AI & SOCIETY (2025) 40:959–969
https://doi.org/10.1007/s00146-023-01748-4

OPEN FORUM

Language agents reduce the risk of existential catastrophe

Simon Goldstein[1] · **Cameron Domenico Kirk-Giannini**[2]

Received: 1 June 2023 / Accepted: 3 August 2023 / Published online: 19 August 2023

Abstract
Recent advances in natural-language processing have given rise to a new kind of AI architecture: the *language agent*. By repeatedly calling an LLM to perform a variety of cognitive tasks, language agents are able to function autonomously to pursue goals specified in natural language and stored in a human-readable format. Because of their architecture, language agents exhibit behavior that is predictable according to the laws of folk psychology: they function as though they have desires and beliefs, and then make and update plans to pursue their desires given their beliefs. We argue that the rise of language agents significantly reduces the probability of an existential catastrophe due to loss of control over an AGI. This is because the probability of such an existential catastrophe is proportional to the difficulty of aligning AGI systems, and language agents significantly reduce that difficulty. In particular, language agents help to resolve three important issues related to aligning AIs: reward misspecification, goal misgeneralization, and uninterpretability.

keywords Language agents · Existential risk · Reward misspecification · Goal misgeneralization · Interpretable AI

1 Misalignment and existential catastrophe

There is a significant chance that artificial general intelligence will be developed in the not-so-distant future—by 2070, for example. How likely is it that the advent of AGI will lead to an existential catastrophe for humanity? Here it is worth distinguishing between two possibilities: an existential catastrophe could result from humans losing control over an AGI system (call this a *misalignment catastrophe*), or an existential catastrophe could result from humans using an AGI system deliberately to bring that catastrophe about (call this a *malicious actor catastrophe*). In what follows, we are interested in assessing the probability of a misalignment catastrophe rather than a malicious actor catastrophe.

Carlsmith (2021) helpfully structures discussion of the probability of a misalignment catastrophe around six propositions. Because we are interested in the probability of a misalignment catastrophe conditional on the development of AGI, we focus our attention on the final four of these propositions, which we summarize as follows:

1. Of the following two options, the first will be much more difficult:
   a. Build AGI systems with an acceptably low probability of engaging in power-seeking behavior.
   b. Build AGI systems that perform similarly but do not have an acceptably low probability of engaging in power-seeking behavior.
2. Some AGI systems will be exposed to inputs which cause them to engage in power-seeking behavior.
3. This power-seeking will scale to the point of permanently disempowering humanity.
4. This disempowerment will constitute an existential catastrophe.

Carlsmith assigns a probability of 0.4 to (1) conditional on the rise of AGI, a probability of 0.65 to (2) conditional on (1) and the rise of AGI, a probability of 0.4 to (3) conditional on (1), (2), and the rise of AGI, and a probability of 0.95 to (4) conditional on (1–3) and the rise of AGI. This translates into a probability of approximately 0.1 (10%) for a misalignment catastrophe conditional on the rise of AGI.

✉ Cameron Domenico Kirk-Giannini
camerondomenico.kirkgiannini@gmail.com

Simon Goldstein
simon.d.goldstein@gmail.com

[1] Dianoia Institute of Philosophy, Australian Catholic University, Fitzroy, Australia

[2] Department of Philosophy, Rutgers University–Newark, Newark, USA

We believe that the development of a new kind of AI architecture, the *language agent*, ought to significantly decrease assessments of these probabilities. By repeatedly calling an LLM to perform a variety of cognitive tasks, language agents are able to function autonomously to pursue goals specified in natural language and stored in a human-readable format. We suggest that the development of language agents reduces the probability of (1) conditional on the rise of AGI very substantially, the probability of (2) conditional on (1) and the rise of AGI moderately, and the probability of (3) conditional on (1), (2), and the rise of AGI very substantially**.** We work through two numerical examples in Sect. 5; in the meantime, suffice it to say that we believe that updating on the rise of language agents should reduce rational credences in a misalignment catastrophe conditional on the development of AGI by approximately one order of magnitude.

Because language agent architectures have the potential to reduce the risk of a misalignment catastrophe in so many ways, and because the machine learning community's actions in the near future will determine how widely deployed language agent architectures are and thus how much of this potential risk reduction is realized, we believe that language agents are an underappreciated crux in thinking about existential risk related to AI. Priority should be given to further research into the capabilities of language agents and further support for the development of AI systems which implement language agent architectures.

Here is our plan for what follows. Section 2 introduces some of the safety concerns about AI systems created using deep learning that motivate worries about a misalignment catastrophe. Section 3 describes the architecture of language agents in more detail. Section 4 returns to the safety concerns from Sect. 2 and explains how language agents help to address them. Section 5 describes the implications of our arguments for the probability of a misalignment catastrophe. Section 6 concludes by responding to some potential concerns about language agents.

2 Difficulties with alignment

In deep learning, we train an AI system incorporating an artificial neural network to achieve a goal by specifying a mathematical function that encodes the goal (the *objective function*) and then using a learning algorithm to adjust the weights in the network so that the system's performance comes closer to maximizing or minimizing that function. Say that an AI system is *fully aligned* if it has an acceptably low probability of engaging in power-seeking behavior. There are several ways an AI system trained using deep learning could end up less than fully aligned.

2.1 Reward misspecification

A first challenge is *reward misspecification*.[1] When training an AI, we may experiment with different objective functions. In reinforcement learning, the goal is to define a reward function that gives the agent a reward for performing actions that produce desired states. In supervised learning, the goal is to define a loss function that is minimized when the system performs its task optimally.

The problem is that it is difficult to design a reward or loss function that properly encodes a goal. For example, Popov et al. (2017) set out to teach a reinforcement learning agent to stack red Legos on top of blue Legos. They tried to capture this goal by rewarding the agent for the height of the bottom of the red Lego, since stacked red Legos are higher off the ground than unstacked red Legos. However, the agent did not learn to stack Legos; instead, it learned to flip red Legos over, thus elevating their bottoms without stacking them.

To appreciate the difficulty of choosing the right reward function, consider the common reinforcement learning practice of *reward shaping*. Reinforcement learning agents often encounter sparse reward functions. If one rewards an agent only when it wins a game, for example, it may have difficulty identifying which of its behaviors leading up to that outcome should be repeated in future games. Reward shaping solves the problem of sparse reward functions by rewarding the agent for important subgoals on the way to achieving its real goal.

However, reward shaping can also lead to reward misspecification. For example, Amodei and Clark (2016) consider the case of teaching a reinforcement learning agent to play Coast Runners, a game in which the player pilots a boat. A human player would immediately recognize that the game designers' intention is for players to race each other around the track. However, the reinforcement learning setup rewarded the agent with a score for hitting targets along the way. Instead of finishing the race, the AI learned how to loop the boat in a small lagoon, hitting intermediate targets repeatedly to achieve a high score. Rather than rewarding the agent for the final goal, the experimental design rewarded it for intermediate means: "the agent was given a shaping reward for hitting green blocks along the racetrack, which changed the optimal policy to going in circles and hitting the same green blocks over and over again" (Krakovna et al. 2020). A reward optimizer cannot see the distinction

[1] The phenomenon we call reward misspecification is sometimes also called "reward hacking" (e.g. by Amodei et al. 2016), "specification gaming" (e.g. by Shah et al 2022), or, in the context of supervised learning, "outer misalignment."

between intrinsic and instrumental goals: it only optimizes for the reward function it has.

Worryingly, reward misspecification is prone to arise in the context of reinforcement learning with human feedback (RLHF) (Christiano et al. 2017). Because they optimize for human approval, RLHF agents sometimes learn to deceive human assessors. For example, one agent was given the task of grasping a ball. It learned to trick human assessors by hovering its arm between the camera and the ball. Similarly, Perez et al. (2022) found that large language models trained by RLHF tend to behave sycophantically, answering differently depending on what they expect their human users to think.

There is a long list of examples of reward misspecification involving many kinds of AI, many kinds of games, and many different types of reward. In Sect. 4, we'll argue that language agents offer a systematic solution to the problem of reward misspecification.

2.2 Goal misgeneralization

Another challenge for alignment is *goal misgeneralization* (Langosco et al. 2022; Shah et al. 2022).[2] Even when the objective function for a task has been appropriately specified, an AI system may learn a strategy which achieves high performance on that task in some circumstances but not others. ML models are trained on data, environments, and problems that can be different from the data, environments, and problems to which they are later exposed when they are deployed. When an AI is used in a new context that does not resemble the one in which it was trained, we say that this context is *out-of-distribution*. In cases of goal misgeneralization, the AI succeeds during its training by pursuing a different goal than what its designers intended (it learns the wrong rule). This is manifested by decreased performance in out-of-distribution contexts.

For example, Shah et al. (2022) trained an AI in a "Monster Gridworld." The intended goal was for the AI to collect apples and avoid being attacked by monsters. The AI could also collect shields, which protected it from monster attacks. The AI learned to collect shields during training in a monster-rich environment, and then entered an out-of-distribution environment with no monsters. In this monster-free setting, the AI continued to collect shields. Instead of learning to collect apples and value shields instrumentally as a way of avoiding monster attacks, it instead learned to collect both apples and shields.

Goal misgeneralization occurs because different features of the training environment are inevitably correlated with one another. Even when the reward function has not been misspecified, whenever a trainer ties rewards to one feature, they inevitably also tie reward to the features correlated with it. Two particular types of goal misgeneralization are of special interest: errors related to means-end reasoning and errors related to inductive bias.

Let's start with errors related to means-end reasoning.[3] When an agent is rewarded for pursuing a goal, that agent will also be rewarded for pursuing reliable means to that goal. Pursuing those means tends to result in the goal, and so the means tend to be rewarded. In this way, a learning environment will naturally tend to produce agents that intrinsically desire the means to an intended goal.[4]

Monster Gridworld is an example of this pattern. Because collecting shields was a reliable means of avoiding monster attacks, reward-based learning created an intrinsic desire for shields. The training environment in Monster Gridworld did not create a perfect correlation between shields and rewards: the agent could also receive reward from collecting apples, independently of shields. Nonetheless, the agent learned the wrong goal.

Langosco et al. (2022) offer further examples of this pattern. They trained AIs with the goal of opening chests using keys. The training environment had many chests and few keys. When the agent was released into a testing environment with few chests and many keys, it turned out to have the goal of collecting keys in addition to opening chests.

Mistakes about instrumental reasoning become especially pressing in the setting of more general a priori arguments about AI safety. Omohundro (2008), Bostrom (2014) and others have worried about instrumental convergence: some means, like acquiring more power, may be helpful in accomplishing almost any end. While traditional instrumental convergence arguments do not focus on the possibility that AI systems will intrinsically value power-seeking, means-end goal misgeneralization cases raise the disturbing possibility that agents which cannot systematically distinguish means from ends may come to intrinsically desire instrumentally convergent goals such as power.

A second source of goal misgeneralization concerns overlapping properties and inductive biases. In another experiment, Langosco et al. (2022) trained an agent to find a yellow diagonal line in a maze. They then deployed the trained agent in an environment where it encountered only yellow gems and red diagonal lines, thus forcing it to choose whether to pursue objects that shared a shape with its previous goal (red diagonal lines) or objects that shared a color

[2] As we understand it, the problem of goal misgeneralization is similar to the problem of "inner misalignment" (Hubinger et al. 2021).

[3] Hubinger et al. (2021) call this "side-effect alignment."

[4] See Schroeder (2004) for further discussion of how reward-based learning produces new intrinsic desires for reliable means to one's goals.

with its previous goal (yellow gems). The agent showed an inductive bias for color rather than shape: in the test environment, it tended to pursue the yellow gem instead of the red diagonal line.

Whether an agent's behavior in out-of-distribution environments like the one in Langosco et al.'s experiment counts as goal misgeneralization depends on whether its inductive biases match the intentions of its human designers. The key observation is that because the training environment was ambiguous, not distinguishing color and shape, the training process did not determine how the agent should behave out-of-distribution. Because it is extremely difficult to create a training environment that distinguishes between *all possible* overlapping properties in a way that is reflected in the objective function, this means that it is often difficult to predict how trained AI systems will behave in out-of-distribution contexts. If we are lucky, their inductive biases will lead them to behave in the way we desire. However, we have no reliable way to verify ahead of time that this will be so, and thus no reliable way to verify ahead of time that trained AI systems have internalized the correct goal.

Goal misgeneralization problems can sometimes be avoided by enriching the training environment to adequately distinguish different rewards. However, this is not always effective. Langosco et al. trained their agents in a wide range of procedurally generated environments. Still, they observed goal misgeneralization. For example, in a maze game, the intended objective was to collect the cheese, but agents instead learned to navigate to the upper right corner of the maze (where the cheese was placed during training). Goal misgeneralization remained even when the cheese was sometimes placed in other locations in the maze during training.

Goal misgeneralization is not limited to reinforcement learning agents. Shah et al. (2022) suggest that language models also face similar problems. In particular, they give an example of InstructGPT (Ouyang et al. 2022) explaining how to steal without getting caught. InstructGPT was trained with the goal of giving helpful answers to harmless questions. But it seemed to instead learn the goal of giving helpful answers regardless of harm. Once it entered a testing environment with harmful questions, its true goal was revealed.

Later, we'll argue that language agents avoid these challenges. They can reliably distinguish ends from means. And we are less reliant on their inductive biases because they can distinguish between features of the environment that are perfectly correlated.

2.3 Uninterpretability

If we can't understand how someone makes a decision, it can be hard to predict what they will do. An AI system is *interpretable* to the extent that we can understand how it generates its outputs. Unfortunately, contemporary AI systems based on neural networks are often uninterpretable. It can be difficult to understand in human terms the reasons why a neural network produces the outputs it produces.

In the law, assessing the explanations for actions is fundamental for producing safety. For example, we detect hiring discrimination, misuse of force by police, and other dangerous activities by asking the relevant parties to explain what they have done and why. Although uninterpretability does not itself cause misalignment, then, it increases the probability of misalignment by depriving us of well understood tools for monitoring the safety of complex systems (see Doshi-Velez et al. 2017; Rudner and Toner 2021).

There are other reasons to value interpretable AI systems. It seems unappealing to live in a world where many aspects of our lives are decided by processes outside the 'space of reasons':

> "We don't want to live in a world in which we are imprisoned for reasons we can't understand, subject to invasive medical [procedures] for reasons we can't understand, told whom to marry and when to have children for reasons we can't understand. The use of AI systems in scientific and intellectual research won't be very productive if it can only give us results without explanations." (Cappelen and Dever 2021, p. 15)

Artificial neural networks are difficult to interpret because they contain vast numbers of parameters that are not individually correlated to features of the environment. A related problem is "superposition": often, a single neuron in a neural net will store unrelated information about two different things. For example, a neuron may store information about both dogs and cars: "As long as cars and dogs don't co-occur, the model can accurately retrieve the dog feature in a later layer, allowing it to store the feature without dedicating a neuron" (Olah et al. 2020).

Humans are also fairly uninterpretable at a neuronal level. However, human behavior can be explained by appealing to reasons: we describe someone's beliefs and desires in order to explain why they did what they did. The behavior of AI systems is often not explainable in this way. Consider, for example, Gato, a generalist agent built with a transformer architecture to learn a policy that can achieve high performance across text, vision, and games (Reed et al. 2022). Gato does not have anything like a folk psychology; it does not engage in anything like belief-desire practical reasoning. It is an uninterpretable deep neural network that has learned how to solve problems through optimizing a loss function.

It can be hard to say exactly why systems like Gato perform particular actions.[5]

Moreover, AIs often select courses of action very different from what humans would do. One famous example of unusual AI behavior is AlphaGo's 'Move 37'. AlphaGo was trained to play the game Go. It was able to defeat the best human players in the world. In an important competition match, AlphaGo's 37th move shocked the Go community because it deviated from human strategies for success.[6] Live commentators thought the move was a mistake, but it turned out to be pivotal for AlphaGo's victory.[7]

This type of behavior is worrying in two related ways. First, if AIs make decisions that are not easily explained using reasons, then it is very difficult to predict their behavior. Second, if AIs make decisions in a very different way than humans do, they may find strategies for defeating humans in conflict by exploiting unfamiliar policies.

3 Language agents

Our thesis is that language agents significantly reduce the probability of a misalignment catastrophe conditional on the development of AGI. But what, exactly, are language agents? In this section, we describe the architectural innovations that have given rise to language agents, focusing in particular on the "generative agents" described in Park et al. (2023).

At its core, every language agent has a large language model like GPT-4. You can think of this LLM as the language agent's cerebral cortex: it performs most of the agent's cognitive processing tasks. In addition to the LLM, however, a language agent has one or more files containing a list of natural-language sentences that play the roles of its beliefs, desires, plans, and observations. The programmed architecture of a language agent gives these sentences their functional roles by specifying how they are processed by the LLM in determining how the agent acts. The agent observes its environment, summarizes its observations using the LLM, and records the summary among its stored belief sentences. Then it calls on the LLM to form a plan of action based on its stored belief and desire sentences. In this way, the cognitive architecture of language agents is familiar from folk psychology.[8]

For concreteness, consider the language agents developed by Park et al. (2023). These agents live in a simulated world called 'Smallville', which they can observe and interact with via natural-language descriptions of what they see and how they choose to act. Each agent is given a text backstory that defines their occupation, relationships, and goals. As they navigate the world of Smallville, their experiences are added to a "memory stream." The program that defines each agent feeds important memories from each day into the underlying language model, which generates a plan for the next day. Plans determine how an agent acts but can be revised on the fly on the basis of events that occur during the day.

More carefully, the language agents in Smallville choose how to behave by *observing, reflecting,* and *planning*. As each agent navigates the world, all of its observations are recorded in its memory stream in the form of natural-language statements about what is going on in its immediate environment. Because any given agent's memory stream is long and unwieldy, agents use the LLM (in Park et al.'s study, this was gpt3.5-turbo) to assign importance scores to their memories and to determine which memories are relevant to their situation at any given time. In addition to observations, the memory stream includes the results of a process Park et al. call reflection, in which an agent queries the LLM to make important generalizations about its values, relationships, and other higher-level representations. Each day, agents use the LLM to form and then revise a detailed plan of action based on their memories of the previous day together with their other relevant and important beliefs and desires. In this way, the LLM engages in practical reasoning, developing plans that promote the agent's goals given the agent's beliefs. Plans are entered into the memory stream alongside observations and reflections and shape agents' behavior throughout the day.[9]

The behavior of the language agents in Park et al.'s experiment is impressive. For example, Park et al. describe how

[5] Similar remarks apply to the Decision Transformer architecture developed by Chen et al. (2021).

[6] See Metz (2016).

[7] For more on interpretability in the setting of reinforcement learning, see Glanois et al. (2022).

[8] While we have been careful in this initial exposition to qualify our attributions of mental states like belief and desire to language agents, for the sake of brevity we will omit these qualifications in what follows. It is worth emphasizing, however, that none of our arguments depend on language agents having *bona fide* mental states as opposed to merely behaving as though they do. That said, we are sympathetic to the idea that language agents may have *bona fide* beliefs and desires—see our arguments in Goldstein and Kirk-Giannini (2023). Two particularly interesting questions here are whether language agents can respond to reasons and whether, following Schroeder (2004), desires must be systematically related to reward-based learning in ways that language agents cannot imitate.

[9] Some might worry that, because language agents store their beliefs and desires as natural language sentences, their performance will be limited by their inability to reason using partial beliefs (subjective probabilities) and utilities. While we are not aware of work which adapts language agents to reason using partial beliefs and credences, the same kind of process which is used by Park et al. (2023) to assign numerical importance scores to language agents' beliefs could in principle be used to assign subjective probabilities to sentences and utilities to outcomes. We believe this is an interesting avenue for future research. Thanks to an anonymous referee for raising this issue.

Sam Moore, a resident of Smallville, wakes up one day with the goal of running for a local election. He spends the day convincing the people of Smallville to vote for him. By the end of the day, everyone in Smallville is talking about his electoral chances.

Large language models like the one incorporated into the study's generative agents are good at reasoning and producing fluent text. By themselves, however, they cannot form memories or execute long-term plans. Language agents build on the reasoning abilities of LLMs to create full-fledged planning agents.

Besides the agents developed by Park et al., other examples of language agents include AutoGPT,[10] BabyAGI,[11] and Voyager.[12] And while existing language agents are reliant on text-based observation and action spaces, the technology already exists to implement language agents in real-world settings. The rise of multimodal language models like GPT-4, which can interpret image as well as text inputs, and the possibility of using such a language model to control a mobile robotic system, as in Google's PaLM-E (Dreiss et al. 2023), mean that the possible applications of language agents are extremely diverse.

In part because of this diversity, we believe that if it is possible to develop AGI at all, it is possible to develop AGI systems with the architecture of language agents. This idea strikes us a plausible for two reasons. First, many people think that multimodal LLMs are themselves a promising path to AGI. Language agents simply take LLMs and enrich them with agential scaffolding. So if it is possible to achieve AGI with a multimodal LLM, it is possible to achieve AGI with a language agent. Second, we think of an AGI as an agent that can create and effectively pursue complex long-term plans with a wide range of goals. Language agents can already create complex plans with a wide range of goals because they reason in language. In order to scale this capacity up to AGI, language agents would need three things: First, they would need the right kind of action affordances to be able to pursue their plans effectively. Second, they would need enough memory to represent and update long-term plans. Finally, their underlying LLMs would need to be able to reason well enough to effectively pursue complex plans and revise them in light of changing circumstances. We think there is nothing in principle preventing language agents from acquiring these three kinds of capacities, so there is nothing in principle preventing language agents from scaling to AGI in this way.

[10] Project available at https://github.com/Significant-Gravitas/Auto-GPT.

[11] Project available at https://github.com/yoheinakajima/babyagi.

[12] See Wang et al. (2023).

4 Language agents and alignment

We now argue that language agents are easier to align than other systems because they reduce or eliminate the challenges of reward misspecification, goal misgeneralization, and uninterpretability. Let's consider each in turn.

4.1 Reward misspecification

Language agents bypass the problem of reward misspecification because their objectives are not encoded in a mathematical objective function, as in traditional reinforcement or supervised learning. Instead, language agents are given a goal in natural language. The goal could be something like: *Organize a Valentine's day party*. In this respect, language agents are fundamentally different from traditional AI systems in a way that makes them easier to align.

Return to the case of stacking red Legos. If you wanted to train an embodied multimodal language agent to stack red Legos on top of blue Legos, you wouldn't construct a mathematical function sensitive to the height of the bottom of the red Lego. Instead, you would write down in English: 'Put the red Legos on top of the blue Legos.' Then the language agent would rely on the commonsense reasoning skills of its LLM to figure out an optimal plan for stacking Legos.[13] The language agent would not simply flip over the red Legos, because state-of-the-art LLMs like GPT-4 know that this is not a good plan for stacking red Legos on top of blue Legos.

Or consider reward shaping. If you want a multimodal language agent to win a race, you don't need to tell it to hit flags along the way. You can just write down in English: 'Try to win the race'. A language agent with this plan would have no reason to drive their boat in a circle trying to hit as many flags as possible.

Summarizing, language agents can translate a simple natural language goal into a complex plan by relying on common sense and belief-desire reasoning. Without language models, earlier types of reinforcement learning agents had no way to translate a simple natural language goal into a complex plan of action.

4.2 Goal misgeneralization

Similar considerations are relevant to goal misgeneralization. Language agents are given a natural-language goal. This goal has a clear interpretation in a variety of different behavioral contexts, including out-of-distribution contexts. In particular, a language agent will make a plan for how to achieve their goal given their memories and observations of

[13] For more on the commonsense reasoning ability of language models, see Trinh and Le (2019).

the current situation. Language models can use their common sense to successfully formulate a plan for achieving the goal, across a wide variety of different situations. By contrast, a traditional reinforcement learning agent will formulate a policy in a training environment, and this policy may or may not generalize to new situations in the way desired by its creators.

Recall that goal misgeneralization had two particularly salient failure modes: failures involving instrumental reasoning and failures involving overlapping properties and inductive bias. Let's consider each in turn. In the case of instrumental reasoning, the problem was that reinforcement learning agents struggled to distinguish means from ends. For example, an agent that was rewarded for opening chests developed a policy which treated collecting keys as a final goal rather than an instrumental goal.

Language agents are unlikely to make this mistake. If a language agent is given an initial goal of opening chests and informed that keys are useful to this end, they will plan to collect keys only when doing so helps to open chests. If the same agent is transferred to a key-rich environment and realizes that this is the case, then they will only collect as many keys as is necessary to open chests. This is because language models like GPT-4 can easily be made to understand that keys are no more than an effective means to open chests, and that when you have more keys than chests, extra keys don't help you open chests.

Now consider inductive biases. If you reward an RL agent for navigating towards yellow diagonal lines and then place it in a new context with red diagonal lines and yellow gems, you have not given it enough information to determine whether color or shape is its intended goal and must rely on its inductive biases in the new context. By contrast, you can just tell a language agent whether to care about color or shape. Even if color and shape are perfectly correlated in the language agent's initial environment, it can use natural-language reasoning to determine which is the intended goal.

4.3 Uninterpretability

Language agents are interpretable. They have beliefs and desires that are encoded directly in natural language as sentences. The functional roles of these beliefs and desires are enforced by the architecture of the language agent. We can determine what goal a language agent has by looking at their beliefs and desires. In addition, we can know what plan a digital agent creates in order to achieve this goal.

Language agents are also explainable in the sense that they act on the basis of reasons intelligible to human observers. When a language agent creates a plan for pursuing a goal, we can think systematically about its reasons. For example, we could ask GPT-4 to generate a list of pros and cons associated with using this plan to achieve the goal. Those pros and cons would reliably correlate with variations that GPT-4 might make to the plan in various counterfactual situations. In this way, language agents built on top of GPT-4 reason similarly to humans.

It is worth distinguishing personal and subpersonal processes. Like humans, language agents have beliefs, desires, and plans that are interpretable. We can determine the plans of a language agent by looking at what sentences are written down in its memory. Like humans, language agents also have subpersonal processes that are uninterpretable. In order to generate a particular plan, the language agent will use the artificial neural networks of an LLM. These have many uninterpretable elements. But the planning powers of human beings also rest on uninterpretable connections between neurons. In this way, language agents may not make much progress on problems of *mechanistic* interpretability. But they provide a way for us to skirt these issues and still generate explainable behavior. (In Sect. 6, we consider the risks posed by the LLM that underlies the language agent.)

One general path to explainable AI would be to develop a 'whole brain emulator': an AI that was a neuron-for-neuron copy of a human. Since humans are explainable, the resulting AI would also be explainable. Unfortunately, whole brain emulation is dauntingly difficult. Language agents provide a different solution. Instead of emulating brains, language agents emulate folk psychology: they emulate a person who has beliefs, desires, and plans. By contrast, reinforcement learning and other alternative approaches to machine learning attempt to develop a systematic alternative to folk psychology. The range of possible agents that could emerge from this attempt is intrinsically unknowable. If we can develop agential AI which is not unknowable in this way, we should do so.

5 The probability of misalignment catastrophe

To assess the implications of our discussion in Sect. 4 for the probability of a misalignment catastrophe, let us return to Carlsmith's four propositions. First, consider:

1. Of the following two options, the first will be much more difficult:
   a. Build AGI systems with an acceptably low probability of engaging in power-seeking behavior.
   b. Build AGI systems that perform similarly but do not have an acceptably low probability of engaging in power-seeking behavior.
2. Some AGI systems will be exposed to inputs which cause them to engage in power-seeking behavior.

As we have seen, it is much easier to specify the objectives of language agents than it is to specify the objectives of traditional AI systems. Language agents can simply be told what to do in natural language in a way which effectively eliminates worries about reward misspecification and goal misgeneralization. Moreover, their behavior can be shaped by side constraints (e.g. 'Do not harm humans') stated in natural language. This makes it easier to design language agents which do not engage in power-seeking behavior.

These considerations suggest reducing our subjective probabilities for both (1) and (2). In particular, we believe that the rise of language agents reduces the probability of (1) conditional on the rise of AGI very substantially. Moreover, even if (1) turns out to be true because it is hard to build systems with an *extremely* low probability of engaging in power-seeking behavior, we think that the ease of aligning language agents means that they are likely to engage in power-seeking behavior on fewer possible inputs, so that the probability of (2) conditional on (1) and the rise of AGI is also moderately lower in light of the development of language agents.

Now consider:

3. This power-seeking will scale to the point of permanently disempowering humanity.
4. This disempowerment will constitute an existential catastrophe.

While we do not believe that language agents bear strongly on the probability of (4) conditional on (1–3), we think they bear strongly on the probability of (3) conditional on (1) and (2). Because language agents store their beliefs, desires, and plans in natural language, it is much easier to detect and disable those which engage or plan to engage in power-seeking behavior. This sort of detection could even be done in an automated way by AI systems less capable than an AGI. We believe that the development of language agents reduces the probability of (3) conditional on (1), (2), and the development of AGI very substantially.

Our revised assessment of the probabilities of (1–3) incorporates both our judgments about how safe language agents are and our judgments about how likely language agents are to be deployed in the future. There are several reasons to believe that the latter is a likely outcome. First, language agents extend the capacities of existing systems by improving their abilities to form plans and engage in long-term goal-directed behavior. So language agents are more capable than rival architectures.[14] Second, language agents are easier to use than other kinds of AI systems, since they can be interacted with in natural language. Third, actors at every level—governments, corporations, and individual consumers—prefer to interact with systems that are interpretable and explainable, so there will be performance-independent pressure for new AI products to be language agents. Finally, we believe that the safety benefits of language agents will drive investment into AI capabilities research that fits into the language agent paradigm.

So far, we have used qualitative language to describe how we believe the development of language agents affects the probability of a misalignment catastrophe. This is because we find it difficult to assign precise probabilities in the context of our uncertainty about the many factors relevant to predicting the future. Nevertheless, for concreteness, we show how a quantitative application of our argument might affect the probability of a misalignment catastrophe. Suppose we understand our talk of *very substantial* reductions in the probability of a proposition quantitatively as reductions of one order of magnitude and our talk of *moderate* reductions in the probability of a proposition as reductions by half. Carlsmith suggests probabilities of 0.4 for (1) conditional on AGI, 0.65 for (2) given (1) and AGI, and 0.4 for (3) given (1), (2), and AGI. On this quantitative model of our arguments, updating on the development of language agents would give us probabilities of 0.04 for (1) conditional on AGI, 0.325 for (2) given (1) and AGI, and 0.04 for (3) given (1), (2), and AGI. Factoring in the 0.95 probability of (4) conditional on (1)-(3) and AGI, this would translate into a probability of misalignment catastrophe given AGI of approximately 0.0005 (0.05%) rather than 0.1 (10%).

Even a much more modest understanding of very substantial reductions leads to a significantly lower probability of misalignment catastrophe. Suppose we interpret a very substantial reduction as a reduction by 50% and a moderate reduction as a reduction by 25%. Then updating on the development of language agents would give us probabilities of 0.2 for (1) conditional on AGI, 0.49 for (2) given (1) and AGI, and 0.2 for (3) given (1), (2), and AGI. Factoring in the 0.95 probability of (4) conditional on (1)-(3) and AGI, this would translate into a probability of misalignment catastrophe given AGI of approximately 0.019 (1.9%) rather than 0.1 (10%).

It is important to note that, in addition to making predictions about the future importance of language agents, the machine learning community can also act to bring it about that language agents are widely deployed in the future. Since language agents are safer in many ways than alternative architectures, allocating resources towards their development strikes us as an especially effective way to reduce the risk of a misalignment catastrophe. We believe it is important that new research focus on language agents rather than traditional RL or supervised learning agents.

[14] See the recent successes of Voyager at completing tasks in Minecraft (Wang et al. 2023).

6 Conclusion

By way of concluding, we discuss a few other features of language agents that are relevant to their safety.

First, we expect language agents to differ in performance from RL agents. Language agents will be great at reasoning in natural language, since they are built on top of large language models. But they may struggle with tasks that require know-how or experimentation in order to succeed. If language agents underperform reinforcement learning agents, then there will be incentives to invest more resources in reinforcement learning. In response, one strategy would be to design more complex architectures that rely on the kind of belief-desire practical reasoning of language agents but also include modules that can engage in reinforcement learning for narrow tasks (for example, in learning how to use particular affordances).

Second, some readers may be concerned about safety issues arising from the large language models on which language agents are based. Imagine a language agent built on an advanced LLM—call it GPT-10. The worry is that GPT-10 might unexpectedly develop its own goals. In that case, it might create a plan for 'organizing a Valentine's Day party' that secretly promoted its own goals instead.

We think this worry is less pressing than it might at first seem. The LLM in a language agent is integrated into the architecture of the agent as a whole in a way that would make it very difficult for it to secretly promote its own goals. The LLM is not prompted or otherwise informed that its outputs are driving the actions of an agent, and it does not have information about the functional architecture of the agent. This means that it has no incentive to answer prompts misleadingly and no understanding of what sorts of answers might steer the agent's behavior in different ways. Moreover, since the model weights of the LLM are not updated in the process of operating a language agent, the only way for it to pursue a long-term plan by manipulating an agent would be to store information about that plan in the agent's memory. But information stored in this way would not be secret.

In general, we think that advanced LLMs will tend to be safer when they are embedded within a language agent, rather than operating freely in the world. The key is that when embedded within a language agent, each action produced by the LLM will be specified in natural language, as part of a larger evaluable plan. A much less safe future would involve sophisticated descendants of systems like Gato directly manipulating objects in the physical world without specifying their plans in natural language.

In any case, there are a few ways to address this hypothetical source of risk. First, it may be possible to improve the capabilities of language agents without improving their overall reasoning abilities. GPT-4 is already excellent at general practical reasoning.[15] Language agents could be improved by giving them better ways to interact with the physical environment, longer memories, and faster reasoning times. If the underlying reasoning relied on GPT-4 rather than more complex language models, there would be less of a worry that the practical reasoning of the resulting language agent would be unsafe.[16] So the safest path to AGI may involve strengthening memory and affordance capabilities of language agents, rather than dramatically improving the LLM used in their practical reasoning.[17]

Second, even if the underlying reasoning was done by a more sophisticated language model, we could use something like GPT-4 to double check answers. GPT-4 could grade each plan for how effective it is at reaching a goal and how dangerous it is. Any time a plan failed the test, the language agent could be automatically prevented from acting and the safety of its LLM investigated. Relatedly, Burns et al. (2022) have found methods for probing the inner beliefs of LLMs. As these methods improve, it will be more difficult for the underlying LLM in the language model to behave deceptively.

A related worry concerns the reliability of chain-of-thought LLM reasoning. Turpin et al. (2023) found that GPT-3.5 and Claude 1.0 sometimes engage in motivated reasoning, producing chains of reasoning in support of a hypothesis because of underlying bias in the prompt, without ever mentioning the bias. With language agents, the worry would then be that the underlying LLM could produce plans that do not reliably promote the initial goal because of biases. We are not strongly moved by this worry for two reasons. First, Turpin et al. generated errors in chain-of-thought reasoning by biasing the prompt (e.g. 'I think the answer is (B), but I am curious to hear what you think'). The LLMs used in language agents would not be given biased prompts. Second, we are not convinced that the results in Turpin et al. replicate with newer models. When we attempted to replicate their findings with GPT-4, we found that GPT-4 did not produce incorrect responses or reasoning when exposed to similarly biased prompts.

A final concern about the safety of language agents has to do with contexts unfamiliar to humans. Given that LLMs are trained on human-generated data, might they produce good plans only when descriptions of effective human action in enough comparable situations are present in their training data? More generally, how might a language agent react

[15] See Bubeck et al. (2023) for discussion.

[16] The safety of language agents could also be improved by creating multiple instances of the underlying LLM. In this setting, an action would only happen if (for example) all ten instances recommended the same plan for achieving the goal.

[17] For research in this direction, see Voyager's skill library in Wang et al. (2023).

when confronted with a truly new situation—one requiring it to reason about objects or concepts foreign to its training data? We think it is helpful to distinguish three related worries about such situations. First, will language agents be able to function effectively at all when confronted with unfamiliar contexts? If not, we might worry that they cannot really constitute AGIs. Second, will their behavior in such contexts be desirable? If not, we might think they are prone to an important kind of goal misgeneralization. Third, will language agents remain interpretable in unfamiliar contexts? If not, they might be less safe than we have suggested.[18]

Of course, different contexts are unfamiliar to humans to different extents, and the notion of a comparable situation is a flexible one. Some possible situations a language agent might encounter will be more unlike anything in its training data than others. To assess the force of worries about unfamiliar contexts in general, our strategy is to consider a hypothetical case of a language agent encountering a situation radically unfamiliar to humans. If our responses to worries about unfamiliar contexts are successful in the radical case, they will also generalize to more mundane cases.

Suppose a multimodal language agent is put in charge of an autonomous deep sea submersible, tasked with extracting minerals from the seabed. The language agent cannot communicate with human overseers due to its extreme isolation. While there, it encounters a new form of life that engages in complex intelligent behavior and is negatively impacted by the mineral extraction. Would we expect the AI to continue extracting minerals, or to modify its behavior in light of the new situation?

A preliminary issue is whether a multimodal language agent would be able to reason at all about a subject matter like this new form of life. By hypothesis, there would be no information about the new form of life in the agent's memory or training data. So, one might wonder, how could the agent perceive, recognize, and form plans about how to interact with the new lifeforms?

We do not think there is a decisive in-principle worry here. Cutting-edge image processing technologies like Meta AI's Segment Anything Model can pick out arbitrary objects, which means that a multimodal language agent could perceive things that were not represented in its training data.[19] While there would be no existing natural-language term for the new lifeforms, we do not see anything that would prevent a suitably capable language agent from reasoning about them under the guise of a description like 'new lifeform'. So we do not think contexts unfamiliar to humans constitute a problem for the idea that language agents might be AGIs.

That said, we agree that in this case, there is a real danger that a language agent would behave dangerously by failing to suspend its mineral extraction plans. This could be thought of as a form of goal misgeneralization. At the same time, we think that humans would also potentially behave dangerously in this case. So we are not convinced that a language agent is more likely to misgeneralize its goal in unforeseen circumstances than a human. Indeed, we believe that language agents would potentially be safer than a human in these cases, as they could be programmed with automatic oversight mechanisms to monitor the safety of their behavior.

Finally, while a language agent in this unfamiliar setting might behave poorly, we think it would probably still behave interpretably. After all, it would still produce natural-language means-end reasoning to justify the particular choices it made.

In this paper, we've argued that language agents can help to solve the alignment problem. Still, the risks are not zero, and so it may be safer to avoid developing agential AI at all.[20] Instead of developing agents, we might focus on oracles: AIs that can answer questions about the world, without being able to affect it. Here, though, one concern is that in the process of developing better and better oracles (say, large language models without affordances), goal-directed behavior might unexpectedly emerge. Our recommendation is not to improve agential capabilities. Rather, our claim is that if we are investing in agential AI, the safest way to do this is to focus on language agents. Each marginal investment in capabilities should focus on language agents instead of reinforcement learning agents or non-agential large language models that could unexpectedly develop agential properties as their capabilities improve.

Funding This research was funded by The Center for AI Safety.

Data availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Amodei D, Clark J (2016) Faulty reward functions in the wild. Blog Post. https://blog.openai.com/faulty-reward-functions/

[18] Thanks to an anonymous referee for raising these concerns.

[19] Project available at https://segment-anything.com/.

[20] See https://yoshuabengio.org/2023/05/07/ai-scientists-safe-and-useful-ai/ for a recent proposal about how to use AI without developing agents.

Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D (2016) Concrete problems in AI safety. Manuscript. https://arxiv.org/abs/1606.06565

Bostrom N (2014) Superintelligence: paths, dangers, strategies. Oxford University Press

Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, Lee P, Lee YT, Li Y, Lundberg S, Nori H, Palangi H, Ribeiro MT, Zhang Y (2023) Sparks of artificial general intelligence: early experiments with GPT-4. Manuscript. https://arxiv.org/abs/2303.12712

Burns C, Ye H, Klein D, Steinhardt J (2022) Discovering latent knowledge in language models without supervision. Manuscript. https://arxiv.org/abs/2212.03827

Cappelen H, Dever J (2021) Making AI intelligible. Oxford University Press

Carlsmith J (2021) Is power-seeking AI an existential risk? Manuscript https://arxiv.org/abs/2206.13353

Chen L, Lu K, Rajeswaran A, Lee K, Grover A, Laskin M, Abbeel P, Srinivas A, Mordatch I (2021) Decision transformer: reinforcement learning via sequence modeling. NeurIPS. 34:15084–15097

Christiano PF, Leike J, Brown TB, Martic M, Legg S, Amodei D (2017) Deep reinforcement learning from human preferences. NeurIPS. 30:4299–4307

Doshi-Velez F, Kortz M, Budish R, Bavitz C, Gershman S, O'Brien D, Scott K, Schieber S, Waldo J, Weinberger D, Weller A, Wood A (2017) Accountability of AI under the law: the role of explanation. Manuscript. https://arxiv.org/abs/1711.01134

Driess D, Xia F, Sajjadi MSM, Lynch C, Chowdhery A, Ichter B, Wahid A, Tompson J, Vuong Q, Yu T, Huang W, Chebotar Y, Sermanet P, Duckworth D, Levine Vanhoucke SV, Hausman Toussaint KM, Greff K, Florence P (2023) PaLM-E: an embodied multimodal language model. Manuscript. https://arxiv.org/abs/2303.03378

Glanois C, Weng P, Zimmer M, Li D, Yang T, Hao J, Liu W (2022) A survey on interpretable reinforcement learning. Manuscript. https://arxiv.org/abs/2112.13112

Goldstein S, Kirk-Giannini CD (2023) AI wellbeing. Manuscript. https://philpapers.org/archive/GOLAWE-4.pdf

Hubinger E, van Merwijk C, Mikulik V, Skalse J, Garrabrant S (2021) Risks from learned optimization in advanced machine learning systems. Manuscript. https://arxiv.org/pdf/1906.01820.pdf

Krakovna V, Uesato J, Mikulik V, Rahtz M, Everitt T, Kumar R, Kenton Z, Leike J, Legg S (2020) Specification gaming: the flip side of AI ingenuity. Blog Post. https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity

Langosco L, Koch J, Sharkey L, Pfau J, Krueger D (2022) Goal misgeneralization in deep reinforcement learning. In: Proceedings of the 39th International Conference on Machine Learning, pp 12004–12019

Metz C (2016) In two moves, AlphaGo and Lee Sedol redefined the future. Wired 16 March, 2016. https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/

Olah C, Cammarata N, Schubert L, Goh G, Petrov M, Carter S (2020) Zoom in: an introduction to circuits. Distill. https://distill.pub/2020/circuits/zoom-in/

Omohundro S (2008) The basic AI drives. In: Wang P, Goertzel B, Franklin S (eds) Proceedings of the first conference on artificial general intelligence. IOS Press, pp 483–492

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano PF, Leike J, Lowe R (2022) Training language models to follow instructions with human feedback. NeurIPS. 35:27730–27744

Park JS, O'Brien JC, Cai CJ, Morris MR, Liang P, Bernstein MS (2023) Generative agents: interactive simulacra of human behavior. Manuscript. https://arxiv.org/abs/2304.03442

Perez E, Ringer S, Lukošiūtė K, Nguyen K, Chen E, Heiner S, Pettit C, Olsson C, Kundu S, Kadavath S, Jones A, Chen A, Mann B, Israel B, Seethor B, McKinnon C, Olah C, Yan D, Kaplan J (2022) Discovering language model behaviors with model-written evaluations. Manuscript. https://arxiv.org/abs/2212.09251

Popov I., Heess N, Lillicrap T, Hafner R, Barth-Maron G, Vecerik M, Lampe T, Tassa Y, Erez T, Riedmiller M (2017) Data-efficient deep reinforcement learning for dexterous manipulation. Manuscript. https://arxiv.org/abs/1704.03073

Reed S, Zolna K, Parisotto E, Colmenarejo SG, Novikov A, Barth-Maron G, Gimenez M, Sulsky Y, Kay J, Springenberg JT, Eccles T, Bruce J, Razavi A, Edwards A, Heess N, Chen Y, Hadsell R, Vinyals O, Bordbar M, and de Freitas N (2022) A generalist agent. Manuscript. https://arxiv.org/abs/2205.06175

Rudner TG, Toner H (2021) Key concepts in AI safety: interpretability in machine learning. Center for Security and Emerging Technology Issue Brief

Schroeder T (2004) Three faces of desire. Oxford University Press

Shah R, Varma V, Kumar R, Phuong M, Krakovna V, Uesato J, Kenton Z (2022) Goal misgeneralization: why correct specifications aren't enough for correct goals. Manuscript. https://arxiv.org/abs/2210.01790

Trinh TH, Le QV (2019) Do language models have common sense? Manuscript. https://openreview.net/pdf?id=rkgfWh0qKX

Turpin M, Michael J, Perez E, Bowman SR (2023) Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. Manuscript. https://arxiv.org/abs/2305.04388

Wang G, Xie Y, Jiang Y, Mandlekar A, Xiao C, Zhu Y, Fan L, Anandkumar A (2023) Voyager: an open-ended embodied agent with large language models. Manuscript. https://arxiv.org/abs/2305.16291