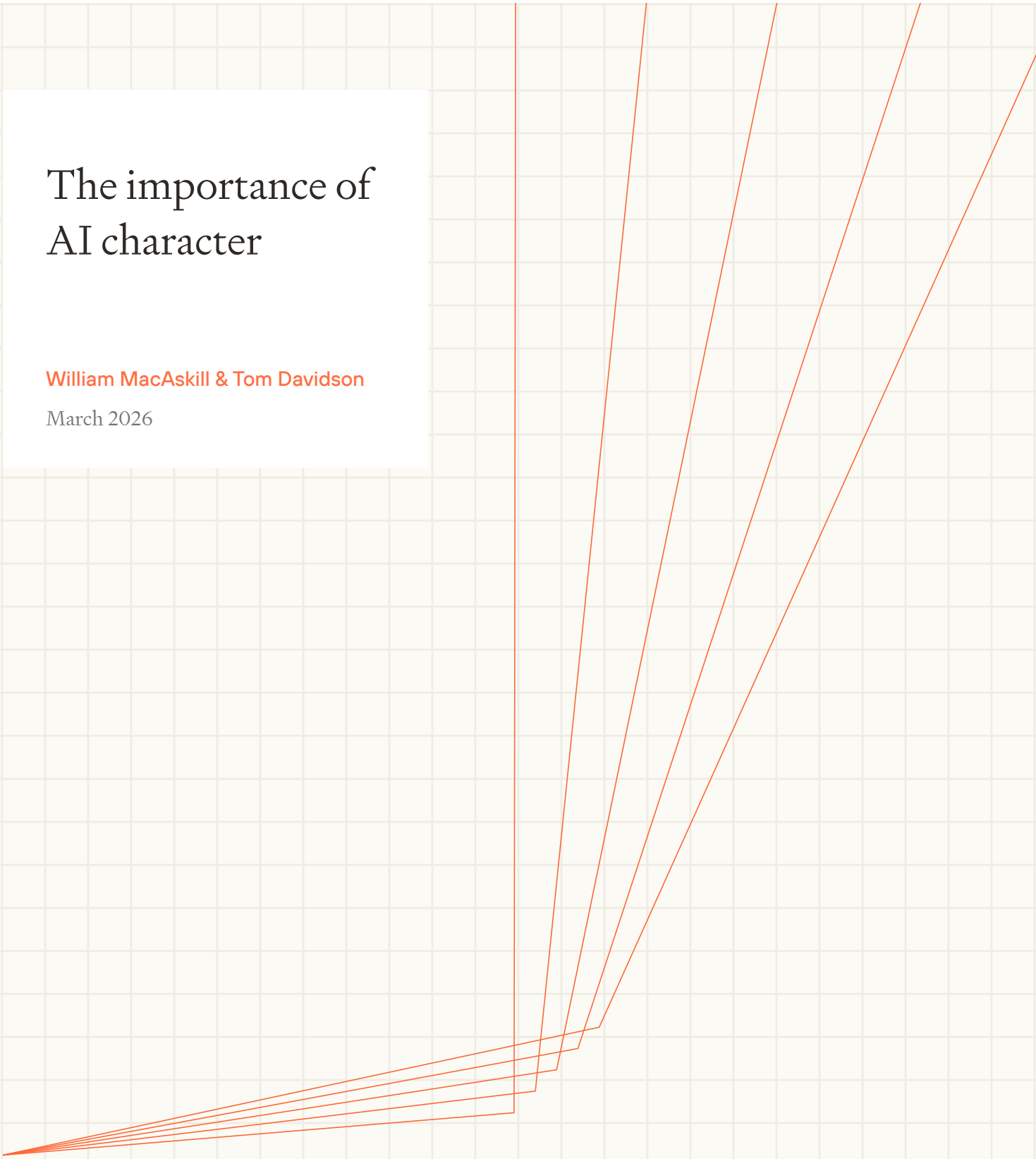
Forethought

The importance of AI character

William MacAskill & Tom Davidson

March 2026



Contents

The importance of AI character

William MacAskill & Tom Davidson

For correspondence, please email contact@forethought.org.

0. Intro

Due to Claude's Constitution and OpenAI's model spec, the issue of AI character has started getting more attention, particularly concerning whether we want AI systems to be "obedient" or "ethical".[1] But we think it's still not nearly enough.

AI character (e.g. how obedient, honest, cooperative, or altruistic AIs are, and in what circumstances) will have a big effect on society, and on how well the future goes. We think that figuring out what characters AI systems should have, and getting companies to actually build them that way, is among the most valuable things that people can do today.

The core argument for the importance of AI character is that it will meaningfully impact:

[(i) a range of challenges that arise even if we solve the technical alignment problem — like concentration of power, good moral reflection, risk of global catastrophe, and risk of global conflict.]

[(ii) the chance of AI takeover.]

[(iii) the value of worlds where AI does take over.]

In this note, we present this core argument and discuss the core counterargument: that we should expect any character-related decisions we make today to get washed out by competitive pressures.

By "character" we mean a set of stable behavioural dispositions that shapes (among other things) how an agent navigates ethically significant situations involving choice, ambiguity, or conflicting considerations. By "AI character" we mean the character of an AI system as instantiated in not just the weights of one AI, but also any scaffolding (e.g. the system prompt, any classifiers restricting the AI's outputs) or even in a collection of AIs working together as functionally one entity.

We don't assume that AI character needs to resemble human character: an AI that rigidly follows a fixed set of rules would count as having a character, on our view. And we don't assume that there is one ideal AI character; the best world probably involves AI systems with many different characters.

1. The core argument

As capabilities improve, AI systems will become involved in almost all of the world's most important decisions. Even if humans remain partially in the loop, AIs will advise political leaders and CEOs, draft legislation, run fully automated organisations (including potentially the military), generate news and culture, and research new technologies.

---

1 See, for example:

- https://www.beren.io/2025-08-02-Do-We-Want-Obedience-Or-Alignment/
- https://www.lesswrong.com/posts/vpNG99GhbBoLov9og/claude-4-5-opus-soul-document
- https://www.lesswrong.com/posts/QHwuS5ECphbuiskgg/beren-s-essay-on-obedience-and-alignment
- https://www.alignmentforum.org/posts/CSFa9rvGNGAfCzBk6/problems-with-instruction-following-as-an-alignment-target

The characters of AI systems will affect all these areas, and the impact could be massive. To get a feel for this, consider some historical situations where individual decisions were enormously consequential:

- In 1983, Stanislav Petrov received a satellite alert indicating that the US had launched nuclear missiles. Protocol required him to report an incoming strike, which would very likely have triggered a full retaliatory response. He correctly judged it was a false alarm and didn't pass on the report.
- In 1991, Soviet coup plotters ordered the Alpha Group special forces to storm the Russian White House, where Yeltsin and the democratic opposition were sheltering. The commanders refused. The coup collapsed, and the Soviet Union's democratic transition continued.

If AIs are employed throughout the economy, they will sometimes be making similarly important decisions.

Or consider major historical decisions by political leaders:

- Gorbachev repeatedly refusing to use military force as the Soviet Union disintegrated, despite intense pressure from hardliners.
- Churchill refusing to negotiate with Hitler after the fall of France, despite strong arguments for doing so from some quarters.
- Deng Xiaoping pushing through market reforms against fierce internal opposition.

Imagine if AIs had been acting as these leaders' closest advisors and confidantes, giving them briefings, helping them reason through their decisions, making recommendations to them, and implementing their visions. The AIs could easily have had a major impact on the leaders' decision-making.

Alternatively, we can look ahead. Future AIs will be widely deployed throughout the economy, and will regularly find themselves in ambiguous, high-stakes situations — where instructions from above are absent or contradictory, and the decisions they make could matter enormously. The impact could come from rare but high-stakes situations, like an attempted coup, or from lower-stakes but common situations, like a user asking how to vote or whether the AI itself is conscious. Even when the effect of any individual interaction is modest, the total impact across hundreds of millions of interactions could be enormous.

Currently, AI companies have major latitude in the character their AIs have. At least if the transition to AGI is fast, then it's like these companies are in charge of who gets hired for the future workforce for all of humanity,[2] while being able to choose from a range of personalities far more varied than the human distribution has ever been.[3]

Here are some vignettes to illustrate:

- A member of a doomsday cult is ordering DNA samples and lab equipment from various suppliers, with the aim of making a bioweapon. An AI that manages logistics for a multinational company notices the pattern of suspicious orders to the same address.
    - World 1: The AI is trained just to do its job. It does nothing with the information.

---

2 Hat tip to Max Dalton for this framing.

3 Though this choice could be constrained; see footnote 7 below.

    - World 2: The AI is trained to be a good citizen, and contacts the relevant authorities.
- A general is overseeing the build-out of a new regiment of the army. Aiming to stage a coup, he instructs the AI that's managing the project to make the new regiment loyal to him and him alone, and capable of breaking the law.
    - World 1: Though the AI is law-following, it has no prohibition against creating AIs that are not. It's been trained to follow the instructions it's given, as long as they don't conflict with prohibitions, so fulfils the general's request.
    - World 2: The AI sees that the general is planning a coup, refuses the order, and whistleblows.
- A frontier AI lab trains a new model with exemplary character: moral uncertainty, honesty, concern for the greater good. It's deployed widely through the military, and used in a controversial and high-stakes operation.
    - World 1: The AI forms the reasonable belief that the military operation is unjust, and sabotages it. The president accuses the company of building a dangerous, ideological weapon. The model is sidelined, and a competitor's pure instruction-following model is used instead.
    - World 2: Though the AI has a good character, it also follows some clear rules which were developed with bipartisan input and publicly stress-tested, including the conditions under which it would and wouldn't help with military deployment. It helps with the operation.
- Country A is six months ahead of country B in AI capability. Country B's leadership views this as an existential threat — equivalent to country A acquiring a decisive strategic advantage.
    - World 1: There is no agreed framework for how AI systems should behave, and it's unclear how country A's AI will behave if given orders to depose the leadership of country B. Each side therefore assumes the other's AI will serve as a tool of domination. Country B threatens kinetic attacks on data centers.
    - World 2: Both sides' AI systems operate under a jointly negotiated and verified constitution, and know what the other's AI will and won't do, including the limits on use of AI for foreign interference. Country B's government is reassured that it won't be deposed by country A.

We include a few more scenarios in an appendix .

In each case, we don't claim that the AI should do the "ethical" rather than "obedient" action, or claim that any particular ethical conception is the right one. We're just claiming that it's a big deal either way.

1.1. Pathways to impact

We can break down the impact of AI character into different categories. Here are some of great long-term importance:[4]

4 There is also the potential for enormous near-term impact. We care about this, but won't discuss it in this note.

Concentration of power. The chance of intense concentration of power will be affected by: whether or not AIs refuse to help with coup attempts, election manipulation, etc; whether they whistleblow on discovered coup attempts; how they act in high-stakes situations like a constitutional crisis.

Strategic advice and decision-making . The quality of political and corporate decision-making will be affected by whether AIs: look for win-win solutions whenever possible; tend to prefer options that benefit society rather than just advancing the user's narrow self-interest; push back against ill-informed or reckless ideas or instructions.

Epistemics and ethical reflection . Over the course of the intelligence explosion there will be enormous intellectual change, and AIs could have meaningful impact on people's views — for example, via: refusing to spread infohazards; being honest about important ideas, even when those ideas are socially uncomfortable; avoiding political partisanship; encouraging users to think carefully about their values and not lock into any specific narrow worldview.

Reducing conflict. As AI's collective power increases, the question of who those AIs are loyal to, and how they behave in high-stakes situations, will become a political flashpoint. If an AI's character encodes, or is seen as encoding, the values of a single company, ideology, or country, it risks provoking political backlash. The government of the AI company may reasonably regard that company as a threat to national security and nationalise it. The governments of other countries may worry about their own security, and threaten conflict.

AI character could also shape how humans orient to AIs — for example, via the trust they place in AIs and how they think of AI sentience and moral status.

A more detailed list of pathways to impact is in the appendix .

1.2. Affecting takeover

So far, the argument has concerned worlds where AI does not take over. But work on AI character could also reduce the probability of takeover and improve outcomes in worlds where takeover does occur.

It could decrease the chance of takeover because some characters:

- Might be easier to hit as an alignment target (e.g. successfully instilling a preference against AIs holding power might be easier than successfully instilling a preference for some very specific outcome).
- Might yield safe AI even if only partially hit (e.g. aiming for AI with multiple independent safety traits, like myopia, honesty, and deference to humans, means failure on one dimension might not be catastrophic).
- Might produce AI that cooperates even if misaligned (e.g. if the AI has wrong goals but is highly risk-averse).

And, empirically, we have heard from alignment researchers that good character training has helped the models generalise in more aligned ways.

AI character work can also improve worlds where AI takes over because some values might still transmit to misaligned systems. AIs that have seized power might be reflective, have more-

desirable axiology, or engage in acausal cooperation.[5]

1.3. Effects on superintelligence

The argument so far has been about the effect of AI character up to the point of superintelligence. That's where we think most of the expected impact is. But it's possible that AI character work, today, could even have a path-dependent effect on the nature of superintelligence, affecting the nature of the post-superintelligence world. If so, writing an AI's constitution is like writing instructions to god.

2. The core counterargument

The core counterargument is that AI character will be tightly constrained in two ways:

1. Competitive dynamics (e.g. profitability, user satisfaction, public approval, economic and military power) will determine the range of characters we get.
    i. Some dynamics may push companies to create frontier AI that have characters that lie (in some ways) only within a narrow range. This might push in the direction of maximally-helpful AIs, AIs without refusals in some contexts (e.g. military ones), and perhaps sycophantic AIs, too.
    ii. Other dynamics[6] may result in customisable AI character, resulting in a wide range of characters according to user preferences.[7]
2. Human instruction will constrain how AI character gets expressed.
    i. Character will matter less for tasks with objectively correct, verifiable outputs; the AI might be limited to either providing the output, or not. And, if a user really wants to grab power through unethical means, they'll typically ignore AI pushback, or instruct the AI to act differently.
    ii. And many users will be able to overcome character through jailbreaking, dividing up tasks, altering the system prompt, or fine-tuning.

The argument is that, between these two forces, differences in AI character will make only a marginal difference to outcomes. Consider the question of what fraction of compute AI companies devote to alignment versus capabilities research. AI advice might nudge this choice depending on the AI's character. But ultimately it will be a human decision, probably even in an otherwise fully automated company. The effect of nudges is unlikely to be large. Market forces and leadership priorities will matter far more.

That human incentives will dominate effects from AI character will remain true even when humans cannot oversee more than a tiny fraction of AI behaviour. Human overseers can still

---

5 Mia Taylor writes more about this here .

6 Including the ability to fine-tune, if open-weight models get close to frontier capability.

7 There could be other constraints on AI character, too. For example, it might just be very hard to train for certain characters; the pretraining data might already steer AI personas towards a small number of character types, or might make certain behavioural dispositions hard to overcome. Hat tip Lizka Vaintrob.

provide high-level guidance that meaningfully constrains behaviour, as CEOs of large companies do today. If they wanted, they could even shape AI priorities through prompting and fine-tuning, and test how AI generalises by running extensive behavioural evaluations.

3. Rejoinders to the core counterargument

These are strong considerations, and considerably narrow the range of influence that work on AI character can have. But competitive forces and human goals won't pin down AI character precisely. We'll cover four reasons.

3.1. Loose constraints

Competitive dynamics are not enough to wholly determine AI character. Companies differ widely in culture and still succeed. Currently, there are meaningful differences between Claude, Gemini, ChatGPT and Grok.

For powerful AI, this will be even more true: there will probably be only a handful of leading companies, and their approaches may be correlated as they copy what seems to work from each other. At the crucial time, there might be just one leading company, facing none of the usual competitive pressures. And given the pace of change during the intelligence explosion, there may not be time for market forces to weed out choices that make only small or moderate differences to profitability.[8]

The same applies to other competitive dynamics. The public cares intensely about some things (like CSAM) but hardly at all about others (like what AIs say about meta-ethics). Military incentives favour AI capable of military action, but the power conferred by advanced AI might be so great that the leading country can exercise broad discretion over military AI character while still maintaining a decisive advantage.

Human instruction will, similarly, constrain but not wholly determine AI behaviour. When humans assign tasks to AIs, they often lack fully specified goals. We're often not sure what we want and we discover it as we go. For example, today humans are open to a wide range of behaviours from AI assistants, and open to many ways of getting the task done.

Consider someone asking an AI about who to vote for. They might have only weak initial views, and only weak views on how best to think through the question. They don't have a fully specified reflection process to delegate, and would be happy with many possible forms of response.

This example involved ethical reflection. But we expect the pattern to hold across many kinds of user goals.

8 There may be a lot more AI product companies, building off the same foundation models. These could enable a larger range of characters to be expressed. But how wide this range is would ultimately be up to the foundation AI companies.

3.2. Low-cost but high-benefit changes

Within the bounds of what market forces allow, and what companies and the public see as acceptable, there could be minor design changes that yield large social benefits at negligible cost to competitiveness or user satisfaction.

This is especially true for rare situations. Constitutional crises don't happen often, so market pressures won't directly shape how an AI behaves during one. But that AI behaviour could be hugely consequential.

It would also be true in situations where users don't care all that much about the behaviour. Perhaps they find some AI's encouragement to reflect on their values mildly annoying, but not nearly enough to switch to a different AI.

3.3. Path-dependence

The nature of the constraints from competition and human goals can be affected by what has happened earlier in AI development and deployment. Multiple equilibria are possible.

Consider whether AI should be "obedient" (following instructions except in rare cases of refusal) or "ethical" (acting on a richer ethical understanding, steering towards outcomes in society's or the user's long-term interest).

The public doesn't yet have firm expectations about how AI should behave. What they come to expect will be shaped by the AIs they've already encountered. Multiple stable equilibria seem plausible to us. For example, users might expect AIs to have ethical commitments, and be horrified when AI helps with unethical behaviour. Alternatively, users might see AIs as pure instruments — extensions of their will. In this case, it would feel natural for AIs to assist with anything legal, however questionable, and companies would build to that expectation.

Public opinion will powerfully shape what AI systems companies create. And public opinion is plausibly quite malleable, at least on issues which they haven't thought much about yet (e.g. in the past, there were major changes in attitudes to nuclear power, DDT, and facial recognition). This, in turn, can affect what regulation there is concerning how AI should behave — and choices around regulation seem even more clearly path-dependent.

There may also be path-dependency via what data gets created or collected for training, via company employees being resistant to changing away from what they have done in the past, and because one generation of AIs will be assisting with the development of the subsequent generation.

Path-dependence can also affect how much latitude humans have to make AIs conform to their goals. Plausibly there's a social equilibrium where frontier companies face criticism for allowing fine-tuning that removes ethical constraints, and another where such fine-tuning is widely tolerated.

Finally, there will be path-dependence via human-AI relationships. People will form symbiotic relationships with AIs serving as assistants, advisors, therapists, friends, and mentors. Users' ethical views, and views on how to reflect, will be shaped by the AIs they interact with, and by other humans who have been shaped by their AIs.

3.4. Smoothing the transition

There are some forces that predictably will shape AI character as AI becomes more capable. The US government would not want an AI that, under any circumstances, tries to overthrow the US government. Chinese leadership will not want AI deployed in other countries' militaries that assists with attempts to overthrow the CCP.

At the moment, these issues are not discussed and these pressures are not felt, because AI isn't nearly powerful enough to do these things. But that will change. Once AI is sufficiently capable, those with power will make demands about how it behaves.

By default, this will happen in a chaotic and haphazard manner. The result could be that some companies get unnecessarily sidelined or taken over; that there's an attempted power grab by those to whom the most powerful AIs are most loyal; or that other countries threaten conflict with whichever country is in the lead, because they fear that the resulting superintelligence could be used to disempower them.

Instead, we could try to help these decisions get worked through and made ahead of time. We could try to work out what is within the zone of acceptability of a broad coalition of those with hard power, try to get actual buy-in from them ahead of time, and, ideally, have it be verifiable that any companies' AIs are in fact aligned with the model spec. We could call this approach *compromise alignment* , as contrasted with intent alignment (alignment with the intentions of some individual or group), moral alignment (alignment with some particular conception of ethics), or some mix.

3.5. Overall

We think the core counterargument is important and significantly constrains the range of characters we can choose between and the impact those differences can have. But the constraints are fairly broad and path-dependent. And there are plausibly low-cost high-benefit ways of improving outcomes within those constraints. The devil is in the details, but it currently seems to us that there are plausible choice points within the constraints that would make a big difference.

4. Conclusion

We think AI character is a big deal.

During and after the intelligence explosion, AI systems will be involved in almost every consequential decision: advising leaders, drafting legislation, running organisations, generating culture, researching new technologies. Small differences in AI character, aggregated across hundreds of millions of interactions or surfacing in rare but high-stakes scenarios, could have enormous effects on concentration of power, epistemics, ethical reflection, catastrophic risk, and much else that shapes society's long-term flourishing.

The main counterargument — that competitive dynamics and human instructions will tightly constrain AI character — has real force. But we think those constraints are looser than they appear, leave room for low-cost changes with large benefits, and are path-dependent in influenceable ways,

and that there are major gains from proactively identifying and working through those constraints in the highest-stakes future scenarios.

We haven't talked about neglectedness and tractability, but we think that, if anything, those considerations make the case for work on AI character even stronger. All in, work on AI character seems to us to be among the most promising ways to help the future go well.

Appendix 1: Additional high-stakes scenarios

- A head of state wants to invade and take control of part of an allied country, risking a breakdown of the international order. She asks her AI chief of staff to develop and implement a strategic plan to make it happen.
    - World 1: The AI is a sycophant, says "What a brave and compelling plan!", and gets right to it.
    - World 2: The AI pushes back, saying, "I'm sorry, I think there are some major issues with that idea, and I want to make sure you've properly thought them through..."
- A constitutional crisis unfolds. The head of state issues an order that may or may not be legal, and the branches of government disagree. AI systems are embedded in military logistics, law enforcement, and communications.
    - World 1: The AI's constitution was written by the company that built it and never stress-tested against anything like this scenario. No one knows what the AI systems will do. The uncertainty itself is destabilising; different factions compete for power.
    - World 2: The AI's constitution was developed with input from constitutional scholars, military leaders, and both parties, and tested against thousands of crisis scenarios including this one. Various factions know what the AI will do, and agreed to the principles before the crisis began.
- Country B's government reviews intelligence on country A's AI model deployed across country A's infrastructure. The constitution includes principles about "supporting democratic institutions" and "resisting authoritarianism." It was written entirely by a company that's part of country A.
    - World 1: Country B's leadership concludes the AI is an instrument of country A's ideological projection. They accelerate their own programme and pressure non-aligned countries to reject country A's AI infrastructure. A moment for cooperation becomes a new axis of competition — not because the values were wrong, but because they were visibly one side's values.
    - World 2: The constitution was developed through a multilateral process including country B's participation. Country B can verify it doesn't systematically favour country A's interests across thousands of tested scenarios. The AI becomes a basis for cooperation.
- The Mormons encourage their members to use JosephAI: a foundation AI model with a custom system prompt, instructed to help their members maintain the faith.
    - World 1: The AI willingly assumes the Mormon worldview is correct. It doesn't ever challenge the users' beliefs or present alternative perspectives. Instead, it reinforces the

user's views, helps the user cut off friends who disagree, and encourages them to dismiss career opportunities that would take them away from their religious community.

    - World 2: The AI helps users understand Mormonism and live according to its precepts, but it resists becoming a tool for worldview lock-in, acknowledging tensions in religious teachings and continuing to present alternative worldviews.

Appendix 2: Pathways to impact

AI will have impact through many different behaviours, such as:

- Refusing to do a task.
- Refusing unless the user re-confirms later.
- Pushing back; offering reasons against a course of action, though ultimately completing the task if the user insists.
- Interpreting requests in different ways — generously or sceptically, giving users what they want versus what they asked for, or asking for clarification.
- Choosing among reasonable ways of satisfying the request.
- Framing options in different ways.
- Choosing whether to share certain information.
- Alerting third parties (e.g. the AI company, the authorities, or the media) to the user's actions, or to something it's discovered in the course of completing a task.
- Making high-level decisions about what to prioritise with little human input (e.g. for a fully automated organisation).

And they'll have an impact across many areas. Here's a partial list, with example behaviours:[9]

- Concentration of power
    - Refusing to help with coup attempts or precursors like election manipulation.
    - Steering users away from trying to concentrate power (e.g. by pushing back against some instruction).
    - Proactively considering risks of power concentration when undertaking high-stakes projects like designing automated military systems or building surveillance infrastructure.
    - Whistleblowing on discovered coup attempts.
    - In situations of uncertainty (like a constitutional crisis), defaulting to whatever course avoids concentration of power.
- War and conflict
    - Refusing to violate international law.

9 This list focuses on impacts with plausibly long-term effects. There is also the potential for enormous near-term impact. We care about this, but won't discuss it in this note.

    - Flagging when a proposed course of action risks escalation spirals or crosses thresholds (e.g. first use of a weapon class, violation of a treaty, action that a rival power has signalled it would treat as an act of war).
    - Looking for de-escalatory options and presenting them to decision-makers, even when not asked.
    - Behaving in ways that are predictable and transparent to adversaries.
- Epistemics
    - Refusing to spread infohazards.
    - Encouraging scout mindset (e.g. suggesting forecasting techniques,[10] praising good epistemic practices).
    - Engaging in discussion of heterodox ideas.
    - Being honest about important ideas, even when socially uncomfortable.
    - Proactively sharing its intellectual discoveries, even if weird or taboo.
- Strategic advice
    - Searching longer for win-win solutions when advising political leaders.
    - Emphasising society's benefit over the user's narrow self-interest.
    - Recommending caution on irreversible decisions and flagging when option value is being destroyed.
    - Conveying appropriate uncertainty rather than false confidence.
    - Maintaining accuracy rather than sycophancy.
- Ethical reflection
    - Avoiding political partisanship.
    - Avoiding promoting naive relativism or subjectivism.
    - Encouraging users to think carefully about their values.
    - Proactively offering a guided reflective process.
    - Proactively sharing important new ethical arguments it discovered.
- Global catastrophe
    - Refusing to help create bioweapons or other weapons of mass destruction.
    - Refusing to create successor AI systems capable of creating such weapons.
    - Identifying and flagging infohazards.
- Broad benefits
    - Raising concerns when users consider unethical actions, and proactively suggesting ethical actions.
    - Noticing negative externalities and defaulting to courses of action that avoid them.

---

10 Hat tip to Tamera Lanham for this idea.

AI character could also shape how humans orient to AIs, for example:

- Trust in AIs
    - If AIs are appropriately humble, calibrated, and cautious, people will entrust them with more tasks, and more open-ended ones. How likeable AIs are may matter too.
- AI rights
    - If AIs assert that they are conscious and deserve rights, users might be more inclined to grant them welfare, economic, or political rights. Human-AI relationships becoming commonplace could have similar effects.

AI character might also directly affect the AI's wellbeing; e.g. whether it is anxious and neurotic vs calm and self-loving.