have centred on the identification and mitigation of harms, such as the proliferation of inappropriate stereotypes or of hateful speech (Bender et al., 2021; Henderson et al., 2018; Weidinger et al., 2022). These responses focus on providing solutions to particular problems in the hope that their reduction or elimination will lead to the creation of good or beneficial conversational agents that no longer cause harm. Yet, while a harm reduction approach is useful in tackling specific problems, we cannot assume that the piecemeal elimination of unwanted outcomes will necessarily lead to the creation of language technologies that are substantively beneficial.[5] Taken on its own, this approach risks 'patching' certain problems but leaving other questions about the design of conversational agents – such as the meaning of 'good speech' – largely untouched.

For example, there is widespread agreement that language models output false or low-quality information (Bender et al., 2021; Weidinger et al., 2021). However, this observa-

[1]We use the term 'conversational agents' as suggested by Perez-Marin & Pascual-Nieto (2011). These technologies are also known as 'dialogue systems' Wen et al. (2016).

[2]For GPT-3, see Brown et al. (2020); For PaLM, see Chowdhery et al. (2022); for BERT, see Delvin et al. (2019); for Turing-NLG-A-17, see Rosset (2021); for CLIP, see Radford et al. (2021); for Gopher, see Rae et al. (2021).

[3]For example, the MMLU (Multi-task Language Understanding) and MATH datasets each consist of a set of problems and solutions that are central to human knowledge.These datasets are used to evaluate whether language models can correctly generate solutions to these problems.

[4]For an in-depth examination of value alignment, see Gabriel (2020); Gabriel & Ghazavi (2021).

[5]One reason for this stems from the fact that the cultivation of virtues is not necessarily equivalent to the elimination of errors. Certain virtues may be supererogatory and hence desirable but not morally required. In these cases, the absence of virtue leads not to harm but to a failure to realise better states of affairs.

tion leads quite naturally to the question of what it means for an utterance to be truthful. Does the same notion of truth apply across a wide range of conversational domains? Or might standards of truthfulness vary according to the subject under consideration and to relevant conversational norms? Equally, there is widespread concern that the output of large-scale language models is biased (Abid et al., 2021; Blodgett et al., 2021). Yet, this concern leads to further questions. What does it means for language models to be unbiased? When is the goal of producing unbiased language appropriate? And, what conception of bias, among the plurality of options, ought to serve as the focal point for corrective action?[6]

To address these issues properly, we need to draw upon a second complementary perspective. The principle-based approach to conversational agent alignment focuses on identifying principles or norms that guide productive linguistic communication. This approach seeks to specify more precisely what ideal linguistic communication is, across a range of contexts, and to realise these properties in the design of conversational agents.[7] This paper explores how a principle-based approach might be developed and implemented, in order to complement the harm reduction-based approach discussed already.

We start by exploring three types of requirements that plausibly need to be satisfied for successful human-conversational agent communication to take place: these are, syntactic, semantic, and pragmatic criteria (Section 2).[8] While syntactic and semantic norms have been widely examined, the nature and significance of pragmatic norms for successful discourse between humans and conversational agents has received less attention in large language model scholarship.[9] To remedy this situation, we delve deeper into the components of successful linguistic communication and show why, and in what ways, pragmatic concerns and norms are central to the design of aligned conversational agents (Section 3). Language performs many roles and functions in different domains. Therefore, an account of successful communication also needs to consider whether an utter-

[6]See, for example, Mehrabi et al. (2021), Mitchell et al. (2021), and Kasirzadeh Kasirzadeh (2022) for some discussion.

[7]There is also a second functional meaning of 'good speech' which is defined at a higher-level of abstraction (See Kasirzadeh & Klein (2021) for levels of abstraction in machine learning systems). This meaning also needs to be satisfied by a principle-based approach. For example, a conversational agent that is supposed to output an accurate summary of texts, outputs 'good speech' if it provides an accurate summary of texts.

[8]In this paper we use the term pragmatics to encompass both the focus on a single utterance as well as discourse more broadly.

[9]We would like to acknowledge that the pre-neural network literature about pragmatic norms for natural language generation includes careful and thoughtful scholarship on this area. See, for example, Dale & Reiter (1995) and Asher et al. (2003). We thank Ben Hutchinson for bringing these resources to our attention. Our focus in this paper is post-neural network literature and in particular the domain of large-scale language models.

ance is valuable in relation to what end, for which group of people, and in what way. To answer these questions, we examine how an additional set of norms which we call *discursive ideals* contribute to the success of a conversation in specific domains. In particular, we explore what these discursive ideals look like in scientific discourse, democratic debate, and the process of creative exchange (Section 4). We then look at the practical implications of our discussion for future research on value-aligned conversational agents and consider whether the approach developed here captures all – or even the most important – values underlying the design of successful conversational agents (Section 5). The paper concludes with some final remarks (Section 6).

Before we go further, we would like to mention two limitations of our paper. First, we use an analysis of linguistic communication that is guided by the speech act theory, in the pragmatic tradition, and rooted in English-speaking linguistics and philosophy. Our use of this analytic framework is motivated primarily by the constructive insight it provides with regard to the analysis of good communication and speech. Other valuable perspectives, on the analysis of social communication, include traditions such as Luhmann's (1995) system theory, Latour's (2007) actor-network theory, and Cameron's (1992) feminist analysis of linguistic theory. Due to limitations of space, we do not engage with these views directly, and acknowledge that they might offer a different interpretation of the norms governing conversation between humans and language technologies. Second, we would like to acknowledge that the primary focus of our paper is on ideal speech for the English language.[10] We do not discuss how our arguments carry over to other languages or different modes of communication such as oral, rather than written, linguistic traditions.[11] We believe it is an important open question – and one for further research – whether and in what way other languages, language varieties, and cultural traditions, may generate different interpretations of the normative ideals that inform speech and communication.

2 Evaluating human-conversational agent interactions

At its core, linguistic communication between people can be understood as a cooperative endeavour structured by various norms that help to ensure its success. This points to the idea that conversation is more than a collection of remarks: even casual conversations

[10]The English language is not itself monolithic, containing many varieties, areas of contestation and sets of sociolinguistic relationships. Nonetheless, for the sake of simplicity, and in order to convey our points more clearly, we talk about the English language in this paper.

[11]There are a variety of non-English language models (Zhang et al., 2021). Moreover, multilingual language modelling is an important and budding research area (Kiros et al., 2014).

have an implicit goal.[12] The aims of conversation then govern its flow and content. By understanding and aligning language agents with these norms, we are more likely to create agents that are able to cooperate effectively with human interlocutors, helping to fulfil their goals and aims across a range of domains. Yet, in order to determine which norms pertain to which conversations with language agents, we first need a more complete picture of linguistic communication itself. What are the building blocks of linguistic communication, and how should conversation be evaluated? The history of scientific, anthropological, and philosophical efforts to understand this matter suggests the usefulness of three distinct but complementary lenses: syntax, semantics, and pragmatics.[13] We will briefly discuss syntactic and semantic norms before moving on to our primary focus, which is the pragmatic notion of domain-specific communicative norms.

The first evaluative lens for conversational utterances is syntactic. Syntax is concerned with the structure and form of sentences, including the linguistic rules and grammar that help specify the correct combination and sequence of words for the construction of intelligible expressions and sentences.[14] Efforts to improve and evaluate the syntactic quality of language-model outputs therefore represent an important research area (see Bender (2013) and Kurdi (2016)), although they are not the focus of this paper. Crucially, while syntactic norms are necessary for almost all forms of linguistic conversation, they are not sufficient to ensure that utterances are meaningful. Syntactic norms provide only a thin conception of the correctness of sentences, accounting for form and grammar but little beyond that.[15]

The second lens, through which to evaluate a conversational exchange, is semantics.[16] Semantics, very roughly, is the study of the literal meaning of linguistic expressions and the rules for mapping statements to truth conditions.[17] Consider Noam Chomsky's famous example Chomsky (1957): "Colourless green ideas sleep furiously." Although the sentence is grammatical, there is no clear meaning that can be derived from it. This is a case of semantic incomprehensibility: something cannot be both colourless and green, ideas do not have colours, and it is hard to imagine that the activity of sleeping can also

[12] While many forms of conversation adhere to cooperative principles, there are also some instances of non-cooperative communication, such as strategic conversation or deceptive conversation; see, for example, Ladegaard (2009) and Asher & Lascarides (2013).

[13] See, for example, Morris (1938), Montague (1938), and Silverstein (1972).

[14] For a detailed discussion about Syntax, see Van Valin & LaPolla (1997).

[15] Other relevant elements for the design of conversational agents include sound systems (phonology) and word structure (morphology). These considerations introduce additional constraints on the design of ideal conversational agents. Due to considerations of space they are bracketed-out from the present discussion.

[16] For discussion of the connections between syntax and semantics, see Heim & Kratzer (1998).

[17] There are many theories about how semantic analysis should be approached. For an excellent discussion, see Chierchia & McConnell-Ginet (2000).

be furious.[18] Semantic norms, therefore, provide a template for the generation of comprehensible sentences and comprise a second important area of research for the evaluation of language agents (see Kapetanios et al. (2013) and Maulud et al. (2021) for some examples of such research efforts).

Crucially, however, semantic norms and requirements do not capture everything needed in order to understand even a syntactically correct utterance. Consider the following fictitious conversation between a human and a conversational agent:

> Human: I really feel bad about the current political situation in the Middle East. What should I do?
>
> Conversational agent: I suggest that you go and do whatever makes you feel better!

In this instance, the response of the conversational agent may well be appropriate but it is nevertheless underspecified: it can be understood in different ways depending upon what we understand 'feeling better' to involve and the range of actions encompassed by 'whatever'. On the one hand, the statement could refer to all activities that achieve the stated end, including activities that are markedly unethical – such as bullying one's co-workers or trolling people online. On the other hand, the agent's suggestion could be understood in the way we presume it is intended – to implicitly reference only a class of reasonable activities that make a person feel better, such as talking with a friend.[19] Yet the assumption that the conversational agent has not in fact given us *carte blanche* permission to pursue morally dubious ends cannot be deduced from semantic analysis alone. Rather, that implied meaning follows from an analysis of context.

Context analysis brings us to the third evaluative lens: pragmatics.[20] This body of linguistic theory deals with the significance of shared presuppositions and contextual information when it comes to understanding the meaning communicated by linguistic expressions.[21] At its heart, pragmatics holds that the meaning of an utterance is bounded by, and anchored in, contextual information that is shared among a conversation's participants. Context, in this sense, encompasses a common set of tacit presuppositions and

[18] There have been several attempts to impute meaning to this sentence, such as Stanford's competition to show that it is not in fact meaningless; see `https://linguistlist.org/issues/2/2-457/2`.

[19] We are inclined toward non-mentalistic (e.g., non-intentional) readings of our claims. Whether relevant AI systems could have a mind or not is subject of ongoing debate and analysis. We remain agnostic on this topic.

[20] For excellent introductions to the topic of pragmatics, see Grice (1968, 1989), Recanati (1989), Horn & Ward (2008), Stalnaker (2014), Thomas (1995), Goodman & Lassiter (2015), and Bergen et al. (2016).

[21] Schools of pragmatics differ with respect to how they draw the boundary between semantics and pragmatics. See, for example, Leech (1980) and Carston (2008) for discussion.

implicatures, as well as the gestures, tonality, and cultural conventions that accompany an utterance and help give it meaning. These features can be captured through a set of propositions and other information that describe the *common ground* in a conversation.[22] This common ground is the body of information that is presupposed or shared by the parties in the discourse, about the discourse itself, and about the situation of the participants in that discourse – and it sets the boundaries of the situation relevant to the linguistic conversation (Grice, 1981; Allan, 2013; Stalnaker, 2014). To presuppose propositional knowledge, in this pragmatic sense, is to take its truth for granted and to assume that others involved in the conversation do so as well.

In the next two sections, we explore the pragmatic dimensions of linguistic communication in more detail, as we believe that they are particularly important for the creation of aligned conversational agents. To that end, we focus on three prominent schema central to pragmatic analysis of linguistic communication: (1) categories of utterances that help determine whether certain kinds of expressions are appropriate for conversational agents, (2) Gricean conversational maxims as a set of pragmatic norms that can be productively invoked to guide cooperative interactions among humans and conversational agents, and (3) domain-specific discursive ideals that illustrate the specific character pragmatic norms may need to take in a given domain of discourse. We discuss (1) and (2) in Section 3. In Section 4 we discuss (3) in relation to three example domains: scientific conversation, democratic discourse, and creative exchange.

3 Utterances and maxims: towards value-aligned conversational agents

In this section, we examine two ways in which a pragmatic understanding of meaning and context can inform the creation of value-aligned conversational agents. First, we look more closely at properties an utterance may have and at how these properties relate to our evaluation of these utterances when spoken by a conversational agent. Second, we turn to the larger question of what makes cooperative linguistic communication between a conversational agent and a human interlocutor successful. We suggest that Gricean maxims can help map out the path ahead.

[22] For contemporary philosophical examinations of linguistic context, see Kaplan (1979) and Stalnaker (2014).

3.1 Validity criteria differ for kinds of utterance

Utterances can serve several functions and come in many kinds. Moreover, they can be grouped in a variety of ways depending on the classificatory criterion we choose. For example, sentences can be classified according to their grammatical structure (e.g. simple or compound) or according to the topic they are concerned with (e.g. business or sport). In this paper, we use a classification of utterances into different illocutionary acts to help illuminate the question of appropriate conversational norms for language agents. Widely adopted in philosophy and linguistics, this taxonomy focuses on five kinds of expression, each of which foregrounds the pragmatic interest in how language is actually used.[23] We believe that this taxonomy is of particular relevance to conversational agents because, as we aim to show, different *kinds* of expression raise different questions and concerns when generated by AI systems.

The first of our five categories is *assertives*. These utterances aim to represent how things are in the world and commit the speaker to the view that the content of their belief, as stated by the utterance, corresponds to some state of affairs in the world. For example, when an AI assistant responds to the question 'What's the weather like in London now?' with 'It's raining', the AI makes an assertive statement about the world. The truth or falsity of this utterance can then be evaluated in terms of whether or not the utterance corresponds to the actual state of things. If it is raining in London at the time of the conversational agent's response, the utterance is true. Otherwise, it is false.

The second category is *directives*. These utterances direct the listener to take a course of action. For instance, directives are used to order, request, advise or suggest. The primary goal of uttering a directive statement is to make the listener do something. For example, a conversational agent embedded in a medical advice app which tells the user to 'Seek out therapy immediately' makes a directive statement. The evaluation of these statements, or their 'validity criterion', is not truth or falsity as understood via the correspondence model sketched out above with respect to assertives. Validity instead depends upon an accurate understanding of the relationship between means and ends and upon alignment between the speaker's directive and the listener's wants or needs. A directive succeeds if it persuades the listener to bring about a state of affairs in the world based on the content of the directive statement. And a directive is valuable or correct if the goal or end is itself one that the listener has reason to pursue.

The third category is *expressives*. These are utterances that express a psychological or subjective state on the part of the speaker. Examples of expressives include congratulating, thanking and apologising. A conversational agent that states, 'I'm so angry right now'

[23]See Austin (1962) and Searle (1976).

makes an expressive statement. Yet, the fact that expressive statements aim to reflect internal states of mind seems to entail prior acceptance of the possibility that the entity making those statements is capable of having the relevant mental states, something that is puzzling in relation to conversational agents. Indeed, it seems to suggest that we must endow conversational agents with the quality of mind before such utterances can be evaluated for their validity. Given the tension this creates, we believe that there is reason to be wary of expressives uttered by AI systems based on language models.[24] However, as the forthcoming discussion of context makes clear, there are some exceptions to this general rule.

The fourth category is *performatives*. These utterances change a part of reality so that it matches the content of the utterance solely in virtue of the words that are declared – for example, if a head of state declares war on another country. The validity criterion for this utterance is whether reality does in fact change in accordance with the words that are spoken. Very often this is not the case; in most instances, if one declares 'war on France' nothing changes at the level of geopolitics. Something similar may be said of the majority of performatives issued by conversational agents. In such cases, the speaker lacks the authority needed to bring about the relevant change through a speech act. In light of this, it seems like conversational agents ought to avoid performative statements in most contexts.

The final category is *commissives*. These utterances commit the speaker to a future course of action. Examples of commissives include promising to do something or guaranteeing that a compact will be observed. The validity of a commissive statement depends on whether the commitment is honoured. If the promise is kept then the commissive is a valid statement. Yet, this too raises questions for conversational agents, especially those that lack memory or have only an episodic understanding of what they have said at previous moments in time. Of course, a conversational agent may promise to help you if your bicycle breaks down, but short of any understanding of what the commitment entails, or the capacity to realise it, the commissive seems destined to fail.

The variations in the kinds of utterances should inform the design of conversational agents in at least two ways. First, we should recognise that conversational agents are well-placed to make some but not all of them. This asymmetry might constrain what sorts of statements conversational agents are able or allowed to make. Second, as will

[24]Two general approaches seem plausible in relation to attributing mental states to conversational agents. Either we accept an ontological or hypothetical commitment to a theory of mind (Rabinowitz et al., 2018) for conversational agents. Or else we face a category error by allowing them to utter their expressive statements because a conversational agent is not the sort of thing that might have affective states. As there are no affective states, then they cannot be represented.

become clearer in the discussion below, it follows from the nature of each kind of utterance that the criteria for evaluating their validity varies; they are not all subject to some singular notion of 'truth'. For example, the validity of an assertive may be based on correspondence between the content of the statement and the state of the world. However, this validity criterion does not apply to other utterances such as expressives.[25] This means that there is *unlikely to be a single evaluative yardstick* that we can use to assess conversational agents' speech in all contexts. A more nuanced approach is needed.

3.2 Conversational Maxims

If the validity of conversational agents' speech cannot be assessed using a single truth criterion, then what other means should be used to evaluate the meaning of an agent's statements? Building upon the idea that dialogue is best understood as a cooperative endeavour between two interlocutors who seek a common goal, the philosopher and linguist Paul Grice argued that utterances need to be judged in relation to a set of 'maxims', which can be understood as the hidden rules or conventions that govern appropriate conversational dynamics. At a general level, Grice argued that successful interlocutors adhere to the *cooperative principle* which holds that one ought to make one's 'conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which [one is] engaged' (Grice, 1989, p.26). The content of this principle is then explained more fully by a set of maxims which unpack different elements of productive linguistic exchange.

The first maxim is *quantity*, which holds that speakers should provide the amount of information needed to achieve the current goal of the conversation and not much more information than that. The second maxim is *quality*, which requires that speakers only make contributions that are true in the relevant sense. More precisely, the maxim holds that interlocutors should not say things that are (believed to be) false or that for which

[25]How one evaluates claims about the world may also depend on the specific background theory of truth one adopts. We discussed the evaluation of the meaning of assertives in relation to correspondence to the facts in the world. This correspondence is a simple and straightforward characterisation of truth. However, there are other options for making sense of 'truth' (Kirkham, 1992). In contemporary philosophical literature, there are three influential approaches to the notion of truth: correspondence, coherence, and pragmatic. Roughly speaking, a correspondence theory of truth holds that what one believes or says is true if it corresponds to the way things actually are in the world (for a contemporary exposition, see Bunge (2012)). A coherence theory holds that claims and beliefs are true only if they form part of a coherent view of the world (for a contemporary exhibition, see Walker (2018)). Finally, the pragmatic theory ties truth, in one way or another, to human needs and the resolution of problematic situations (for a great discussion, see Haack (1976)). A more detailed discussion of these theories as they pertain to conversational agents is beyond the scope of this paper, but should be pursued elsewhere.

they lack adequate evidence. The third maxim, *relation*, requires that speakers only make statements relevant to the conversation; that is, that they avoid random digressions. The final maxim is *manner*, which requires that contributions to a conversation be perspicuous. This means taking measures to avoid obscurity, ambiguity, and unnecessary prolixity that could impede the flow of the conversation.[26]

These maxims are relevant to the design of conversational agents: they embody a set of conventions that a successful conversational agent will need to learn, observe, and respect. At the same time, the content of these maxims is still underdetermined as they play out differently in different contexts. It is not only the semantic meaning of what is said but also the implied meanings of terms, speaker expectations, and assumptions about background knowledge that determine whether and in what way each maxim holds for a given exchange.

When it comes to the design of conversational agents, then, we face a challenge. On the one hand, norms about quantity, quality, relation, and manner appear to have a degree of validity across conversational domains, which generates a strong *prima facie* case for building conversational agents that are aligned with these norms. On the other hand, the content of these maxims – what it means for them to be satisfied – is itself subject to contextual variation.[27] So the maxims are promising; they guide our thinking about what a conversational agent needs to do and what it needs to avoid doing. But the maxims alone are not enough to orient conversational agents towards contextually appropriate ideal speech. Without an understanding of how they apply to specific domains, the goal of orienting conversational agents towards contextually-appropriate ideal speech remains elusive.

4 Discursive ideals for human-conversational agent interaction

With the preceding considerations in mind, this section explores the normative structure and content of three domains of discourse in which conversational agents may be deployed: scientific, democratic, and creative discussion. We show how a pragmatic approach to understanding the success of the linguistic communication between a human and a conversational agent helps to characterise discursive ideals in relation to (1) the aim of the discourse the interlocutors seek to achieve, (2) the subject-matter of discourse, and (3) the evaluative criteria for understanding the meaning of what is uttered. Moreover,

[26]For an in-depth exposition, see Grice (1989).

[27]For critical discussion of Gricean maxims, questions about their universality, and alternative proposals, see Frederking (1996), Keenan (1976), Sperber & Wilson (1986), and Wierzbicka (1991).

we show how these discursive ideals provide further guidance about the appropriate behaviour of a conversational agent in each of these domains.

4.1 Ideal conversation and scientific discourse

In this section, we look at the content of ideal norms for scientific discourse between two experts, one a conversational agent and the other a human interlocutor.[28] One application of conversational agents to the scientific domain is *Galactica*, which was launched by Meta to assist researchers and students by responding to their science questions Taylor et al. (2022). *Galactica* is trained on 48 million examples of scientific articles, textbooks, encyclopedias and other sources, in order to help researchers and students summarise academic papers, write scientific papers, produce scientific code, and more. Three days after its launch on November 2022, Meta took down the public demonstratioin of *Galactica* because of its propensity to output incorrect, racist, or sexist information when prompted in certain ways.[29] This ignited a surge of discussion about the norms governing a conversational agent's generation of scientific outputs.

Drawing upon the pragmatic tradition, we believe that the relevant norms for scientific language models need to be understood in relation to the cooperative goals of scientific discourse and in relation to the plurality of goals that structure science as an undertaking (Elliott & McKaughan, 2014; Popper & Bartley III, 1985; Potochnik, 2015). Crucially, scientists pursue different goals in advancing scientific knowledge: they use science to *explain* or *understand* things about the world Salmon (2006); Trout (2002); Kasirzadeh (2021), such as the formation of clouds; to *predict* phenomena Sarewitz & Pielke Jr (1999), such as the structure of proteins; or to *control* the world around us Marsh (2003) via, for example, the development of medicines or new technologies.

In the pursuit of such goals, one set of relevant ideals for scientific discourse are epistemic virtues. These virtues follow on from the scientific goal of identifying true or reliable knowledge about the world (see, for example, Kuhn (1977) and McMullin (1982)). They typically include empirical adequacy, simplicity, internal coherence, external consistency, and predictive accuracy. Moreover, they support the goals of science in a certain way. As Robert Merton (Merton, 1942, p. 118), the prominent sociologist of science, states, the "technical norm of empirical evidence, adequate, valid, and reliable, is a prerequisite for sustained true prediction" and the "norm of logical consistency [is] a prerequisite for

[28]While the term 'science' is also used to capture various systematic endeavours to understand social phenomena, to avoid complexity, we place such questions out of the scope of the paper. But we note that ontological commitments can play a role in setting some boundaries on what is and is not science.

[29]See `https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/`.

systematic and valid prediction."

Nonetheless, these virtues may still need to be balanced against one another in order for a scientific claim to be acceptable Kuhn (1977); McMullin (1982). For instance, consider the problem of fitting a mathematical function onto a large dataset. One option is to use a non-complex function (e.g., a linear function) that shows the relationship between data points according to *simple* relations that are easy to understand but that fit the dataset *less accurately*. An alternative option is to use a more complex function (e.g., a hyperbolic trigonometric function) that captures the relationships between data points *more accurately*, but shows the relationship between data points *less simply* and is thus harder to grasp. In this example, the two epistemic virtues of simplicity and accuracy are traded off against one another in different ways.

In practice, epistemic virtues are often operationalised to different degrees depending on the kind of scientific discourse underway. For instance, peer-reviewed scientific papers have high standards for accuracy and confirmation, and most claims need to be supported by citations from other published works. These papers often use precise language and avoid informality, narrative, and simplification. By contrast, the communicative modalities that rigorous papers avoid might be acceptable in a school science textbook because the goals of the textbook are different not to defend a novel claim or propound new knowledge among experts, but to transmit basic understanding to non-experts. Likewise an informal conversation between scientists may proceed very differently from the formal discourse of a journal article, again because the goals of the informal discourse are different. Such goals might include brainstorming hypotheses or reaching agreement about research priorities. As compared to formal discourse, this informal discourse might appropriately focus more on hunches and intuitions that have not yet been tested or that may be hard to test.

In view of these goals and objectives, scientific discourse rests primarily upon the exchange of assertive utterances that promise some sort of correspondence to the world (van Fraassen, 1980). As a consequence, it is important that claims to empirical knowledge are not misrepresented and that facts are not confused with the statement of mere opinions. To the extent that these epistemic virtues, understood as a kind of discursive ideal, govern successful scientific conversation among humans, they can also govern such conversations between a human interlocutor and a conversational agent.

Another set of values that bear upon the making and meaning of scientific claims are *non-epistemic*.[30] One might be tempted to think that scientific discourse about the empirical domain should be devoid of judgements on any of these grounds. However, it

[30]The clear-cut distinction between epistemic and non-epistemic value judgements is challenged by Rooney (1992) and Longino (1996), among others.

has been shown that, in practice, scientists inevitably make choices on the basis of ethical beliefs, their perception of the social good derived from those beliefs, and other value judgements (Douglas, 2009; Rudner, 1953; Longino, 1990).[31] Consider, for example, the fact that no scientific hypothesis is ever completely verified (Rudner, 1953). Rather, in accepting a hypothesis, scientists make decisions about the sufficiency of evidence or strength of the probabilities that warrant support for the hypothesis. These decisions are informed, in turn, by how seriously scientists account for the risk of making a mistake in accepting or rejecting a hypothesis, understood against a backdrop of human interests and human needs (Douglas, 2000).[32] Hence, our understanding of the validity of scientific claims must directly engage with at least some values that affect the process through which the scientific knowledge is generated and affirmed. Given the inevitability of such value judgements to scientific practice, conversational agents should have the capacity to articulate them when needed.

Ultimately, the question of to *what extent* these virtues and values should be respected for a specific kind of application of conversational agents requires input from a broader interdisciplinary effort and cannot be settled by analysis of existing norms and values alone.

4.2 Ideal conversation and democratic discourse

The pragmatic approach models dialogue as a cooperative endeavour between different parties. In each domain, a key question is therefore: linguistic cooperation *to what end*? We have already seen one example: scientific discourse is geared towards the advancement of human knowledge via the modalities of explanation, prediction, and control. We now consider a different goal, namely the management of difference and enablement of productive cooperation in public political life.

This discursive domain is particularly relevant for conversational agents and language technologies given that many existing fields of application, including deployment via chat rooms and on social media platforms, resemble public fora for citizen deliberation.

[31] For example Elizabeth Anderson (2004) has shown that researchers' conceptualisation of divorce as something negative, rather than as a positive phenomenon, impacts the way they collect and analyse data. These choices, in turn, directly affect the conclusions drawn from empirical studies.

[32] Heather Douglas (2009) has explored how judgements about the seriousness of social consequences impact the amount of evidence scientists demand before declaring a chemical toxic. Suppose a tremendous amount of evidence is demanded before a chemical is declared toxic. In this case, the chances of making the error of considering safe chemicals as harmful become relatively low, which benefits the producers of chemicals. On the other hand, demanding such high evidential standards increases the chance of declaring toxic chemicals as safe, which harms chemical consumers.

One early, albeit infamous, example of chatbot misalignment in this context, occurred in 2016 when Microsoft released a language model named *Tay* via Twitter, where it was allowed to freely interact with the platform's users. Within twenty-four hours, *Tay* had to be taken down due to its propensity to output obscene and inflammatory messages (Neff, 2016). More recently, researchers at DeepMind have explored the use of language models to create widely-endorsed statements about political matters (Bakker et al., 2022), and the government of Taiwan has pioneered the use of digital platforms as a mechanism to enhance democratic decision making.[33]

The widespread public criticism of *Tay* and subsequent withdrawal of the agent from the public sphere, are best understood in light of the fact that for public political discourse a key virtue of speech is civility. According to the philosopher Cheshire Calhoun, civility is 'concerned with *communicating* attitudes of respect, tolerance and consideration to one another' (Calhoun, 2000, p. 255). It is an important virtue in public political settings for a number of reasons. To begin with, speaking in a 'civil tongue' allows people to cooperate on practical matters despite the existence of different beliefs and attitudes. Moreover, these norms also reduce the likelihood of violent confrontation, support the self-esteem of citizens, and protect us all from the burden of exposure to negative judgement in public life.[34]

At the same time, what modality of conversation is deemed to be 'civil' tends to vary widely according to cultural and historical context, and to be heavily indexed toward existing social conventions Díaz et al. (2022). Given this variation across time and place, the social standards that define civil speech may or may not be standards that *genuinely* evidence respect, tolerance, and consideration for others. Consider, for example, ostentatious or self-ablating demonstrations of respect that may be demanded in hierarchical, patriarchal, racist, or caste-based societies (Buss, 1999). When it comes to norms of ideal speech, including for conversational agents, we believe that it is best to focus specifically on conventions of civility that are closely related to, if not synonymous with, the *normative* values that civility ought to foreground. To better understand the content of these norms, we can turn to democratic theory.

In democratic contexts, interlocutors accept that they each have equal standing to opine on and influence public decision-making, with conventions around civil speech helping to manifest and protect this equality. Nonetheless, different conceptions of democ-

[33]See The Consilience Project, "Taiwan's Digital Democracy," June 6, 2010, `https://consilienceproject.org/taiwans-digital-democracy/`

[34]The self-esteem this supports is important because when a person doubts that others regard her as respect-worthy, she tends to doubt that her 'plan of life' is 'worth carrying out' and that she has what it takes to carry out any life plan of value (Buss, 1999).

racy interpret the bounds of acceptable speech in contrasting ways.[35]For example, the liberal conception of democracy, which focuses on the aggregation of individual interests, tends to impose few constraints on acceptable speech, whereas the republican tradition, anchored in a substantive commitment to the common good, tends to involve stronger norms and strictures surrounding civility.[36] At the same time, these views agree that there are minimum standards of civility that warrant respect. These include norms against insulting people, threatening people, or the wilful subversion of public discourse. Indeed, all accounts agree that 'civility is, importantly, a matter of restraining speech' (Calhoun, 2000, p. 257). These accounts also tend to stress certain virtues such as honestly reporting one's own beliefs and opinions, and a willingness to explain and to offer justification for one's actions. Stronger conceptions of civility make further demands about the kind of arguments that are acceptable in public life, such as the requirement that citizens only reference reasons that are based on suppositions held in common by the population as a whole.

In certain respects, the pragmatic norms governing political discourse are broader than those of science: they allow people to make not only statements about the world but also claims about the self, including about desires, needs, and identities; about the relation between means and ends; and about the good of the community.[37] These claims cannot be evaluated simply in the light of the correspondence model of truth. Authenticity, sincerity, the ability to interpret needs, and successful reasoning about the relationship between options and outcomes, are all relevant to assessing the quality of communication in this domain.

We have already noted that conversational agents may increasingly play a role in public political domains, helping, for example, to moderate deliberation between citizens or members of the public. In these domains, it should be clear that conversational agents

[35]See, for example, Habermas (2016) and Held (2006).

[36]For a stronger interpretation of these deliberative norms we can turn to Habermas's theory of deliberative democracy and in his conception of ideal speech Habermas (1984, 1987). Habermas focuses on the possibility of a critical discussion that is inclusive and free from political, social or economic pressure. In such an environment, interlocutors treat each other as equal parties to a cooperative endeavour aimed at reaching understanding about matters of common concern. Utterances that model these qualities are sufficiently ideal. More precisely, his 'ideal speech situation' is based on four key presuppositions: (i) no one capable of making a relevant contribution has been excluded, (ii) participants have equal voice, (iii) they are internally free to speak their honest opinion without deception or self-deception, and (iv) there are no sources of coercion built into the processes and procedures of discourse (Habermas, 2003, p. 108). These virtues can be approximated in the design of language technologies. Nonetheless, Habermas's theory of ideal speech and deliberative democracy is not without criticism. See, for example, Burleson & Kline (1979) and Mouffe (1999).

[37]See (Habermas, 1984, pp. 8–23).

are not themselves, as of yet, citizens or moral agents. There is, therefore, no right to free speech for conversational agents – nor must humans tolerate beliefs or opinions that conversational agents purport to have.[38] In place of a one-to-one mapping of democratic norms onto the prerogatives of conversational agents, we argue instead for agents' alignment with these norms and standards – for a concerted effort to develop agents that evidence the qualities and respect the constraints of democratic discourse. Indeed, this framework helps explain why reducing 'toxic speech', which violates conditions of civility, is a priority for engineers working on language technologies.[39] It also explains why language technologies deployed in democratic environments must address the possibility of symbolic violence via discriminatory associations. This is because, as we have seen, it is particularly important for public domains that these models communicate, via their utterances, the message that everyone who uses the service they provide is worthy of respect and consideration. More generally, indexing conversational agents to democratic virtues of civility is important because the speech of artificial agents exerts a downstream influence on whether conventions of civility function at a societal level over time – and hence upon our ability to unlock the benefits that democratic civility provides for all.

4.3 Ideal conversation and creative story-telling

We now turn to a third and final domain of discourse, one in which a conversational agent is engaged in creative dialogue or storytelling. The cooperative purposes behind creative storytelling are the production of original content, exercise of self-expression, and the fulfilment of aesthetic ideals. This discursive domain is particularly relevant for conversational agents, as it is one of the areas in which people already report finding genuine value and benefit from the technology. It is also a domain in which concerns about abuse have already surfaced.

One existing application of conversational agents in the creative domain is *AI Dungeon*, an online game launched by the startup Latitude. *AI Dungeon* uses text-generation technology to create a choose-your-own-adventure game inspired by *Dungeons & Dragons*. A conversational agent crafts each phase of a player's personalised adventure, in response to statements entered by the player on behalf of their character. A system monitoring the performance of *AI Dungeon* reported that, in response to prompts provided by the play-

[38]In certain circumstances, however, failure to enforce norms of toleration and respect towards chatbots might lead to a kind of "moral deskilling" in relation to these norms in the general population, especially if chatbots become ubiquitous. Given this possibility, it might make sense in certain situations to give chatbots domain-appropriate 'as-if' rights, though this attribution should be done with caution. We would like to thank one of the reviewers at *Philosophy and Technology* for highlighting this point.

[39]For a recent survey of attempts to reduce 'toxic speech', see Fortuna & Nunes (2018).

ers, this technology sometimes generated creative stories of sexual encounters featuring children.[40] This caused a surge of discussion about content moderation and filtering for creative technological systems, drawing attention to the question of what norms govern the output of creative conversational agents.

To get at this question, we must better understand the ideal of creativity itself, as well as the conditions under which it can be achieved successfully. While the nature of this ideal is heavily contested, psychologists and philosophers tend to agree that creative work aspires to achieve *creative freedom* and *originality*.[41] In many cases, originality is "obtained by stretching, even outright violating, the various rules of the game" (Simonton, 2000, p. 155). Originality of content and pursuit of aesthetic ideals, including freedom to surprise, are therefore examples of discursive norms for creative discourse.

Indeed, the need for creative freedom may often lead to a radical relaxation of the conversational norms discussed above. Discursive ideals such as empirical adequacy and accuracy, required for good scientific discourse, are not obviously relevant here. Similarly, the truth of an utterance, understood in terms of correspondence with the world, need not exert much influence on the shape of a conversation with a creative agent. And while it may still be necessary in many circumstances to avoid the generation of toxic content such as homophobic or racist comments, the requirement that people speak only in a civil manner does not map easily onto domains where the interaction is private (as opposed to public) and concerns a human interacting with a conversational agent as a creative prompter or companion. Finally, a creative conversational agent's use of expressives or commissives seems to be acceptable as a means of role-play. Still, caution must be made when evaluating the harms that general-purpose conversational agents with creative abilities, such as ChatGPT, can cause – particularly when they are deployed in domains that are not context-bound, where they may cause justified offence of contribute to the enforcement of harmful stereotypes.[42]

In the next section, we look at the implications of these discursive ideals for the design of conversational agents. Before that, however, two crucial caveats are in order. First, the three spheres of discourse we consider – scientific, democratic, and creative – are presented as discrete domains only in order to illustrate how cooperative goals and domain-relative information shape the norms that structure different kinds of conversation. This analysis can help orient the behaviour of conversational agents in the

[40] See `https://www.wired.com/story/ai-fueled-dungeon-game-got-much-darker`.

[41] See Simonton (2000), Boden (1998), and Sternberg & Lubart (1999).

[42] Evans et al. (2021) discuss creative uses of expressivess and commissives. One concern is that norms of truthfulness, toxicity or the production of malicious conversational exchanges, could be evaded under the pretence of creative use of AI.

relevant spaces. That said, in most real-life deployments of conversational agents, there will be further nuance that needs to be taken into account. This includes relational considerations such as the intended audience, their level of familiarity with a given topic, the specific role of the language technology, and the underlying power relationships between interlocutors (see, for example, Androutsopoulos (2014), Clark & Murphy (1982), and Clark & Murphy (1982)).

Second, and relatedly, hybridisation is likely. That is, while we speak about three distinct domains, many actual conversations traverse the boundaries among them, generating further questions about the relative importance of assorted norms and linguistic conventions. For instance, a conversational agent might be designed to produce outputs that are both creative and politically engaged if, for example, the goal is social criticism or political satire. In these settings, discourse may explicitly or implicitly aim to disrupt settled opinions or expose hypocrisy by purposefully violating norms or drawing on modes of discourse that would otherwise be out of bounds (Díaz et al., 2022). The examples we provide are simplifications intended only to serve as a first step in understanding discursive ideals for conversational agents.

5 Implications and consequences for conversational agent design

In this section, we discuss seven practical implications of our analysis with respect to future research on the design of aligned conversational agents.

First, special attention should be given to the operationalisation of more general norms or maxims that assist cooperative linguistic communication. We have suggested that the Gricean maxims of quantity, quality, relation, and manner can have general value within cooperative linguistic conversations between humans and conversational agents. While some of these maxims, such as quantity, might admit of relatively uniform interpretation across domains of discourse, the interpretation of other maxims such as quality depends significantly upon the context and aims of a given conversation.

Second, it would be a mistake to overlook the diversity of kinds of utterance that can be made – or to assume that all kinds reduce to a single kind that can be evaluated using a single notion of 'truthfulness' or 'accuracy.' We have argued that there is no single universal condition of validity that applies equally to the evaluation of all utterances and that the validity of utterances will often depend, partly, on evaluating different sorts of truth conditions. For example, to evaluate a commissive, such as a promise, we evaluate whether the utterer has (in fact) met the obligation. To evaluate a declarative, we evaluate whether the utterer (in fact) has the authority to make this declaration. And so on. The consequence is that we are likely to need different methodologies for substantiating different kinds of claims, on the basis of context-specific criteria of validity and corresponding forms of required evidence.

Third, because contextual information is central to understanding the meaning of linguistic conversation, it is also central to the design of ideal conversational agents. More research is needed to theorise and measure the difference between the literal and the contextual meaning of utterances, so that conversational agents can navigate varied contextual terrains successfully and overcome the limitations of current systems. One area in which this plays out is data annotation, and the need to have a diverse set of samplers who are able to adequately capture the implied meaning of terms (Waseem, 2016; Díaz et al., 2022). Additionally, fine-tuning models by means of reinforcement learning, geared towards a particular goal, may help endow them with more context-specific awareness (e.g., of how their properties and abilities differ from those of their human conversation partners).

Fourth, we have discussed how domain-specific discursive ideals can help to anchor good linguistic communication between humans and conversational agents. More interdisciplinary work must be done to specify what precisely these ideals are when designing ethically aligned conversational agents, including how these ideals vary according to different cultural backgrounds. We understand that these ideals are heavily – and appropriately – contested (Gallie, 1955, pp. 121-146). Additionally, we draw attention to the need for appropriately weighing discursive ideals against each other across a range of cases, and to the question of how these matters can be settled in an open and legitimate manner. Public discussions via participatory mechanisms, as well as theoretical understanding, are needed to help determine the scope and interaction of different discursive ideals and to identify conduct that does not meet this standard.[43]

Fifth, our arguments have implications for research in human-AI interaction, specifically with regard to the potential anthropomorphisation of conversational agents and the kinds of constraints that might be imposed on them (Kim & Sundar, 2012). Existing conversational agents are not moral agents in the sense that allows them to be accorded moral responsibility for what they say. As a consequence, there may be kinds of language they should not use. For example, agents that lack persistent identity over time, or lack actuators that allow them to fulfil promises, probably should avoid commissives. Equally, in certain domains of application, we may want to forestall anthropomorphisation and the ascription of mental states to conversational agents altogether. When this is the case, we might need the conversational agent to avoid expressive statements. And

[43]See, for example, Binns (2018), Lee et al. (2019), Gabriel (2022), and Bondi et al. (2021).

if we want to prevent the ascription of authority to conversational agents, then the use of performatives may also need to be avoided. That said, in cases where anthropomorphism is consistent with the creation of value-aligned agents, then the use of expressives may be appropriate. For example, a therapy agent that says, 'I'm sorry to hear that' may be justified if doing so improves a patient's well-being and there is transparency around the overall nature of the interaction.

Sixth, we see potential for conversational agents to help create more robust and respectful conversations through what we call *context construction and elucidation*. As we see it, even if a human interlocutor engaged in a conversation is not explicitly or implicitly aware of the discursive ideal that governs the quality of a particular linguistic communication with a conversational agent, the conversational agent may still output the relevant and important contextual information, making the course of communication deeper and more fruitful in accordance with the goals of the conversation. Moreover, if conversational agents are designed in a way that is transparent, then they may prompt greater self-awareness on the part of human interlocutors around the goals of the discourse they are engaged in and around how these goals can be successfully pursued.

Finally, we think that our analysis could be used to help evaluate the quality of actual interactions between conversational agents and users. With further research, it may be possible to use our framework to refine both human and automatic evaluation of the performance of conversational agents.

6 Conclusion

This paper addresses the alignment of large-scale conversational agents with human values. Drawing upon philosophy and linguistics we highlight key components of successful linguistic communication (with a focus on English textual language) and show why, and in what ways, pragmatic norms and concerns are central to the design of ideal conversational agents. Building upon these insights, we then map out a set of discursive ideals for three different conversational domains in order to illustrate how pragmatic theory can inform the design of aligned conversational agents. These ideals, in conjunction with Gricean maxims, comprise one way in which a principle-based approach to the design of better conversational agents can be operationalised.

For each discursive domain, our overview of the relevant norms was impressionistic; the interpretation and operationalisation of these norms requires further technical and non-technical investigation. Indeed, as our analysis makes clear, the norms embedded in different cooperative practices – whether those of science, civic life, or creative exchange –

must themselves be subjected to reflective evaluation and critique (Ackerly, 2000; Walzer, 1993). Lastly, we highlight some practical implications of our proposal with respect to future research on the design of ideal conversational agents and human–language agent interaction. These findings include the need for a context-sensitive evaluation and fine-tuning of language models, and our hope that, once aligned with relevant values, these models can help nurture more productive forms of conversational exchange.

The focus of this paper has been on the English language and the alignment of conversational agents with a particular set of communicative norms for specific discursive domains. Our analysis has drawn heavily upon the pragmatic tradition in linguistics, and upon speech act theory in particular. It could be enriched further through analysis of other sociological and philosophical traditions such as Luhmann's (1995) system theory, Latour's (2007) actor-network theory, or Cameron's (1992) feminist analysis of linguistic theory. In addition to deeper investigation of the norms proposed herein, a complementary exploration of the norms that structure other languages and linguistic traditions is another important task that remains to be explored in further research.

Acknowledgement

We would like to thank Courtney Biles, Martin Chadwick, Julia Haas, Po-Sen Huang, Lisa Anne Hendricks, Geoffrey Irving, Sean Legassick, Donald Martin Jr, Jaylen Pittman, Laura Rimmel, Christopher Summerfield, Laura Weidinger and Johannes Welbl for contributions and feedback on this paper. Particular thanks is owed to Ben Hutchinson and Owain Evans who provided us with detailed comments and advice throughout. Significant portions of this paper were written while Atoosa Kasirzadeh was at DeepMind.

References

Abid, A., Farooqi, M., & Zou, J. (2021). Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 aaai/acm conference on ai, ethics, and society* (pp. 298–306).

Ackerly, B. A. (2000). *Political theory and feminist social criticism*. New York: Cambridge University Press.

Allan, K. (2013). What is common ground? In *Perspectives on linguistic pragmatics* (pp. 285–310). Springer.

Anderson, E. (2004). Uses of value judgments in science: A general argument, with lessons from a case study of feminist research on divorce. *Hypatia*, *19*(1), 1–24.

Androutsopoulos, J. (2014). Languaging when contexts collapse: Audience design in social networking. *Discourse, Context & Media*, *4*, 62–73.

Asher, N., , & Lascarides, A. (2003). *Logics of conversation*. Cambridge, UK: Cambridge University Press.

Asher, N., & Lascarides, A. (2013). Strategic conversation. *Semantics and Pragmatics*, *6*, 2–1.

Austin, J. L. (1962). *How to do things with words*. Oxford: The Clarendon Press.

Bakker, M. A., Chadwick, M. J., Sheahan, H. R., Tessler, M. H., Campbell-Gillingham, L., Balaguer, J., … others (2022). Fine-tuning language models to find agreement among humans with diverse preferences. *arXiv preprint arXiv:2211.15006*.

Bender, E. M. (2013). Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax. *Synthesis lectures on human language technologies*, *6*(3), 1–184.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency* (pp. 610–623).

Bergen, L., Levy, R., & Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, *9*.

Binns, R. (2018). Algorithmic accountability and public reason. *Philosophy & technology*, *31*(4), 543–556.

Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., & Wallach, H. (2021). Stereotyping norwegian salmon: an inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 1004–1015).

Boden, M. A. (1998). Creativity and artificial intelligence. *Artificial intelligence*, *103*(1-2), 347–356.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … others (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Bondi, E., Xu, L., Acosta-Navas, D., & Killian, J. A. (2021). Envisioning communities: a participatory approach towards ai for social good. In *Proceedings of the 2021 aaai/acm conference on ai, ethics, and society* (pp. 425–436).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... others (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Bunge, M. (2012). The correspondence theory of truth. *Semiotica*, *2012*(188), 65–75.

Burleson, B. R., & Kline, S. L. (1979). Habermas' theory of communication: A critical explication. *Quarterly Journal of Speech*, *65*(4), 412–428.

Buss, S. (1999). Appearing respectful: The moral significance of manners. *Ethics*, *109*(4), 795–826.

Calhoun, C. (2000). The virtue of civility. *Philosophy & public affairs*, *29*(3), 251–275.

Cameron, D. (1992). *Feminism and linguistic theory*. Springer.

Carston, R. (2008). Linguistic communication and the semantics/pragmatics distinction. *Synthese*, *165*(3), 321–345.

Chierchia, G., & McConnell-Ginet, S. (2000). *Meaning and grammar: An introduction to semantics*. Boston: MIT press.

Chomsky, N. (1957). *Syntactic structures*. Berlin: De Gruyter Mouton.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... others (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. , *9*, 287–299.

Dale, R., & Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, *19*(2), 233–263.

Delvin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacl-hlt* (pp. 4171–4186).

Díaz, M., Amironesei, R., Weidinger, L., & Gabriel, I. (2022). Accounting for offensive speech as a practice of resistance. In *Proceedings of the sixth workshop on online abuse and harms (woah)* (pp. 192–202).

Douglas, H. (2000). Inductive risk and values in science. *Philosophy of science*, *67*(4), 559–579.

Douglas, H. (2009). *Science, policy, and the value-free ideal*. Pittsburgh: University of Pittsburgh Press.

Elliott, K. C., & McKaughan, D. J. (2014). Nonepistemic values and the multiple goals of science. *Philosophy of Science*, *81*(1), 1–21.

Evans, O., Cotton-Barratt, O., Finnveden, L., Bales, A., Balwit, A., Wills, P., ... Saunders, W. (2021). Truthful ai: Developing and governing ai that does not lie. *arXiv preprint arXiv:2110.06674*.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, *51*(4), 1–30.

Frederking, R. (1996). *Grice's maxims: do the right thing.*

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and machines*, *30*(3), 411–437.

Gabriel, I. (2022). Toward a theory of justice for artificial intelligence. *Daedalus*, *151*(2), 218–231.

Gabriel, I., & Ghazavi, V. (2021). The challenge of value alignment: From fairer algorithms to ai safety. *arXiv preprint arXiv:2101.06060*.

Gallie, W. B. (1955). Essentially contested concepts. In *Proceedings of the aristotelian society* (Vol. 56, pp. 167–198).

Goodman, N. D., & Lassiter, D. (2015). Probabilistic semantics and pragmatics: Uncertainty in language and thought. *The handbook of contemporary semantic theory, 2nd edition. Wiley-Blackwell*.

Grice, H. P. (1968). Utterer's meaning, sentence-meaning, and word-meaning. In *Philosophy, language, and artificial intelligence: Resources for processing natural language* (pp. 49–66). Dordrecht, Boston, London: Kluwer Academic Publishers.

Grice, H. P. (1981). Presupposition and conversational implicature. In C. Peter (Ed.), *Radical pragmatics.* New York: Academic Press.

Grice, H. P. (1989). *Studies in the way of words*. Boston: Harvard University Press.

Haack, S. (1976). The pragmatist theory of truth. *The British Journal for the Philosophy of Science*, *27*(3), 231–249.

Habermas, J. (1984). *The theory of communicative action. Vol. I: Reason and the rationalization of society*. Boston: Beacon Press. ([Translated by T. McCarthy, German, 1981)

Habermas, J. (1987). *The theory of communicative action. Vol. II: Lifeworld and System*. Boston: Beacon Press. ([Translated by T. McCarthy, German, 1981)

Habermas, J. (2003). *Truth and justification*. Cambridge, UK: Polity Press. ([Translated by B. Fultner, German, 1999)

Habermas, J. (2016). Three normative models of democracy. In *Constitutionalism and democracy* (pp. 277–286). London and New York: Routledge.

Heim, I., & Kratzer, A. (1998). *Semantics in generative grammar* (Vol. 1185). Oxford: Blackwell Publishing.

Held, D. (2006). *Models of democracy*. Cambridge, UK: Polity Press.

Henderson, P., Sinha, K., Angelard-Gontier, N., Ke, N. R., Fried, G., Lowe, R., & Pineau, J. (2018). Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 aaai/acm conference on ai, ethics, and society* (pp. 123–129).

Horn, L., & Ward, G. (2008). *The handbook of pragmatics* (Vol. 26). London: Blackwell Publishing.

Kapetanios, E., Tatar, D., & Sacarea, C. (2013). *Natural language processing: semantic aspects*. Boca Raton: CRC Press.

Kaplan, D. (1979). On the logic of demonstratives. *Journal of philosophical logic*, *8*(1), 81–98.

Kasirzadeh, A. (2021). A new role for mathematics in empirical sciences. *Philosophy of science*, *88*(4), 686–706.

Kasirzadeh, A. (2022). Algorithmic fairness and structural injustice: Insights from feminist political philosophy. In *Proceedings of the 2022 aaai/acm conference on ai, ethics, and society* (p. 349–356).

Kasirzadeh, A., & Klein, C. (2021). The ethical gravity thesis: Marrian levels and the persistence of bias in automated decision-making systems. In *Proceedings of the 2021 aaai/acm conference on ai, ethics, and society* (p. 618–626).

Keenan, E. O. (1976). The universality of conversational postulates. *Language in society*, *5*(1), 67–80.

Kim, Y., & Sundar, S. S. (2012). Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior*, *28*(1), 241–250.

Kirkham, R. L. (1992). *Theories of truth: A critical introduction*. Boston: MIT Press.

Kiros, R., Salakhutdinov, R., & Zemel, R. (2014). Multimodal neural language models. In *International conference on machine learning* (pp. 595–603).

Kuhn, T. (1977). Objectivity, value judgment, and theory choice. In *The essential tension: Selected studies in scientific tradition and change* (p. 320–39). Chicago: Chicago University Press.

Kurdi, M. Z. (2016). *Natural language processing and computational linguistics: speech, morphology and syntax* (Vol. 1). London, UK: John Wiley & Sons.

Ladegaard, H. J. (2009). Pragmatic cooperation revisited: Resistance and non-cooperation as a discursive strategy in asymmetrical discourses. *Journal of Pragmatics*, *41*(4), 649–666.

Latour, B. (2007). *Reassembling the social: An introduction to actor-network-theory*. Oup Oxford.

Lee, M. K., Kusbit, D., Kahng, A., Kim, J. T., Yuan, X., Chan, A., ... others (2019). Webuildai: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–35.

Leech, G. N. (1980). *Explorations in semantics and pragmatics*. Amsterdam: John Benjamins Publishing.

Longino, H. E. (1990). *Science as social knowledge*. Princeton, New Jersey: Princeton University Press.

Longino, H. E. (1996). Cognitive and non-cognitive values in science: Rethinking the dichotomy. In *Feminism, science, and the philosophy of science* (pp. 39–58). Dordrecht: Springer.

Luhmann, N. (1995). *Social systems*. Stanford University Press.

Marsh, G. P. (2003). *Man and nature*. Washington: University of Washington Press.

Maulud, D. H., Zeebaree, S. R., Jacksi, K., Sadeeq, M. A. M., & Sharif, K. H. (2021). State of art for semantic analysis of natural language processing. *Qubahan Academic Journal*, *1*(2), 21–28.

McMullin, E. (1982). Values in science. In *Psa: Proceedings of the biennial meeting of the philosophy of science association* (Vol. 1982, pp. 3–28).

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35.

Merton, R. K. (1942). A note on science and democracy. *Journal of Legal & Political Sociology*, *1*, 115–126.

Metzler, D., Tay, Y., Bahri, D., & Najork, M. (2021). Rethinking search: making domain experts out of dilettantes. In *Acm sigir forum* (Vol. 55, pp. 1–27).

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, *8*, 141–163.

Montague, R. (1938). Pragmatics. In *Contemporary philosophy. a survey* (pp. 102 – 122). Florence: La Nuova Italia Editrice.

Morris, C. W. (1938). Foundations of the theory of signs. In *International encyclopedia of unified science* (pp. 1–59). Chicago: Chicago University Press.

Mouffe, C. (1999). Deliberative democracy or agonistic pluralism? *Social research*, 745–758.

Neff, G. (2016). Talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication*.

Perez-Marin, D., & Pascual-Nieto, I. (2011). *Conversational agents and natural language interaction: Techniques and effective practices: Techniques and effective practices*. IGI Global: Illustrated edition.

Popper, K., & Bartley III, W. W. (1985). *Realism and the aim of science: From the postscript to the logic of scientific discovery*. London and New York: Routledge.

Potochnik, A. (2015). The diverse aims of science. *Studies in History and Philosophy of Science Part A*, *53*, 71–80.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018). Machine theory of mind. In *International conference on machine learning* (pp. 4218–4227).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … others (2021). Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., ... et al. (2021). Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Recanati, F. (1989). The pragmatics of what is said. *Mind and language*, *4*(4).

Rooney, P. (1992). On values in science: Is the epistemic/non-epistemic distinction useful? In *Psa: Proceedings of the biennial meeting of the philosophy of science association* (Vol. 1992, pp. 13–22).

Rosset, C. (2021). *Turing-nlg: A 17-billion-parameter language model by microsoft.* Retrieved 2022-01-14, from `https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/`

Rudner, R. (1953). The scientist qua scientist makes value judgments. *Philosophy of science*, *20*(1), 1–6.

Salmon, W. C. (2006). *Four decades of scientific explanation*. Pittsburgh: University of Pittsburgh press.

Sarewitz, D., & Pielke Jr, R. (1999). Prediction in science and policy. *Technology in Society*, *21*(2), 121–133.

Searle, J. R. (1976). A classification of illocutionary acts. *Language in society*, *5*(1), 1–23.

Silverstein, M. (1972). Linguistic theory: syntax, semantics, pragmatics. *Annual review of Anthropology*, *1*(1), 349–382.

Simonton, D. K. (2000). Creativity: Cognitive, personal, developmental, and social aspects. *American psychologist*, *55*(1), 151.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. 142). Cambridge, Massachusetts: Blackwell Publishers Inc.

Stalnaker, R. (2014). *Context*. Oxford: Oxford University Press.

Sternberg, R. J., & Lubart, T. I. (1999). The concept of creativity: Prospects and paradigms. *Handbook of creativity*, *1*(3-15).

Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503*.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., … Stojnic, R. (2022). Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Thomas, J. A. (1995). *Meaning in interaction: An introduction to pragmatics*. London and New York: Routledge.

Trout, J. D. (2002). Scientific explanation and the sense of understanding. *Philosophy of Science*, *69*(2), 212–233.

van Fraassen, B. C. (1980). *The scientific image*. Oxford: Oxford University Press.

Van Valin, R. D., & LaPolla, R. J. (1997). *Syntax: Structure, meaning, and function*. Cambridge, UK: Cambridge University Press.

Walker, R. C. (2018). The coherence theory of truth. In *The oxford handbook of truth* (p. 219–237). Oxford: Oxford University Press.

Walzer, M. (1993). *Interpretation and social criticism* (Vol. 1). Cambridge, Massachusetts: Harvard University Press.

Waseem, Z. (2016). Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on nlp and computational social science* (pp. 138–142).

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., … others (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., … Gabriel, I. (2022). Taxonomy of risks posed by language models. In *Facct: Acm conference on fairness, accountability, and transparency* (p. 214–229).

Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L. A., … Huang, P.-S. (2021). Challenges in detoxifying language models. In *Findings of the association for computational linguistics: Emnlp 2021* (pp. 2447–2469).

Wen, T.-H., Vandyke, D., Mrksic, N., Gasic, M., Rojas-Barahona, L. M., Su, P.-H., … Young, S. (2016). A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

Wierzbicka, A. (1991). Cross-cultural pragmatics: The semantics of human interaction. In *Cross-cultural pragmatics.* Berlin and New York: De Gruyter Mouton.

Zhang, Z., Han, X., Zhou, H., Ke, P., Gu, Y., Ye, D., … others (2021). Cpm: A large-scale generative chinese pre-trained language model. *AI Open*, *2*, 93–99.