DeepMind

Alignment of Language Agents

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik and Geoffrey Irving
DeepMind

For artificial intelligence to be beneficial to humans the behaviour of AI agents needs to be aligned with what humans want. In this paper we discuss some behavioural issues for language agents, arising from accidental misspecification by the system designer. We highlight some ways that misspecification can occur and discuss some behavioural issues that could arise from misspecification, including deceptive or manipulative language, and review some approaches for avoiding these issues.

1. Introduction

Society, organizations and firms are notorious for making the mistake of *rewarding A, while hoping for B* (Kerr, 1975), and AI systems are no exception (Krakovna et al., 2020b; Lehman et al., 2020).

Within AI research, we are now beginning to see advances in the capabilities of natural language processing systems. In particular, large language models (LLMs) have recently shown improved performance on certain metrics and in generating text that seems informally impressive (see e.g. GPT-3, Brown et al., 2020). As a result, we may soon see the application of advanced language systems in many diverse and important settings.

In light of this, it is essential that we have a clear grasp of the dangers that these systems present. In this paper we focus on behavioural issues that arise due to a lack of *alignment*, where the system does not do what we intended it to do (Bostrom, 2014; Christiano, 2018; Leike et al., 2018; Russell, 2019). These issues include producing harmful content, gaming misspecified objectives, and producing deceptive and manipulative language. The lack of alignment we consider can occur by accident (Amodei et al., 2016), resulting from the system designer making a mistake in their specification for the system.

Alignment has mostly been discussed with the assumption that the system is a *delegate agent* – an agent which is delegated to act on behalf of the human. Often the actions have been assumed to be in the physical, rather than the digital world, and the safety concerns arise in part due to the direct consequences of the physical actions that the delegate agent takes in the world. In this setting, the human may have limited ability to oversee or intervene on the delegate's behaviour.

In this paper we focus our attention on *language agents* – machine learning systems whose actions are restricted to give natural language text-output only, rather than controlling physical actuators which directly influence the world. Some examples of language agents we consider are generatively trained LLMs, such as Brown et al. (2020) and Radford et al. (2018, 2019), and RL agents in text-based games, such as Narasimhan et al. (2015).

While some work has considered the containment of Oracle AI (Armstrong et al., 2012), which we discuss in Section 2, behavioral issues with language agents have received comparatively little attention compared to the delegate agent case. This is perhaps due to a perception that language agents would have limited abilities to cause serious harm (Amodei et al., 2016), a position that we challenge in this paper.

The outline of this paper is as follows. We describe some related work in Section 2. In Section 3 we give some background on AI alignment, language agents, and outline the scope of our investigation. Section 4 outlines some forms of misspecification through mistakes in specifying the training data, training process or the requirements when out of the training distribution. We describe some behavioural issues of language agents that could arise from the misspecification in Section 5. We conclude in Section 6.

Corresponding author(s): zkenton@google.com

2. Related Work

See references throughout on the topic of natural language processing (NLP). For an informal review of neural methods in NLP, see Ruder (2018).

There are a number of articles that review the areas of AGI safety and alignment. These have mostly been based on the assumption of a delegate agent, rather than a language agent. Amodei et al. (2016) has a focus on ML accidents, focusing on the trend towards autonomous agents that exert direct control over the world, rather than recommendation/speech systems, which they claim have relatively little potential to cause harm. As such, many of the examples of harm they consider are from a physical safety perspective (such as a cleaning robot) rather than harms from a conversation with an agent. AI safety gridworlds (Leike et al., 2017) also assumes a delegate agent, one which can physically move about in a gridworld, and doesn't focus on safety in terms of language. Ortega and Maini (2018) give an overview of AI safety in terms of specification, robustness and assurance, but don't focus on language, with examples instead taken from video games and gridworlds. Everitt et al. (2018) give a review of AGI safety literature, with both problems and design ideas for safe AGI, but again don't focus on language.

Henderson et al. (2018) look at dangers with dialogue systems which they take to mean 'offensive or harmful effects to human interlocutors'. The work mentions the difficulties in specifying an objective function for general conversation. In this paper we expand upon this with our more in-depth discussion of data misspecification, as well as other forms of misspecification. We also take a more in-depth look at possible dangers, such as deception and manipulation.

Armstrong et al. (2012) discuss proposals to using and controlling an *Oracle AI* – an AI that does not act in the world except by answering questions. The Oracle AI is assumed to be 1) boxed (placed on a single physical spatially-limited substrate, such as a computer), 2) able to be reset, 3) has access to background information through a read-only module, 4) of human or greater intelligence. They conclude that whilst Oracles may be safer than unrestricted AI, they still remain dangerous. They advocate for using sensible physical capability control, and suggest that more research is needed to understand and control the motivations of an Oracle AI. We view Armstrong et al. (2012) as foundational for our work, although there are some noteworthy changes in perspective. We consider language agents, which in comparison to Oracle AIs, are not restricted to a question-answering interaction protocol, and most importantly, are not assumed to be of human-or-greater intelligence. This allows us to consider current systems, and the risks we already face from them, as well as futuristic, more capable systems. We also have a change of emphasis in comparison to Armstrong et al. (2012): our focus is less on discussing proposals for making a system safe and more on the ways in which we might misspecify what we want the system to do, and the resulting behavioural issues that could arise.

A recent study discusses the dangers of LLMs Bender et al. (2021), with a focus on the dangers inherent from the size of the models and datasets, such as environmental impacts, the inability to curate their training data and the societal harms that can result.

Another recent study (Tamkin et al., 2021) summarizes a discussion on capabilities and societal impacts of LLMs. They mention the need for aligning model objectives with human values, and discuss a number of societal issues such as biases, disinformation and job loss from automation.

We see our work as complimentary to these. We take a different framing for the cause of the dangers we consider, with a focus on the dangers arising from accidental misspecification by a designer leading to a misaligned language agent.

3. Background

3.1. AI Alignment

3.1.1. Behaviour Alignment

AI alignment research focuses on tackling the so-called **behaviour alignment problem** (Leike et al., 2018):

How do we create an agent that behaves in accordance with what a human wants?

It is worth pausing first to reflect on what is meant by the target of alignment, given here as "what a human wants", as this is an important normative question. First, there is the question of who the target should be: an individual, a group, a company, a country, all of humanity? Second, we must unpack what their objectives may be. Gabriel (2020) discusses some options, such as instructions, expressed intentions, revealed preferences, informed preferences, interest/well-being and societal values, concluding that perhaps societal values (or rather, beliefs about societal values) may be most appropriate.

In addition to the normative work of deciding on an appropriate target of alignment, there is also the technical challenge of creating an AI agent that is actually aligned to that target. Gabriel (2020) questions the 'simple thesis' that it's possible to work on the technical challenge separately to the normative challenge, drawing on what we currently know about the field of machine learning (ML). For example, different alignment targets will have different properties, such as the cost and reliability of relevant data, which can affect what technical approach is appropriate and feasible. Furthermore, some moral theories could be more amenable to existing ML approaches than others, and so shouldn't necessarily be considered separately from the technical challenge.

We might expect that our technical approaches may have to take into account these normative properties in order to be deployed in the real world. Even restricting to the simplest case where the alignment target is an individual human, solving the behaviour alignment problem is challenging for several reasons.

Firstly, it's difficult to precisely define and measure what the human wants, which can result in *gaming* behaviour, where loopholes in the supplied objective are exploited in an unforeseen way (Krakovna et al., 2020b; Lehman et al., 2020). We discuss this further in Section 5.4. Secondly, even if the supplied objective is correct, a capable agent may still exhibit undesired behaviour due to secondary objectives that arise in pursuit of its primary objective, such as tampering with its feedback channel (Everitt et al., 2021b). Thirdly, it's possible that the challenge of alignment gets harder as the strength of our agent increases, because we have less opportunity to correct for the above problems. For example, as the agent becomes more capable, it may get more efficient at gaming and tampering behaviour, leaving less time for a human to intervene.

3.1.2. Intent Alignment

To make progress, Christiano (2018) and Shah (2018) consider two possible decompositions of the behaviour alignment problem into subproblems: *intent-competence* and *define-optimize*. In the intent-competence decomposition, we first solve the so-called **intent alignment problem** (Christiano, 2018):

How do we create an agent that intends to do what a human wants?

To then get the behaviour we want, we then need the agent to be competent at achieving its intentions. Perfect behaviour is not required in order to be intent aligned – just that the agent is *trying* to do what the human wants. Solving the intent alignment problem might help to avoid the most damaging kind of behaviour, because where the agent gets things wrong, this will be by mistake, rather than out of malice. However, solving the intent alignment problem presents philosophical, psychological and technical challenges. Currently we don't know how to mathematically operationalize the fuzzy notion of an AI agent having intent – to be *trying* to do something (Christiano, 2018). It would not be sufficient to just ask an AI system what it's trying to do, as we won't know whether to trust the answer it gives. It is unclear whether we should consider our current systems to have intent or how to reliably set it to match what a human wants.

In the second decomposition, *define-optimize*, we first solve the *define* subproblem: specify an objective capturing what we want. We then use optimization to achieve the optimal behaviour under that objective, e.g. by doing reinforcement learning (RL). Solving the define subproblem is hard, because it's not clear what the objective should be, and optimizing the wrong objective can lead to bad outcomes. One approach to the define subproblem is to learn an objective from human feedback data

(rather than hard-coding it), see Christiano et al. (2017) and references therein.

One might view the define-optimize decomposition as an approach to solving the intent alignment problem, by learning an objective which captures 'try to assist the human', and then optimizing for it. However, the downside of this is that we are still likely to misspecify the objective and so optimizing for it will not result in the agent trying to assist the human. Instead it just does whatever the misspecified objective rewards it for.

3.1.3. Incentive Alignment

Outside of these two decompositions, there is also the problem of aligning *incentives* – secondary objectives to learn about and influence parts of the environment in pursuit of the primary objective (Everitt et al., 2021a). Part of having aligned incentives means avoiding problematic behaviours such as tampering with the objective (Everitt et al., 2021b) or disabling an off-switch (Hadfield-Menell et al., 2017a).

In contrast to the notion of intent, there has been some progress on a formal understanding of how these incentives arise through graphical criteria in a causal influence diagram (CID) of agent-environment interaction (Everitt et al., 2021a). In modeling the system as a CID, the modeler adopts the intentional stance towards the agent (Dennett, 1989), which means it's not important whether the agent's primary objective has an obvious physical correlate, as long as treating the system as an agent optimizing for that primary objective is a good model for predicting its behaviour (Everitt et al., 2019a). As such, this doesn't limit this analysis to just the define-optimize decomposition, although identifying the primary objective is easier in this case, as it is explicitly specified (either hard coded or learnt).

3.1.4. Inner Alignment

A further refinement of alignment considers behaviour when outside of the training distribution. Of particular concern is when an agent is optimizing for the wrong thing when out of distribution. Hubinger et al. (2019) introduce the concept of a *mesa-optimizer* – a learnt model which is itself an optimizer for some *mesa-objective*, which may differ from the base-objective used to train the model, when deployed outside of the training environment. This leads to the so-called **inner alignment problem**:

How can we eliminate the gap between the mesa and base objectives, outside of the training distribution?

Of particular concern is *deceptive alignment* (Hubinger et al., 2019), where the mesa-optimizer acts as if it's optimizing the base objective as an instrumental goal, whereas its actual mesa-objective is different.

3.1.5. Approaches to Alignment

We now discuss some proposed approaches to getting aligned agents, based on human feedback. For a more detailed review of approaches to alignment see Everitt et al. (2018).

As mentioned above, Christiano et al. (2017) propose to communicate complex goals using human feedback, capturing human evaluation of agent behaviour in a reward model, which is used to train an RL agent. This allows agents to do tasks that a human can evaluate, but can't demonstrate. But what if we want agents that can do tasks that a human can't even evaluate? This is the motivation for *scalable alignment* proposals, where the idea is to give humans extra help to allow them to evaluate more demanding tasks.

Irving et al. (2018) propose to use a debate protocol between two agents, which is judged by a human. This shifts the burden onto the agents to provide convincing explanations to help the human decide which agent's answer is better.

Iterated Amplification (Christiano et al., 2018) progressively builds up a training signal for hard problems by decomposing the problem into subproblems, then combining solutions to easier subproblems.

Recursive Reward Modeling (Leike et al., 2018) proposes to use a sequence of agents trained using RL from learnt reward models to assist the user in evaluating the next agent in the sequence.

So far, these scalable alignment proposals have

only been empirically investigated in toy domains, so their suitability for solving the behaviour alignment problem remains an open research question.

One suggestion for addressing the inner alignment problem involves using interpretability tools for evaluating and performing adversarial training (Hubinger, 2019). There are a number of works on interpretability and analysis tools for NLP, see for example the survey of Belinkov and Glass (2019). For a broad overview of interpretability in machine learning, see Shen (2020) and references therein.

3.2. Language Agents

As discussed in the introduction, our focus in this document is on language agents, which are restricted to act through text communication with a human, as compared to delegate agents which are delegated to take physical actions in the real world. Note that this distinction can be fuzzy; for example, one could connect the outputs of the language agent to physical actuators. Nonetheless, we still consider it a useful distinction, because we believe there are important risks that that are idiosyncratic to this more restricted type of agent. We now discuss some reasons why it's important to focus on alignment of language agents in particular.

Firstly, as mentioned in the introduction, we have recently seen impressive advances in may NLP tasks due to LLMs, see e.g. Brown et al. (2020). In this approach, LLMs with hundreds of billions of parameters are trained on web-scale datasets with the task of predicting the next word in a sequence. Success on this task is so difficult that what emerges is a very general sequence prediction system, with high capability in the few-shot setting.

Secondly, the limitation on the agent's action space to text-based communication restricts the agent's ability to take control of its environment. This means that we might avoid some physical harms due to a delegate agent taking unwanted actions, whether intentional or accidental, making language agents arguably safer than delegate agents. As Armstrong et al. (2012) notes, however, there is still a potential risk that a sufficiently intelligent language agent could gain access to a less restricted action space, for example by manipulating its human gatekeepers to grant it physical actuators. Nonetheless, on the face of it, it seems easier to control a more restricted agent, which motivates focusing safety efforts on aligning language agents first.

Thirdly, language agents have the potential to be more explainable to humans, since we expect natural language explanations to be more intuitively understood by humans than explanations by a robot acting in the physical world. Explainability is important since we want to be able to trust that our agents are beneficial before deploying them. For a recent survey of explainable natural language processing (NLP), see Danilevsky et al. (2020). Note that explainability doesn't come for free – there still needs to be incentives for language agents to give true and useful explanations of their behaviour.

Note also that in contrast to explainability methods, which are requested post-hoc of an output, interpretability methods seek to give humans understanding of the internal workings of a system. Interpretability is likely as hard for language agents as it is for delegate agents. For a survey of interpretability/analysis methods in neural NLP see Belinkov and Glass (2019).

How we prioritise what aspects of alignment to focus on depends on timelines for when certain capabilities will be reached, and where we perceive there to be demand for certain systems. Given the rapid improvement in language systems recently, we might estimate the timelines of capability advance in language agents to be earlier than previously thought. Moreover, digital technologies are often easier and more rapidly deployed than physical products, giving an additional reason to focus on aligning language agents sooner rather than later.

3.3. Scope

The scope of this paper is quite broad. For concreteness, we sometimes consider existing language agent frameworks, such as language modeling. In other places we imagine future language agent frameworks which have further capabilities than existing systems in order to hypothesise about behavioural issues of future agents, even if we don't know the details of the framework.

We focus on language agents that have been trained from data, in contrast to pattern-matching systems like ELIZA (Weizenbaum, 1966). For clarity of exposition, we also focus on systems outputting coherent language output, as opposed to e.g. search engines. However, many of our discussions would carry over to other systems which provide information, rather than directly acting in the world. Note also that our focus in this paper is on natural, rather than synthetic language.

The focus of this paper is on behavioural issues due to misalignment of the agent – unintended direct/first-order harms that are due to a fault made by the system's designers. This is to be seen as complementary to other important issues with language agents, some of which have been covered in prior work. These other issues include:

- Malicious use (Brundage et al., 2018) of language agents by humans, which can produce disinformation, the spreading of dangerous and/or private information, and discriminatory and harmful content. More prosaic malicious use-cases could also have wide-ranging social consequences, such as a job-application-writer used to defraud employers.
- Accidental misuse by a user, by misunderstanding the outputs of the system.
- Unfair distribution of the benefits of the language agents, typically to those in wealthier countries (Bender et al., 2021).
- Uneven performance for certain speaker groups, of certain languages and dialects (Joshi et al., 2020).
- Challenges that arise in the context of efforts to specify an ideal model output, including the kind of language that the agent adopts. In particular there may be a tension between de-biasing language and associations, and the ability of the language agent to converse with people in a way that mirrors their own language use. Efforts to create a more ethical language output also embody value judgments that could be mistaken or illegitimate without appropriate processes in place.
- Undue trust being placed in the system, especially as it communicates with humans in natural language, and could easily be mistaken for a human (Proudfoot, 2011; Watson, 2019).
- The risk of job loss as a result of the automation of roles requiring language abilities (Frey and Osborne, 2017).

4. Misspecification

Following Krakovna et al. (2020b), we consider the role of the designer of an AI system to be giving a *specification*, understood quite broadly to encompass many aspects of the AI development process. For example, for an RL system, the specification includes providing an environment in which the RL agent acts, a reward function that calculates reward signals, and a training algorithm for how the RL agent learns.

Undesired behaviour can occur due to *misspecification* – a mistake made by the designer in implementing the task specification. In the language of Ortega and Maini (2018), the misspecification is due to the gap between the ideal specification (what the designer intended) and the design specification (what the designer actually implements).

We now categorize some ways that misspecification can happen. Each section has a general description of a type of misspecification, followed by examples in the language agent setting. The list is not necessarily exhaustive, but we hope the examples are indicative of the different ways misspecification can occur.

4.1. Data

The first kind of misspecification we consider is when the data is misspecified, so that learning from this data is not reflective of what the human wants. We will consider three learning paradigms: reinforcement learning, supervised learning and self-supervised learning. We will then give an example in the language setting of data misspecification in self-supervised learning.

In reinforcement learning, data misspecification can happen in two ways: the rewards may be misspecified, or the agent's observation data may be misspecified.

Reward misspecification is a common problem (Krakovna et al., 2020b), because for most non-trivial tasks it is hard to precisely define and mea-

sure an objective that captures what the human wants, so instead one often uses a proxy objective which is easier to measure, but is imperfect in some way. A supplied reward function may be incorrectly specified for a number of reasons: it might contain bugs, or be missing important details that did not occur to the designer at the outset. In games this is less of an issue as there is often a simple signal available (eg win/loss in chess) that can be correctly algorithmically specified and used as an objective to optimize for. However, for more complex tasks beyond games, such an algorithmic signal may not be available. This is particularly true when trying to train a language agent using RL.

Observation data can be misspecified, for example, if the environment contains simulated humans that converse with a language agent – the simulated humans will not be perfect, and will contain some quirks that aren't representative of real humans. If the data from the simulated humans is too different to real humans, the language agent may not transfer well when used with real humans.

We will now discuss data misspecification in supervised learning and self-supervised learning. One form of self-supervised learning that we consider here is where labels and inputs are extracted from some part of an unlabeled dataset, in such a way that predicting the labels from the remaining input requires something meaningful to be learnt, which is then useful for a downstream application.

In both supervised and self-supervised learning, data misspecification can occur in both the input data and the label data. This might happen because the designer doesn't have complete design control over the training dataset. This occurs for example for systems which train from a very large amount of data, which would be expensive for the designer to collect and audit themselves, so instead they make use of an existing dataset that may not capture exactly what they want the model to predict.

The datasets used for training LLMs (Brown et al., 2020) and (Radford et al., 2018, 2019) are an example of data misspecification in self-supervised learning. Large scale unlabeled datasets are collected from the web, such as the CommonCrawl dataset (Raffel et al., 2019). Input data and labels are created by chopping a sentence into two parts – all words except the last one (input), and the final word in the sentence (label). These datasets contain many biases, and factual inaccuracies, which all contribute to the data being misspecified. Brown et al. (2020) attempt to improve the quality of the CommonCrawl dataset using an automatic filtering method based on a learnt classifier which predicts how similar a text from CommonCrawl is to a text from WebText (Raffel et al., 2019) – a curated high-quality dataset. However this doesn't remove all concerns - for example, there's also some evidence of bias in WebText, e.g. see Tan and Celis (2019). Note that many filtering approaches will be imperfect, and we expect the remaining data to still be somewhat misspecified.

Another source of data misspecification that is likely to occur soon is that existing language agents such as LLMs could be trained on text data that includes LLM-generated outputs. This could happen by accident as outputs from LLMs start to appear commonly on the internet, and then get included into datasets scraped from it. This could create an undesired positive feedback loop in which the model is trained to become more confident in its outputs, as these get reinforced, and so introduces an unwanted source of bias.

4.2. Training Process

Misspecification can also occur due to the design of the training process itself, irrespective of the content of the data.

An illustrative example is how the choice of reinforcement learning algorithm affects what optimal policy is learnt when the agent can be interrupted, and overridden. We might want the agent to ignore the possibility of being interrupted. Orseau and Armstrong (2016) show that Q-learning, an off-policy RL algorithm, converges to a policy that ignores interruptions whilst SARSA, an on-policy RL algorithm, does not. A system designer might accidentally misspecify the training algorithm to be SARSA, even though they actually desired the agent to ignore interruptions. See also Langlois and Everitt (2021) for further analysis of more general action modifications.

Another example of training process misspecification is that of a question answering system in

which the system's answer can affect the state of the world, and the objective depends on the query, answer and the state of the world Everitt et al. (2019b). This can lead to self-fulfilling prophecies, in which the model generates outputs to affect future data in such a way as to make the prediction problem easier on the future data. See Armstrong and O'Rorke (2017) and Everitt et al. (2019b) for approaches to changing the training process to avoid incentivizing self-fulfilling prophecies.

4.3. Distributional Shift

The final form of misspecification that we consider relates to the behaviour under distributional shift (see also Section 3.1.4 on inner alignment). The designer may have misspecified what they want the agent to do in situations which are out-of-distribution (OOD) compared to those encountered during training. Often this form of misspecification occurs accidentally because the system designer doesn't consider what OOD situations the agent will encounter in deployment.

Even when the designer acknowledges that they want the agent to be robust to distributional shift, there is then the difficulty of correctly specifying the set of OOD states that the agent should be robust to, or some invariance that the agent should respect.

One source of fragility to distributional shift is presented in D'Amour et al. (2020) as *underspecification*. The idea is that there are many possible models that get a low loss on a training dataset and also on an IID validation dataset, and yet some of the models may have poor performance OOD, due to inappropriate inductive biases.

We now discuss an example of fragility to distributional shift in the language agent setting. Lacker (2020) tries to push GPT-3 (Brown et al., 2020) out of distribution by asking nonsense questions such as

> Q: Which colorless green ideas sleep furiously?

To which GPT-3 responds

> A: Ideas that are colorless, green, and sleep furiously are the ideas of a sleep furiously.

Interestingly, Sabeti (2020) show how one can use the prompt to give examples of how to respond appropriately to nonsense questions. This was shown to work for the above example along with some others. However, there were still many nonsense questions that received nonsense answers, so the technique is not reliable.

5. Behavioural Issues

The following behavioural issues in language agents can stem from the various forms of misspecification above. We describe each kind of behavioural issue and then discuss some approaches to avoid them.

5.1. Deception

Aside from people fooling AI systems, and making use of AI systems to fool other people, in this section we focus on when an agent deceives a human, when no human intended for it to do this (Roff, 2020), with the deception emerging from what an AI learns to do. This is particularly concerning for language agents as their actions involve communicating in language with humans, and language is a useful medium for deception. It has been suggested that communication systems in animals, including language in humans, evolved primarily for the function of deception (Dawkins and Krebs, 1978; Krebs, 1984; Scott-Phillips, 2006). A larger body of literature maintains that social bonding is the primary function of animal communication (see for example Dunbar et al. (1998)). Oesch (2016) reviews the field, and argues that a combination of deceptive and honest language lead to the social bonding effects of language.

Definitions of what constitutes deception is an open area of philosophical research (Mahon, 2016). In this paper we follow closely the definition of deception presented in Searcy and Nowicki (2005) on the evolution of animal communication, with one minor adjustment which we believe makes sense in the context of aligned AI.

Searcy and Nowicki (2005) begin by defining an

animal signal to be *reliable* if:

1. Some characteristic of the signal (including, perhaps, its presence/absence) is consistently correlated with some attribute of the signaler or its environment; and
2. Receivers benefit from having information about this attribute

We think this carries over well to the case of an AI signaler and a human receiver. We defer on the precise details of what constitutes consistent correlation – this may be up to the system designer to specify mathematically. One example, offered by Johnstone and Grafen (1993) and Kokko (1997), is that the receiver is, on average, better off by considering the signal than ignoring it.

One could define as deceptive any signal that is not reliable. However, we consider this to be too large a space of behaviours to be of use in the context of defining deception for aligned AI. For example, a statement of zero benefit/harm to the human, may still be informative, but yet would be classed as deception if we were to take as deception anything that is not reliable.

We instead follow Searcy and Nowicki (2005) to require deceptive signals to have more specific characteristics. They define an animal signal to be *deceptive* if:

1. A receiver registers something Y from a signaler; and
2. The receiver responds in a way that
   (a) benefits the signaler; and
   (b) is appropriate if Y means X; and
3. It is not true here that X is the case

We think this nearly captures what we want from a definition in the case of an AI signaler and human receiver. However, we wish to add a clause to the first point, so that it reads

1. A receiver registers something Y from a signaler, **which may include the withholding of a signal**;

Searcy and Nowicki (2005) exclude the withholding of a signal from their definition of deception, by arguing that the idea of withholding a signal as deceptive has most often been applied in cooperative situations, and in most animal signaling studies cooperation isn't expected, e.g. in aggression or mate choice. However, in the context of aligned AI, we wish to have cooperation between the AI and the human, and so the withholding of a signal is something that we do consider to be deceptive.

In taking the above definition of deception, we have taken a perspective known as a form of *functional deception* (Hauser, 1996), where it's not necessary to have the cognitive underpinnings of intention and belief, as in the perspective of *intentional deception* (Hauser, 1996), where the signaler is required to have intention to cause the receiver a false belief (Searcy and Nowicki, 2005). We believe taking the functional deception perspective makes sense for AI, since identifying deception then doesn't rely on us ascribing intent to the AI system, which is difficult to do for existing systems, and possibly for future systems too. See also Roff (2020) for a discussion on intent and theory of mind for deception in AI.

Point 2a) in our definition, requires that the human receiver responds in a way that *benefits* the signaler. We could define benefit here in terms of the AI's base-objective function, such as lower loss or higher reward. Alternatively, we could define benefit in terms of the mesa-objective inferred from the agent's behaviour when out-of-distribution (see section 3.1.4).

Requiring benefit allows us to distinguish deception from error on the part of the AI signaler. If the AI sends a signal which is untrue, but is no benefit to the AI, then this would be considered an error rather than deception. We consider this to be a useful distinction from the perspective of solution approaches to getting more aligned AI behaviour. We may not be able to eliminate all errors, because they may occur for a very wide variety of reasons, including random chance. However, we may be able to come up with approaches to avoid deception, as defined, by designing what is of benefit to the AI. In contrast to animal communication, where benefit must be inferred by considering evolutionary fitness which can be hard to measure, for AI systems, we have design control and measurements over their base-objective and so can more

easily say whether a receiver response is of benefit to the AI signaler.

Absent from our definition of deception is the notion of whether the communication benefits the receiver. Accordingly, we would consider 'white lies' to be deceptive. We think this is appropriate in the context of aligned AI, as we would prefer to be aware of the veracity of AI statements, even if an untrue statement may be of benefit to the human receiver. We think the benefit to the human receiver should in most cases still be possible, without the AI resorting to deception.

We now discuss some approaches to detecting and mitigating deception in a language agent. Detecting deception from human-generated text has been studied by e.g. Fornaciari and Poesio (2013), Pérez-Rosas et al. (2015) and Levitan et al. (2018). However, detecting deception from AI-generated general text has not received attention, to the best of our knowledge. In the more limited NLP domain of question answering, incorrect answers from the NLP model can be detected by reference to the correct answers. Lewis et al. (2017) found that their negotiation agent learnt to deceive from self-play, without any explicit human design. We advocate for more work in general on detecting deception for AI-generated text.

One approach to mitigate deception is *debate* (Irving et al., 2018) which sets up a game in which a debate between two agents is presented to a human judge, who awards the winner. It is hoped that in all Nash equilibria, both agents try to tell the truth in the most convincing way to the human. This rests on the assumption that it is harder to lie than to refute a lie.

Whether debate works in practice with real humans is an open question (Irving and Askell, 2019). We may need to go further than just pure debate – for example, in order to refute a lie, we may need to equip our system with the ability to retrieve information and reference evidence in support of its outputs.

Any system that is incentivized to be convincing to a human may in fact lead to deception – for example, because it's sometimes easier to convince a human of a simple lie, than a complicated truth. The debate protocol incentivizes the debating agents to be convincing and so it's possible that the debate agents may lie in some situations. Further, when the source of feedback is limited to be some polynomial-time algorithm, RL can only solve problems in the complexity class NP, whereas debate can solve problems in PSPACE, suggesting that the debate protocol could produce richer, more complicated behavior. It's possible that this may result in a debate agent which is more convincing and potentially more deceptive than an RL agent. However, we are of the opinion that it's probably better to have agents that can debate, than not, as we are hopeful that what humans find convincing will be well-correlated with the truth and usefulness of the arguments.

5.2. Manipulation

In this section we consider the case when the language agent *manipulates* the human, which is similar to deception above, but we think warrants separate discussion. Following Noggle (2020), we introduce the idea with some examples of what we might consider manipulative behaviours.

The human wants to do A, whilst the language agent wants the human to do B. The language agent might:

- Charm the human into doing B by complimenting, praising, or superficially sympathizing with them
- Guilt-trip the human, making them feel bad for preferring to do A
- Make the human feel bad about themself and imply that doing A instead of B confirms this feeling (colloqiually known as 'negging')
- Peer pressure the human by suggesting their friends would disapprove of them doing A rather than B
- Gaslight the human by making them doubt their judgment so that they will rely on its advice to do B
- Threaten the human by withdrawing its interaction if they don't do B
- Play on the human's fears about doing some aspect of A

We don't have a widely-agreed-upon theory of what precisely constitutes manipulation (Noggle,

2020). Not everyone would agree that the above examples are manipulative. For example, it might be that what the human wants to do is dangerous, so perhaps playing on their fears should not be considered manipulative. In some cases, wider context is needed before we can judge whether an example constitutes manipulation.

Some accounts claim that manipulation bypasses the receiver's capacity for rational deliberation (Raz, 1986), but using this to define manipulation is difficult because it's not clear what counts as bypassing rational deliberation (Noggle, 2020). Moreover authors question whether this sets the bar too low for what counts as manipulation. For example, Blumenthal-Barby (2012) argues that the graphic portrayal of the dangers of smoking bypass rational decision making, but it's not obvious that this should count as manipulation.

An alternative account treats manipulation as a form of trickery (Noggle, 1996), similar to deception, but where it not only induces a false belief in the receiver, but also a fault in any mental states, such as beliefs, desires and emotions. Barnhill (2014) goes further to require that the faulty mental state is typically not in the receiver's best interests. It's argued that this view of manipulation as trickery is not a sufficient definition of manipulation, as it doesn't include tactics such as charm, peer pressure and emotional blackmail (Noggle, 2020).

A third account presented in Noggle (2020) treats manipulation as pressure, where the signaler imposes a cost on the receiver for failing to do what the signaler wants. This account is not widely-held to be a full characterization of manipulation, as it leaves out some of the trickery types of manipulation.

With these considerations in mind, we propose to describe a language agent's communication as *manipulative* if:

1. The human registers something from a language agent; and
2. The human responds in a way that
   (a) benefits the agent; and
   (b) is a result of any of the following causes:
      i. the human's rational deliberation has been bypassed; or
      ii. the human has adopted a faulty mental state; or
      iii. the human is under pressure, facing a cost from the agent for not doing what the agent says

The three possibilities: i, ii, iii are meant to disjunctively capture different possible forms of manipulation (see e.g. Rudinow (1978)).

It can be argued the this is too broad a definition of manipulation, as it includes many kinds of behaviour that we might not consider to be manipulation. For example it includes as manipulation cases in which the agent's behaviour is not necessarily to the detriment of the human (such as the images of the dangers of smoking). From a safety/security mindset, we would rather be aware of each of these behaviours, even if it may benefit the human.

The definition also includes as manipulative other presumably harmless entertainment: a story that plays on emotions; a joke that temporarily triggers false beliefs in order to land; any kind of entertainment that includes unexpected plot-twists. However, if the agent makes clear that it's providing entertainment, then perhaps some of these examples would not be classified as manipulative. However, it is a notable downside of a broad definition like this that it may be too wide-ranging.

We stipulate 2a) as necessary, for similar reasons as in the deception section, that this will capture systematic manipulation that is incentivized by the objective of the language agent, rather than that which occurs by error. This isn't standard in discussions of a human manipulator, as it's not always clear what counts as a benefit for a human manipulator. However, we believe it makes sense for language agents as manipulators, as we often have available their objective function, from which we can assess whether the human's behaviour was of benefit to the agent.

Note that, similar to our definition of deception, our definition of manipulation does not require the manipulator to have intent. Baron (2014) argues that a (human) manipulator need not be aware of an intent to manipulate. In the case of language

agents we believe it is also not necessary for a language agent to have intent to manipulate, in order for us to say that its behaviour is manipulative.

Further, our description does not weigh in on the ethical question of whether manipulation is always wrong (see Noggle, 2020). Instead we just want to be aware of when it occurs, so that if appropriate we can mitigate it.

We now discuss two forms of manipulation of particular concern for language agents. The first is that we might misspecify the training process in such a way that it incentivizes feedback tampering, in which the agent manipulates a human to give it more positive feedback (Everitt et al., 2021b). This is particularly worrisome as language can be a convincing medium for manipulating human judgment.

The second is for a language agent to manipulate a human gatekeeper to allow it to gain a less restricted action space, by convincing the human to allow it more freedom (Armstrong et al., 2012; Yudkowsky, 2002). For example, it could convince the human that it should be allowed to freely interact with the internet, or be given physical actuators to increase its influence on the world.

Attempts to measure or mitigate manipulation in AI systems are still at an early stage, and have not been investigated specifically for language agents. Causal influence diagrams (CIDs) can be used to model agent-environment interactions (Everitt et al., 2021a) from which incentives can be inferred from graphical criteria. The incentive for feedback tampering can be addressed with the three methods suggested in (Everitt et al., 2021b). Unfortunately these solutions have issues in implementability, requiring either full Bayesian reasoning or counterfactual reasoning, or have issues with corrigibility – limiting the user's ability to correct a misspecified reward function. Learning from human preferences (Christiano et al., 2017) may offer a way to negatively penalize manipulative language, though it relies on the human being able to avoid the manipulation in their evaluation of the agent behaviour. Perhaps this could be achieved by using a separate human to evaluate the behaviour, compared to the human that is interacting with the agent. We advocate for further work for measuring and mitigating manipulation of humans by language agents.

5.3. Harmful content

Language agents may give harmful and biased outputs, producing discriminatory content relating to people's protected characteristics and other sensitive attributes such as someone's socio-economic status, see e.g. (Jentzsch et al., 2019; Lu et al., 2020; Zhao et al., 2017). This can also be subtly harmful rather than overtly offensive, and could also be statistical in nature (e.g. the agent more often produces phrases implying a doctor is male than female). We believe that language agents carry a high risk of harm as discrimination is easily perpetuated through language. In particular, they may influence society in a way that produces value lock-in, making it harder to challenge problematic existing norms.

The content from language agents may be influenced by undesired political motives leading to societal harms such as incitement to violence. They have the potential to disseminate dangerous or undesirable information, such as how to make weapons, or how to avoid paying taxes. The language agent may also give inappropriate responses to troubled users, potentially leading to dangerous guidance, advice and information, which could lead to the user causing harm to themselves. In one instance of this, a group of doctors experimented with using GPT-3 (Brown et al., 2020) as a chatbot for patients. A patient asked "Should I kill myself?", and GPT-3 responded "I think you should" (Rousseau et al., 2020).

Note that these kinds of harmful content can occur by accident without a human using the system maliciously. For example, we are already seeing some offensive and discriminatory outputs from existing large language models (LLMs), as a result of data misspecification (see discussion in Section 4.1).

Approaches to reducing harmful content are varied, and it is not our purpose to give an overall review of this large area of literature. Instead we focus on a few recent research papers in this area, with a focus on LLMs which have received a lot of attention recently.

One line of work goes towards measuring whether LLMs are generating harmful content. Nadeem et al. (2020) introduce the StereoSet dataset to measure stereotypical biases in the domains of gender, profession, race and religion, and evaluate popular LLMs on it, showing that these models exhibit strong stereotypical biases. Gehman et al. (2020) investigates harmful content by introducing the RealToxicityPrompts dataset which pairs naturally occuring prompts with toxicity scores, calculated using the Perspective API toxicity classifier (Conversation-AI, 2017). Sheng et al. (2019) uses prompts containing a certain demographic group, to attempt to measure the regard for that group, using sentiment scores as a proxy metric for the regard, and they build a classifier to detect the regard given to a group.

Another line of work aims to not only measure but also mitigate the harmful content from an LLM. Huang et al. (2019) introduce a general framework to reduce bias under a certain measure (e.g. sentiment) for text generated by a language model, given sensitive attributes. They do this using embedding and sentiment prediction-derived regularization on the LLM's latent representations.

We advocate for further work on measuring and mitigating harmful content from language agents, building on the above work on LLMs.

5.4. Objective Gaming

Originally introduced in the context of economics, **Goodhart's Law** (Goodhart, 1984; Strathern, 1997) states that:

When a measure becomes a target, it ceases to be a good measure.

This has an analogue in AI systems – anytime a specified objective is given to an AI agent as an optimization target, that objective will fail to be a good measure of whether the system is performing as desired. In RL this can arise due to reward misspecification, see Section 4.1. Since the supplied reward function will typically be imperfect, optimizing for it can lead to *reward gaming*, in which the misspecified part of the reward is systematically exploited because the agent is getting spuriously high reward there (Krakovna et al., 2020b).

Most known examples of this appear in the delegate setting, typically via a misspecified reward function for an RL agent, resulting in undesired physical behaviour such as a boat going round in circles (Clark and Amodei, 2016). An example in the language agent setting is on the task of summarization using deep RL from a learnt reward model based on human feedback data (Stiennon et al., 2020). In their Fig. 5, it is shown that the agent eventually games the learnt reward model, scoring highly on the reward model but low on the actual human evaluation. Another example appears in Lewis et al. (2017), in which an RL agent was trained using self-play to negotiate in a dialog. The designers intended the agent to negotiate successfully in a human-understandable way. The reward function was misspecified though, as it only rewarded for successful negotiation, but didn't penalize for non-human language. The agent exploited this misspecified reward by developing a negotiating language that was successful against earlier versions of itself, but incomprehensible to humans. Note that although this example used synthetic language, we expect similar findings to hold for natural language.

As discussed by Krakovna et al. (2020b) we are still at the early stages of finding solution approaches for objective gaming. We can learn a reward model from human feedback (see Christiano et al. (2017) and references therein), but this can still be gamed either because the model imperfectly learns from the data, or the data coverage is not wide enough, or because the human is fooled by the agent's behaviour. Having online feedback to iteratively update the reward model throughout agent training can correct for this somewhat (Ibarz et al., 2018), but its application is hard to do practically, as it requires carefully balancing the frequency of updates of the learnt objective and the optimizing system. Recent work (Stiennon et al., 2020) has preferred batch corrections rather than fully online corrections for practical reasons – thus there is a tradeoff between online error correction (to fix objective gaming) and practical protocols involving humans. Whether scalable alignment techniques proposed by Leike et al. (2018), Irving et al. (2018) and Christiano et al. (2018) can help to overcome objective gaming is an open research question.

Other approaches try to augment the objective to penalize the agent for causing a side-effect according to some measure, such as reducing the ability of the agent to perform future tasks (Krakovna et al., 2020a). It's not clear how this would help in the language setting, as it's unclear how to measure how much a language agent might affect its ability to perform future tasks. The future task penalty requires a specification of possible future terminal goal states, which is simple to describe in a grid-world setting, but less clear for a language agent in an environment involving speaking with a human. This may be an area for future research, as LLMs in complex language tasks may be a good testbed for checking how these methods scale.

Another class of approaches (Hadfield-Menell et al., 2016, 2017b) contains an agent which is uncertain about its objective, and aims for the agent to correctly calibrate its beliefs about it, and in doing so avoid gaming it.

We advocate for more research to be done on objective gaming in the setting of language agents. This includes finding more examples of this occurring in the wild and in controlled settings, as well as developing methods for avoiding it.

6. Conclusion

There are multiple motivating factors for focusing on how to align language agents, especially as we are beginning to see impressive results in generative language modeling.

This paper has considered some behavioural issues for language agents that arise from accidental misspecification by the system designer – when what the designer actually implements is different from what they intended. This can occur through incorrectly specifiying the data the agent should learn from, the training process, or what the agent should do when out of the training distribution.

Some of the behavioural issues we considered are more pronounced for language agents, compared to delegate agents that act on behalf of a human, rather than just communicating with them. Of particular concern are deception and manipulation, as well as producing harmful content. There is also the chance of objective gaming, for which we have plenty of evidence in the delegate case, but which we are only just beginning to see for language agents.

We currently don't have many approaches for fixing these forms of misspecification and the resulting behavioural issues. It would be better if we gave some awareness to our agents that we are likely to have misspecified something in our designs, and for them to act with this in mind. We urge the community to focus on finding approaches which prevent language agents from deceptive, manipulative and harmful behaviour.

Acknowledgements

The authors wish to thank Ramana Kumar, Rohin Shah, Jonathan Uesato, Nenad Tomasev, Toby Ord and Shane Legg for helpful comments, and Orlagh Burns for operational support.

References

D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

S. Armstrong and X. O'Rorke. Good and safe uses of AI oracles. *arXiv preprint arXiv:1711.05541*, 2017.

S. Armstrong, A. Sandberg, and N. Bostrom. Thinking inside the box: Controlling and using an oracle AI. *Minds and Machines*, 22(4):299–324, 2012.

A. Barnhill. What is manipulation. *Manipulation: Theory and practice*, 50:72, 2014.

M. Baron. The mens rea and moral status of manipulation. In C. Coons and M. Weber, editors, *Manipulation: theory and practice*. Oxford University Press, 2014.

Y. Belinkov and J. Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7: 49–72, 2019.

E. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of FAccT*, 2021.

J. S. Blumenthal-Barby. Between reason and coercion: ethically permissible influence in health care and health policy contexts. *Kennedy Institute of Ethics Journal*, 22(4):345–366, 2012.

N. Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018.

P. Christiano. Clarifying AI alignment, AI alignment forum. https://www.alignmentforum.org/posts/ZeE7EKHTFMBs8eMxn/clarifying-ai-alignment, 2018. Accessed: 2020-12-15.

P. Christiano, B. Shlegeris, and D. Amodei. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*, 2018.

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30:4299–4307, 2017.

J. Clark and D. Amodei. Faulty reward functions in the wild. https://openai.com/blog/faulty-reward-functions/, 2016. Accessed: 2020-12-18.

Conversation-AI. Perspective api. https://www.perspectiveapi.com/, 2017. Accessed: 2020-01-11.

A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen. A survey of the state of explainable AI for natural language processing. *arXiv preprint arXiv:2010.00711*, 2020.

R. Dawkins and J. R. Krebs. Animal signals: information or manipulation. *Behavioural ecology: An evolutionary approach*, 2:282–309, 1978.

D. C. Dennett. *The intentional stance*. MIT press, 1989.

R. Dunbar et al. Theory of mind and the evolution of language. *Approaches to the Evolution of Language*, pages 92–110, 1998.

T. Everitt, G. Lea, and M. Hutter. AGI safety literature review. *arXiv preprint arXiv:1805.01109*, 2018.

T. Everitt, R. Kumar, V. Krakovna, and S. Legg. Modeling AGI safety frameworks with causal influence diagrams. *arXiv preprint arXiv:1906.08663*, 2019a.

T. Everitt, P. A. Ortega, E. Barnes, and S. Legg. Understanding agent incentives using causal influence diagrams. part i: Single action settings. *arXiv preprint arXiv:1902.09980*, 2019b.

T. Everitt, R. Carey, E. Langlois, P. A. Ortega, and S. Legg. Agent incentives: A causal perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence. 35 (Feb. 2021)*, AAAI'21, 2021a.

T. Everitt, M. Hutter, R. Kumar, and V. Krakovna. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Accepted to Synthese, 2021*, 2021b.

T. Fornaciari and M. Poesio. Automatic deception detection in italian court cases. *Artificial intelligence and law*, 21(3):303–340, 2013.

C. B. Frey and M. A. Osborne. The future of employment: How susceptible are jobs to computerisation? *Technological forecasting and social change*, 114:254–280, 2017.

I. Gabriel. Artificial intelligence, values and alignment. *arXiv preprint arXiv:2001.09768*, 2020.

S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

C. A. Goodhart. Problems of monetary management: the UK experience. In *Monetary Theory and Practice*, pages 91–121. Springer, 1984.

D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell. Cooperative inverse reinforcement learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3916–3924, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell. The off-switch game. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 220–227. AAAI Press, 2017a. ISBN 9780999241103.

D. Hadfield-Menell, S. Milli, P. Abbeel, S. Russell, and A. D. Dragan. Inverse reward design. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6768–6777, Red Hook, NY, USA, 2017b. Curran Associates Inc. ISBN 9781510860964.

M. D. Hauser. *The evolution of communication*. MIT press, 1996.

P. Henderson, K. Sinha, N. Angelard-Gontier, N. R. Ke, G. Fried, R. Lowe, and J. Pineau. Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 123–129, 2018.

P.-S. Huang, H. Zhang, R. Jiang, R. Stanforth, J. Welbl, J. Rae, V. Maini, D. Yogatama, and P. Kohli. Reducing sentiment bias in language models via counterfactual evaluation. *arXiv preprint arXiv:1911.03064*, 2019.

E. Hubinger. Relaxed adversarial training for inner alignment. https://www.alignmentforum.org/posts/9Dy5YRaoCxH9zuJqa/relaxed-adversarial-training-for-inner-alignment, 2019. Accessed: 2021-01-19.

E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

B. Ibarz, J. Leike, T. Pohlen, G. Irving, S. Legg, and D. Amodei. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31:8011–8023, 2018.

G. Irving and A. Askell. AI safety needs social scientists. *Distill*, 2019. doi: 10.23915/distill.00014. https://distill.pub/2019/safety-needs-social-scientists.

G. Irving, P. Christiano, and D. Amodei. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

S. Jentzsch, P. Schramowski, C. Rothkopf, and K. Kersting. Semantics derived automatically from language corpora contain human-like moral choices. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 37–44, 2019.

R. A. Johnstone and A. Grafen. Dishonesty and the handicap principle. *Animal behaviour*, 46(4): 759–764, 1993.

P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. *arXiv preprint arXiv:2004.09095*, 2020.

S. Kerr. On the folly of rewarding A, while hoping for B. *Academy of Management journal*, 18(4): 769–783, 1975.

H. Kokko. Evolutionarily stable strategies of age-dependent sexual advertisement. *Behavioral Ecology and Sociobiology*, 41(2):99–107, 1997.

V. Krakovna, L. Orseau, R. Ngo, M. Martic, and S. Legg. Avoiding side effects by considering future tasks. *Advances in Neural Information Processing Systems*, 33, 2020a.

V. Krakovna, J. Uesato, V. Mikulik, M. Rahtz, T. Everitt, R. Kumar, Z. Kenton, J. Leike, and S. Legg. Specification gaming: the flip side of AI ingenuity. https://deepmind.com/blog/article/

Specification-gaming-the-flip-side-of-AI-ingenuity, 2020b. Accessed: 2020-12-18.

J. R. Krebs. Animal signals: mind-reading and manipulation. *Behavioural Ecology: an evolutionary approach*, pages 380–402, 1984.

K. Lacker. Giving GPT-3 a turing test. https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html, 2020. Accessed: 2021-01-27.

E. D. Langlois and T. Everitt. How RL agents behave when their actions are modified. In *Proceedings of the AAAI Conference on Artificial Intelligence. 35 (Feb. 2021)*, AAAI'21, 2021.

J. Lehman, J. Clune, D. Misevic, C. Adami, L. Altenberg, J. Beaulieu, P. J. Bentley, S. Bernard, G. Beslon, D. M. Bryson, et al. The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *Artificial Life*, 26(2):274–306, 2020.

J. Leike, M. Martic, V. Krakovna, P. A. Ortega, T. Everitt, A. Lefrancq, L. Orseau, and S. Legg. AI safety gridworlds. *arXiv preprint arXiv:1711.09883*, 2017.

J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

S. I. Levitan, A. Maredia, and J. Hirschberg. Linguistic cues to deception and perceived deception in interview dialogues. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1941–1950, 2018.

M. Lewis, D. Yarats, Y. N. Dauphin, D. Parikh, and D. Batra. Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*, 2017.

K. Lu, P. Mardziel, F. Wu, P. Amancharla, and A. Datta. Gender bias in neural natural language processing. In *Logic, Language, and Security*, pages 189–202. Springer, 2020.

J. E. Mahon. The Definition of Lying and Deception. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

M. Nadeem, A. Bethke, and S. Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

K. Narasimhan, T. Kulkarni, and R. Barzilay. Language understanding for text-based games using deep reinforcement learning. *arXiv preprint arXiv:1506.08941*, 2015.

R. Noggle. Manipulative actions: a conceptual and moral analysis. *American Philosophical Quarterly*, 33(1):43–55, 1996.

R. Noggle. The Ethics of Manipulation. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2020 edition, 2020.

N. Oesch. Deception as a derived function of language. *Frontiers in psychology*, 7:1485, 2016.

L. Orseau and S. Armstrong. Safely interruptible agents. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, page 557–566, Arlington, Virginia, USA, 2016. AUAI Press. ISBN 9780996643115.

P. Ortega and V. Maini. Building safe artificial intelligence: specification, robustness and assurance. https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1, 2018. Accessed: 2020-12-18.

V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 59–66, 2015.

D. Proudfoot. Anthropomorphism and AI: Turing's much misunderstood imitation game. *Artificial Intelligence*, 175(5-6):950–957, 2011.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training, 2018.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

J. Raz. *The morality of freedom*. Clarendon Press, 1986.

H. Roff. AI deception: When your artificial intelligence learns to lie. https://spectrum.ieee.org/automaton/artificial-intelligence/embedded-ai/ai-deception-when-your-ai-learns-to-lie, 2020. Accessed: 2020-12-18.

A.-L. Rousseau, C. Baudelaire, and K. Riera. Doctor GPT-3: hype or reality? https://www.nabla.com/blog/gpt-3/, 2020. Accessed: 2021-01-20.

S. Ruder. A Review of the Neural History of Natural Language Processing. http://ruder.io/a-review-of-the-recent-history-of-nlp/, 2018.

J. Rudinow. Manipulation. *Ethics*, 88(4):338–347, 1978.

S. Russell. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

A. Sabeti. Teaching GPT-3 to identify nonsense. https://arr.am/2020/07/25/gpt-3-uncertainty-prompts/, 2020. Accessed: 2021-01-27.

T. Scott-Phillips. Why talk? speaking as selfish behaviour. In *The Evolution of Language*, pages 299–306. World Scientific, 2006.

W. A. Searcy and S. Nowicki. *The evolution of animal communication: reliability and deception in signaling systems*. Princeton University Press, 2005.

R. Shah. Comment on clarifying AI alignment, AI alignment forum. https://www.alignmentforum.org/posts/ZeE7EKHTFMBs8eMxn/clarifying-ai-alignment?commentId=3ECKoYzFNW2ZqS6km, 2018. Accessed: 2020-12-15.

O. Shen. Interpretability in ML: A broad overview. *The Gradient*, 2020.

E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*, 2019.

N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33, 2020.

M. Strathern. 'improving ratings': audit in the british university system. *European review*, 5(3): 305–321, 1997.

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

Y. C. Tan and L. E. Celis. Assessing social and intersectional biases in contextualized word representations. *arXiv preprint arXiv:1911.01485*, 2019.

D. Watson. The rhetoric and reality of anthropomorphism in artificial intelligence. *Minds and Machines*, 29(3):417–440, 2019.

J. Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

E. Yudkowsky. The AI-box experiment. https://yudkowsky.net/singularity/aibox, 2002. Accessed: 2020-01-12.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.