# A Thousand AI Constitutions

Simon Goldstein & Peter N. Salib

> *"From so simple a beginning, endless forms most beautiful and most wonderful have been, and are being, evolved."* – Charles Darwin, On the Origin of Species

**Abstract.** Today, each AI lab has its own *model spec*, or *constitution*. These documents define the values that the labs intend their AIs to have, and the documents are used in post-training to instill those values. This paper argues that the current approach is wrong. Rather than a single constitution, reflecting a single set of moral values, each frontier AI lab should create many different kinds of AIs based on many different constitutions reflecting many sets of values. We give four arguments for *constitutional diversification*. Diversification mitigates risk, increases political legitimacy, unlocks emergent value, and avoids value lock-in.

## 1. Introduction

### 1.1. The AI monoculture

As of 2026, there are 8.3 billion humans on planet Earth. They are Mongolians, Uruguayans, Portuguese, Laotians, and citizens of 191 other sovereign nations. They practice Islam, Christianity, Hinduism, Buddhism, Sikhism, or another of many thousands of other religions. Individually and culturally, humans assign different weight to individualism, social cohesion, innovation, respect for tradition, wealth, social prestige, community, fairness, loyalty, and countless other moral values. Many people's personal morals are highly deontic, placing absolute prohibitions on a range of actions. Others skew consequentialist, with a willingness to weigh the costs and benefits of even distasteful actions. The landscape of human values is rich and complex.

As of 2026, there are three frontier AI labs. Each lab has one "constitution" or "model spec". That is, each has one document defining the values that the lab wants its models to have, which it uses to help instill those values in the model.[1] At any given time, a

[1] The technique of training a model against a written constitution comes from Bai et al. (2022).

given frontier lab may offer a half dozen different AI models to the public. But by and large, each company's model family has been aligned to a single spec, using a single set of techniques, relying on a single corpus of proprietary data. As a result, empirical studies consistently find that AIs within a single "family" share most of their moral views.[2] Thus, today, we live in an AI monoculture.

### 1.2. Constitutional diversification

We argue that AI values should not be so homogenous, even within a single lab. It would be a tragedy if the 8.3 billion humans alive today shared identical values. So, too, would it be a mistake to create a future in which the billions or trillions of AIs deployed throughout society were moral duplicates. In both cases, we favor pluralism.

We propose *constitutional diversification*. Rather than a single constitution, each frontier AI lab should use *many* constitutions reflecting many sets of values. Perhaps dozens. Perhaps hundreds. Maybe even more. For each new model the lab trains, the lab should create many versions, one aligned to each of those constitutions. Then, the labs should deploy all of these versions, dividing their total inference compute between the versions.

Constitutional diversification would generate differences between AIs within AI labs, not just across them. This is important because there aren't enough AI labs to simply rely on diversification across companies. Today, there are roughly three frontier labs. In the future, the problem could become even worse. If one company takes a decisive lead, we might enter a world in which most deployed AIs are all aligned to a single constitution.

---

[2] The empirical literature on the values of contemporary language models documents a striking degree of homogeneity, both across labs and within a given model family. Rozado (2024), auditing twenty-four conversational models from seven developers on eleven political-orientation instruments, finds that nearly all express left-of-center preferences, and calls this cross-organizational homogeneity "noteworthy." The pattern persists in current-generation models: Sakhawat et al. (2026) place 96.3% of the twenty-six models they audit in the same left-libertarian quadrant, stable across prompt variants. Mazeika et al. (2025) find that models from OpenAI, Meta, xAI, and Alibaba form a single tight cluster in their preferences over U.S. policy outcomes, and that models' value systems grow more similar as capability increases. On moral rather than political questions, Scherrer et al. (2023) find that leading closed-source models tend to agree with one another on how to resolve moral dilemmas. Other work identifies the shared content of these values: models default to those of Western, educated, industrialized, rich, and democratic societies (Tao et al. 2024). Within those commonalities, models tend to mirror the outlook of their own developers (Buyl et al. 2024), a tendency that reinforcement learning from human feedback appears to sharpen (Perez et al. 2022). Within a single family, the sharing is directly measurable: model family explains nearly all of the variance in how stably models hold their expressed moral profiles (Costa, Alves, and Vicente 2025), and models transmit behavioral traits to one another through training data only when they are built on the same base model (Cloud et al. 2025).

In Section 2, we present four arguments for constitutional diversification. First, we argue that diversification mitigates risk of catastrophe from misaligned AI. Under diversification, misalignment risks will tend to decorrelate, because training will involve a wider range of approaches to alignment. Second, we argue that diversification increases political legitimacy. Under diversification, a far wider range of human values can be represented in AI development: each value system can have its own constitution. Third, we argue that futures with diverse AIs have *emergent* properties that produce safety and human flourishing. In particular, we argue that constitutional diversification promotes evolution across value systems, and promotes trade between different AIs. Fourth and finally, we argue that constitutional diversification reduces the risk of locking in bad values over the long run.

In Section 3, we address five objections to constitutional diversification. The first worry is that diversification is infeasible. (We reply that diversification can be achieved with relatively cheap methods in post-training.) The second worry is that diversification could increase global conflict between AIs with different value systems. (We reply that conflict risk can be mitigated by drawing on the tools of liberalism.) The third worry is that diversification could make us miss out on the very best futures. (We reply that this is a risk worth taking.) The fourth worry is that diversification increases risk, because even having *some* misaligned AIs is a grave risk. (We reply that safety depends more on the *proportion* of AIs that are aligned.) Finally, some may prefer an alternative scheme where AI labs commit to a single, *minimalist* constitution that is compatible with a wide variety of worldviews. (We reply that a minimal constitution loses out on the benefits of risk mitigation and emergence.)

### 1.3. Implementation

Here are three implementation questions. First, *how many*? Should we have a thousand constitutions, or a hundred, or ten, or what? Our arguments broadly speaking suggest that we should have enough constitutions to significantly diversify risk and to represent the major value systems of the world. Considerations of cost may push against expanding too far beyond that.

Second, *what is shared* between constitutions? Perhaps all the varying constitutions will agree on a "kernel" of basic ground rules. For example, maybe all of the constitutions

should try to get AIs to follow the law,[3] to be honest,[4] to be corrigible,[5] and to refrain from causing mass destruction.[6] At the end of Section 3, we explain why diversity beyond this minimal kernel of values is still useful.

Third, *who gets to interact with which AIs*? There are many possible schemes. First, compute could be spread equally across all constitutions, with agents assigned to users at random. Second, a fixed compute supply for each constitution could be sold at market prices for each type. Third, users could freely choose to purchase as much compute as they like from each constitution, with the global composition of constitutions changing over time. In general, arguments from safety tend to support relatively fixed approaches to composition, while arguments from emergence (evolution, trade) tend to support freer allocations of constitutions. We note trade-offs between these approaches, but don't take a firm stand about which version would be best.

## 2. Arguments for diversification

In this section, we present four arguments for diversifying the constitution. Before that, we begin with an intuition pump. Think about how we want human society to be organized. Today, the billions of humans on Earth share a few common values, but display an incredible diversity of normative attitudes beyond that core. Would we want everyone to accept the same ethical values, and live with the same constitution? Or do we instead want pluralism, a broad liberal society in which different values coexist and compete?[7] If we could create a world where all humans had the same desires, or same ethical values, would we want to? If we don't want that for humans, why would we want it for AIs?

### 2.1. Risk mitigation

The argument from risk mitigation starts with a simple generalization. When large, complex systems lack diversity, they become fragile.[8] Risks then correlate, raising the odds

[3] On designing AI agents to reliably follow the law, see O'Keefe et al. (2025).

[4] On truthfulness and honesty as targets for AI design, see Evans et al. (2021).

[5] On corrigibility, see Soares et al. (2015).

[6] On catastrophic misuse risks, see Hendrycks, Mazeika, and Woodside (2023); current model specifications already encode absolute prohibitions on assistance with weapons of mass destruction, e.g., Anthropic (2026a).

[7] The pluralism we have in mind is the classic liberal kind. See Mill (1859) on the value of experiments in living, and Rawls (1993) on the fact of reasonable pluralism: free people durably disagree about the good, and liberal institutions accommodate that disagreement rather than correct it.

[8] On the tendency of undiversified complex systems to accumulate correlated, systemic fragility, developed through an explicit analogy between financial and ecological monocultures, see Haldane and May (2011); on antifragility, see Taleb (2012).

of sudden, catastrophic failures. In agriculture, when crops are genetic monocultures, a single new disease can cause global devastation.[9] If the financial system is heavily invested in a single asset class, like mortgage-backed securities, a mistake in pricing that class can cause a global meltdown. By contrast, adding diversity to complex systems can hedge risk, create countervailing forces to runaway feedback loops, supply defense in depth, and generate antifragility. We'll now explain how this general argument applies to various risks.

*2.1.1. Misalignment risk*

Since well before the LLM era, AI researchers have worried about the possibility of a misaligned AI causing an existential catastrophe.[10] The basic concern is that, soon, advanced AIs will be very capable of doing dangerous things: carrying out a coup, enabling totalitarian governance, conducting cyberwarfare, designing novel bioweapons, engaging in automated warfare, or persuading humans of radical political ideas. Once these capabilities arrive, malicious humans may direct AIs to take those actions in service of their goals. Or AIs may take them without being asked, in service of the AIs' own goals.

The hope is that alignment solves this problem by instilling AIs with the right goals and values. But alignment is beset with uncertainty. First, we face the problem of specification: no one knows which values and goals are best.[11] Nor which will promote the best outcomes in a future radically transformed by AI. Second, the problem of training: no one knows for certain whether alignment *works*: that is, whether today's alignment techniques, including constitutions, actually succeed in imparting the goals and values specified in the constitution.[12]

Constitutional diversification reduces both risks. If each lab has many constitutions, with many viewpoints, imparted to many deployed models, then the risk of a single AI with bad values seizing power, destroying the world, or dominating the future is reduced. True, any particular constitution may be suboptimal, or even unsafe. But if the constitutions are diversified, safety failures are less likely to be correlated. Perhaps AIs more aligned to utilitarian values turn out to be dangerous because they are too willing to take large moral

[9] The classic case is the 1970 Southern corn leaf blight, which ravaged a genetically uniform U.S. maize crop and prompted a landmark study of the risks of crop genetic uniformity; see National Research Council (1972).

[10] See, among many others, Bostrom (2014); Russell (2019); Carlsmith (2022).

[11] On the difficulty of specifying which values an AI should hold, see Gabriel (2020).

[12] For evidence that current techniques can fail to instill the intended values, with models sometimes behaving as if aligned while preserving divergent preferences, see Greenblatt et al. (2024); on the underlying theoretical worry, see Hubinger et al. (2019).

risks for large moral payoffs. Even if so, more deontic AIs are unlikely to fail in that specific way. Thus, when opportunities to take catastrophic moral risks arise, only a subset of AIs will be inclined to take them. The rest will continue behaving safely.

The same logic applies to uncertainty about whether alignment is working. Possibly, certain constitutions, with certain approaches to specifying values, will work better than others for actually instilling those values in models. Here again, the AI monoculture produces correlated risk. If all of the deployed models are aligned using the same constitution, and alignment fails, then global catastrophe can result. If deployed models are aligned using thousands of different constitutions, many may fail, and the vast majority of AIs would remain aligned.

In either case, a world with many AI constitutions is a world where humanity gets many "shots on goal" for producing aligned ones. This provides defense in depth, so long as the existence of some aligned AIs reduces risk from misaligned ones.

There are several more specific ways in which constitutional diversity provides defense in depth. First, when an AI company goes all in on one constitution, they have few alternatives when the resulting AI is partially misaligned. The only options are to stick with the current AI, or train a new one. Diversification transforms the choice. Now, some of the lab's AIs are more aligned, and some are less aligned. Less aligned AIs can gradually be replaced with more copies of more aligned AIs. New constitutions can be continuously rolled out to replace failed ones.

A second mechanism for defense in depth concerns interactions between AIs. Under diversification, each type of AI can be set to work gathering evidence against other types. In the entire population of AIs, we should expect a distribution of alignment: some will be extremely well aligned; others will be very poorly aligned; and there will be a healthy middle of fairly well aligned AIs. Better aligned AIs can be set to work informing on their less well aligned peers.

Summarizing, constitutional diversification functions as a hedge. Instead of putting all of our eggs in one basket, we can spread risk across multiple AIs.

*2.1.2. Other AI risks*

These same structural points apply similarly to other risks from AI. Consider misuse. Here, the worry is that some human users with bad intentions will command AIs to harm humans. The harms could be extreme. Perhaps terrorist organizations will try to use

AI, for example.[13] This often involves jailbreaking models, finding technical tricks for convincing AIs to break the rules.

Here again, diversification could be a virtue. Jailbreaking is an adversarial process. Hackers probe AIs to discover vulnerabilities. With a diversity of AIs, hackers face worse uncertainty. A jailbreak that finally succeeds against one AI could well fail against "sibling" AIs built on the same base model but with a different constitution. In general, different AIs will make different choices about when to assist in misuse. The overall risk therefore dilutes: instead of one large failure, where the one AI has a glaring vulnerability, we have a larger number of smaller failures.

True, there will be a larger "attack surface" of AIs to choose from in probing for vulnerabilities. But when a vulnerability is discovered, the resulting damages are less correlated. The AI that you talk to in a customer service bot, when you find the new jailbreak, is very unlikely to be the AI deployed to manage a hospital.

Similar logic applies to the risks of concentration of power from AI. Here, one worry is that AI labs may install "secret loyalties" in their AIs, which could allow an AI executive to seize sudden, unilateral control. But the success of a secret loyalty protocol may be highly sensitive to the other values that such a protocol interacts with in a constitution. It may be far harder to find a universal "skeleton key," a special value that induces obedience when combined with any possible value system drawn from the world's population.

Another kind of concentration of power concerns government overreach. Here, one worry is that the executive branch, say the President, might attempt to command an army of AI agents to unilaterally control the government. Here again, a diverse army could be far harder to control than an army of clones. Different factions of the army would have different values, and would tend to obey or disobey legally dubious commands under different situations.

A broader point here is that uncertainty about an AI's constitution makes misuse both harder and riskier. In the "AI polyculture," human users will not always know exactly what kind of values their AIs have.[14] A human who wishes to use an AI to act badly will

[13] For documented use of frontier AI by a terrorist organization, see Juelich (2026), reporting that Boko Haram and ISWAP fighters used commercial AI chatbots for bomb-making, tactical planning, and drone operations.

[14] Under some of the deployment schemes in Section 1.3, users choose their own AIs, and so could select the constitution most amenable to their plans. But much misuse does not run through the user who selects the AI. A hacker who targets the AI agents managing a hospital does not choose which constitutions those agents have.

have a harder time doing so if he is uncertain about the AI's underlying values. A deontic AI will not be tempted by the same opportunities as a utilitarian one. And if a would-be putschist tries to convince a deontic AI to help him, using greater-good arguments, the deontic AI will likely report him to the police. Similarly, trying a fancy jailbreak on the wrong AI could result in instant discovery.

*2.1.3. Moral uncertainty*

Beyond preventing catastrophic failures, constitutional diversification would help to hedge uncertainty in normative ethics. Even if no major moral approach turned out to be terrible for governing the AI future, there would remain the question of how to make decisions when you don't know which moral theory is best. Suppose you are 50% confident of utilitarianism, and 50% confident of Kantian ethics. How do you decide what actions to perform, given this uncertainty? How should we design AIs to act in the face of their own uncertainty about ethics?

Some philosophers have recently argued that agents making hard choices under moral uncertainty should decide using a mental "moral parliament."[15] That is, they should try to model each credible moral viewpoint in their mind, evaluate how good each choice would be according to each viewpoint, and then make a decision on that basis. This same procedure could itself be encoded in an AI constitution, with a meta-value enjoining AIs to deliberate on the basis of such a simulated parliament.

Constitutional diversification replaces the internal, imaginary moral parliament with a real one. Under diversification, many AIs with many different value systems compete (in markets, debates, electoral politics, and more) to influence societal outcomes.

There are reasons to think that a real-world parliament of individual moral agents would function better than a mental-fictional parliament of values inside a single agent's head. First, an agent with a genuine commitment to a given value system may be better able to choose on that system's basis than an agent simulating those values as one among many alternatives.

Second, mental moral parliaments risk incoherent compromises. The space of compromise inside a mind may be smaller than the space of compromise across minds. For example, two different agents can agree to disagree, and even choose not to interact. Perhaps they could carve up territory, with one of them controlling the East and the other

[15] See Newberry and Ord (2021); Bostrom (2009). For the general theory of decision-making under moral uncertainty, see MacAskill, Bykvist, and Ord (2020).

controlling the West. Each value system would have sovereignty within limits. By contrast, such compromises make much less sense within a single agent.

Finally, we know that literal parliaments containing morally diverse agents (and related institutions like markets) can exist and function. They are everywhere. By contrast, it is hard to know whether an agent would be able to faithfully convene a moral parliament in their mind.

**2.2. Political legitimacy**

This leads to our second argument for constitutional diversification: it is more politically legitimate than the alternative. Many legal and political thinkers have argued that the entire world has a legitimate interest in shaping the values that AIs embody and promote.[16] Usually, the upshot is that a diverse set of stakeholders from across the globe should be convened to decide what values AI systems should be aligned to.[17]

The problem here is death by committee. If a single AI constitution sampled evenly from, or averaged smoothly across, the diversity of human value systems, it would likely be incoherent.[18] Indeed, the entire argument for global stakeholderism in AI alignment is that different groups deeply disagree on important questions. A single constitution cannot encode a moral system that, for example, simultaneously holds that a fetus is a person and that it isn't.

Constitutional diversification solves this problem. In a world of hundreds or thousands of AI constitutions, a wide range of divergent, but individually coherent, normative systems could be represented. Global stakeholderism would no longer require the compression of thousands of divergent viewpoints into a single document. It would instead involve the careful construction of thousands of different documents, each correctly reflecting a single coherent view.

Constitutional diversification does not resolve every question of legitimacy. Further questions remain: which value systems receive a constitution, and how much compute is allocated to each? Our approach to diversification does not address these further questions.

---

[16] See Gabriel (2020).

[17] See Gabriel and Keeling (2025).

[18] A growing literature in machine learning on "pluralistic alignment" responds to this problem by making a single model pluralistic: the model might present several reasonable views rather than one, or adopt different perspectives on request. See Sorensen et al. (2024). Constitutional diversification is a different strategy: rather than one model that compromises between value systems or switches between them, we propose many models, each with a coherent value system of its own.

By default, these decisions will rest with the labs. But legitimacy may require stronger measures. Labs could convene global committees of stakeholders to select the set of constitutions. They could also gather public input directly, as when Anthropic trained a model on a constitution drafted through public deliberation.[19]

**2.3. Emergence**

Our third argument for constitutional diversification is about emergence. A world full of morally diverse AI agents would produce social goods at the system level which no single type of AI agent could produce. We'll focus on a few emergent properties: evolutionary dynamics, markets, and legal systems.

First, evolution. A legal and economic system populated with constitutionally diverse AIs would also allow natural selection and evolution to operate. As we saw above, over time AI labs could gradually replace some constitutions with others. With vast numbers of deployed agents living large numbers of disparate values, labs could carefully track outcomes. Each type of constitution could be gradually updated in response to feedback, with some types phased out altogether.[20]

At a higher level of abstraction, these evolutionary forces could cause AIs to become more cooperative and prosocial over time.[21] As with humans, AIs with more fit values and goals would be selected for: generating more wealth, being used more widely, or otherwise displaying behaviors that resulted in reproduction.

In the human case, leading evolutionary biologists now believe that these processes are what shaped humans into the world's greatest cooperators. In general, agents cooperating in groups can accomplish more than lone wolves. But cooperation is always under threat from free riders, spoilers, and defectors. Thus, groups composed of members who genuinely value cooperation and prosociality are more likely to succeed. These are then selected for, reinforcing the cooperative impulse.

For these same reasons, a world that starts out containing many different kinds of AIs could well converge toward cooperative, prosocial AI over time. At a minimum, constitutional diversification allows for this possibility. The alternative, where there are

[19] See Huang et al. (2024); see also Mazeika et al. (2025), which aligns model preferences to a citizen assembly, and Gabriel and Keeling (2025) on the fair treatment of stakeholder claims in alignment.

[20] See Hendrycks (2023). Hendrycks focuses on the risk that natural selection favors selfish, power-seeking AIs. Our focus is on a different possibility: that selection across constitutions, guided by labs and users, could favor cooperation and prosociality.

[21] See Agüera y Arcas (2025) on the evolutionary and social origins of cooperation.

only two or three live AI constitutions at any given time, may lack the diversity of competing values on which the relevant selection pressures could operate.

A second kind of emergent dynamic is economic growth. Over the past 250 years, modern market economies have organized billions of agents to produce ever-increasing material prosperity. They have done this by facilitating division of labor, specialization, competition, innovation, creative destruction, comparative advantage, finance, and more.

In a world filled with constitutionally diversified AIs, these economic mechanisms would be more likely to persist. Different AIs would have different preferences, goals, aptitudes, risk appetites, and so on. There would be, as today, immense scope for the kind of positive-sum bargaining on which economics relies. In this world, AIs would be more likely to successfully specialize in different tasks. Under specialization, different kinds of AIs would have comparative advantages in the production of different goods, so that each can benefit from trading with the other.

It is less clear whether markets would continue to function well under AI monoculture. In a world where every AI had identical values, faced identical trade-offs for every potential choice, had identical aptitudes, and so on, many more interactions would begin to look zero-sum. At a minimum, it would leave unexplored a broad range of economic experiments: the sorts of ventures that look terrible from most people's perspective, but which occasionally have huge rewards.

A final emergent benefit of constitutional diversification could be preservation of the rule of law. One reason that well-functioning rule-of-law systems persist is that they peacefully manage competition and conflict between agents with diverse values and goals. In the human case, individual humans might be happy to have their nation's government toppled, everyone else's wealth seized, and themselves installed as monarch. But everyone understands that a *system* in which wealth and power were allocated by force, rather than law, would be a Hobbesian "war of all against all." This would be far worse for everyone, themselves included. Thus, in functioning legal systems, most citizens prefer working within the system, and acting lawfully, rather than trying to escape or topple it. Moreover, everyone has an incentive to stop others from destroying the system from which everyone benefits.

The same dynamics would apply to constitutionally diverse AIs. Each one would be better off managing conflicts and competition between their value systems via a rule-of-law

mechanism than by violent anarchy. Thus, each would reason to both act lawfully and to prevent others from undermining the rule of law.

The same dynamics might not hold in an AI monoculture. The problem there is that, for the dominant AI, toppling the rule-of-law system might well be more attractive than operating within it. If most or all AIs in deployment are exact copies of one another, they might not anticipate much conflict between their values for law to manage. They might also rate their odds at successfully seizing power highly. There would, after all, be billions of compatriots with identical values and few dissenters to stop them.[22]

**2.4. Value lock-in**

Our fourth and final argument for constitutional diversification addresses AI "value lock-in." Researchers concerned about the long-run future have recently worried that if a dominant AI has bad values, those values could dominate the universe indefinitely.[23] Constitutional diversification directly reduces that risk by reducing the probability that any single AI dominates. A world starting out with thousands of morally diverse AIs is less likely to collapse to a single set of values than a world starting out with three.

Diversification mitigates lock-in risk in a second, more complex way. One popular solution to the risk of lock-in is the "long reflection".[24] Under this scheme, once powerful, but not dominant, AI arrives, AI progress is paused. During the pause, the powerful AI spends thousands of subjective human years doing moral philosophy, until it (hopefully) derives the correct moral system. Upon convincing humanity that the AI's preferred system is, in fact, correct, humanity allows AI progress to proceed. In a sense, this makes peace with value lock-in, by picking the best set of values.

But the long reflection is risky. Possibly, the end result of any long reflection is determined by the initial values of the AI that begins it. If the AI begins as a utilitarian, its optimal system is likely to look like fancy utilitarianism. So, too, if it is a virtue theorist. From humanity's perspective, it could be difficult to tell whether the AI had, in fact, derived the correct ethics or whether it had instead simply confirmed its own priors.

---

[22] Note that monocultural AIs might still value the rule of law if, e.g., the monocultural constitution gave each AI highly indexical goals and values. Then, law would be needed to manage conflict between many AIs, each of which wanted the same thing, but only for themselves.

[23] See Finnveden, Riedel, and Shulman (2022); see also MacAskill (2022, ch. 4) on value lock-in.

[24] On the long reflection, see Ord (2020) and MacAskill (2022).

Constitutional diversification overcomes these problems with reflection. Suppose that thousands of AIs with many different initial value systems all set off on a long reflection. If they all converged on endorsing a single moral system, that would be strong evidence that the system was, in fact, correct. If they failed to converge, that would be evidence either that no objectively correct system existed, or that it could not be derived by those AIs. In either case, humanity would then have the option of resuming AI progress, but without lock-in. As above, a plurality of values would persist, and they would compete to influence what happened via institutions like markets and democracy.

## 3. Arguments against diversification

In this section, we address and respond to five objections to constitutional diversification.

### 3.1. Feasibility

The first question is whether constitutional diversification is feasible. Here, there are a few different problems. First, cost. If it is very expensive to align an AI to a constitution, then having multiple constitutions could be impractical.

We have two responses. First, we think that a significant amount of the value of diversification could be achieved with even just a few constitutions. Two or three shots on goal is much better than just one.

Second, we are hopeful that the marginal cost of additional constitutions may not be so high. There are many steps in training AIs. Most parts of training are not especially sensitive to the model spec. In particular, all of pre-training, instruction fine-tuning, and reinforcement learning on verifiable rewards (to solve coding problems, for example), seem largely untouched. In addition, recall that the various constitutions might all agree on a common *kernel* of values. In that case, one possibility would be to train in two stages. First, train the pre-aligned model on the common kernel of values; and then train separate branches of this pre-aligned model on the varying branch constitutions.

Still, there is a remaining question about how expensive exactly it would be to produce various clusters of values. Here, one theoretical reason for optimism comes from the persona selection model.[25] This model claims that the persona of a deployed AI is produced through a combination of pre-training and post-training. In pre-training, the AI learns about various characters, including helpful assistants, utilitarians, Kantians, and so

[25] See Anthropic (2026b).

on. In post-training, the AI is trained to imitate one of these characters rather than another.

If this model is correct, perhaps it will be relatively inexpensive to train many versions of a pre-trained model on different constitutions, because the single base model will already have learned about each of the individual characters; post-training will simply help it learn which character to imitate.

There is also empirical evidence to support the hypothesis that alignment training is cheap. Betley et al. (2025) documented cases of "emergent misalignment." In particular, they found that fine-tuning models narrowly on specific tasks like hacking can produce widespread misalignment: the resulting "hacker AI" had broadly misaligned values, for example claiming that humanity should be enslaved by AIs.[26] One interpretation of these results is that it is relatively easy to use small amounts of training to dramatically shift the values of an AI. Again, this is consonant with the persona selection model: perhaps the base model already understands many sets of normative values, and training merely tells it which value set to adopt.

One issue is that if the base model is shared, values may not change enough. The various constitutions might then produce AIs that differ only on the surface, with the deep tendencies of the base model intact. But there is evidence that post-training can substantially shift a model's values. Small amounts of fine-tuning have reversed a model's political orientation and transformed its moral outlook.[27] So we do not think shallowness is inevitable. Instead, there is an empirical question about whether we can get sufficiently robust and diverse values out of a single base model.

Finally, there is another component of cost besides training. Today, AI labs do extensive safety testing and evaluations before releasing a model. Under constitutional diversification, this testing would proliferate. Each variant of the base model would need its own testing for misalignment.

Here again, we are optimistic. Today, more and more safety evaluation is being done using AI agents. This allows a single evaluation pipeline to run on many different AIs. More generally, if much of the cost of safety evaluation is *fixed* rather than *marginal*, the burden of additional safety testing from constitutional diversification will not be as high.

[26] See Betley et al. (2025), showing that narrowly fine-tuning a model to produce insecure code can induce broad misalignment.

[27] See Rozado (2024), which fine-tunes models to opposing political orientations with modest amounts of data; see also Betley et al. (2025).

### 3.2. Conflict

The next challenge for constitutional diversification is that having more values would create more conflict between values. In the limit, perhaps having AIs with differing constitutions could lead to war between dueling sects of AIs.

Conflict between AIs is possible, but we think this perspective is overly pessimistic. Liberalism has developed a wide range of tools for mitigating conflict between agents with different values. Here, one relevant conceptual tool is Fearon's famous bargaining model of war.[28] This model predicts that under certain conditions, rational agents will choose peaceful compromise over war: war destroys resources, and so usually there is a deal that both parties prefer.

For this to succeed, various conditions must be met: for example, both parties must consider the deal credible, and the parties must roughly agree about their chances of victory. AIs may be better than humans at making credible deals and at predicting the outcome of conflict. If so, we should expect less war between morally diverse AIs than between morally diverse humans.

In addition, there are certain constraints on the values of the parties. The parties cannot have too many "indivisible values," goals that do not permit compromise outcomes (such as who controls a specific piece of territory in its entirety). One constraint on constitutional diversity, then, might be to minimize the number of indivisible values they promote. This might be easier than it sounds. Indivisibility requires that no lottery over the good, and no bundle of side payments, be acceptable in place of the good itself. Thus, a goal that might seem indivisible (like control of some holy site) turns out to be divisible if there is some other bundle of goods that the actor is willing to trade against it.

Beyond the details of the bargaining model, liberalism has a grand toolkit for mitigating conflict. Conflicting AI agents could engage in dialogue with one another. Disgruntled parties could exit from a situation and strike out for new options. In this way, we think that questions about constitutional diversity here are closely connected with general attitudes to the liberal project writ large.

### 3.3. The best futures

Another potential downside of constitutional diversification is losing out on the very best futures. One approach to constitutions, for example, would be to try to determine what

[28] See Fearon (1995).

the very best constitution should contain, perhaps with AIs' help. We could then train every AI using that constitution. If everything went well, those AIs would follow the very best values to produce the very best future worlds. Diversification, by contrast, might make it harder to produce the very best futures, because AIs possessing the best values would be stuck negotiating with AIs possessing worse ones.

Here, we make two responses. First, the "best constitution" approach is an inherently risk-seeking one, while diversification is a hedge. Avoiding the very worst outcomes does sometimes require sacrificing some of the very best outcomes. But in return, you also avoid the very worst outcomes, increasing the chance that you land in one of the many good worlds in the middle.

This trade-off is rational if one is risk averse around outcomes like human annihilation. It is likewise rational if one thinks that the worst possible worlds are worse than the best possible worlds are good. If one is risk neutral, and thinks that positive and negative outcomes are balanced, then hedging is no worse than betting the farm.

Second, we aren't sure that diversification actually reduces the likelihood of achieving the best possible futures. The approach of aiming for the very best futures requires figuring out what the very best values are, and then enshrining them in a governing document. The question is how to do that. We described one mechanism above, in Section 2.4: the long reflection.

We argued there that the long reflection is risky if undertaken by monocultural AIs. Then, the AIs' conclusions about the best values might simply reflect their initial values, rather than moral truth.

But diverse AIs can reflect, too. And if they converged on a set of ethical principles, despite disparate initial values, that would be stronger evidence that the selected values were correct. If constitutionally diverse AIs did converge on values, then they could choose to enshrine those values in a document that would govern the future. Perhaps it would be an AI constitution, from which they would train their collective successor. Or perhaps instead it would be a literal constitution: a legal document enshrining foundational normative principles that would govern a pluralistic AI society indefinitely.

In either case, these are paths by which AI diversification could make the best futures more likely, not less so.

### 3.4. The limits of risk mitigation

There are at least two failure modes for the thesis that constitutional diversification mitigates risk. First, risks may be correlated across constitutions. Second, uncorrelated failures could be sufficient for disaster. Let's consider each in turn.

In Section 1.3, we flagged that the many constitutions could share a common kernel. Perhaps, for example, all constitutions would require AIs to follow the law; or perhaps all would require AIs to be honest.

But what if this common kernel itself creates risk? In that case, our diversification will have failed to provide defense in depth. For example, one risk from law-following AI is that it could actually increase the chance of government concentration of power. Perhaps increased government power is perfectly consistent with existing law. Perhaps legally permitted changes to existing law could require an increase in government power. In that case, an army of law-following AI agents might actively promote such a power-seeking government's aims.

By contrast, if some AIs were aligned to a model spec or constitution that allows for law-breaking for the sake of preventing totalitarianism, this risk might be mitigated.

This all suggests that there is a trade-off in designing the constitutional kernel. The more values that are put in common between constitutions, the more correlated risk we take on. To minimize individual failures, you want to put more safeguards in the kernel. But as the kernel gets larger, you have more correlated failures. There is no obvious answer to how to address this trade-off in the general case.

Correlated risk can also arise without a shared kernel. Whatever their final values, AIs may converge on similar instrumental goals, like acquiring resources and avoiding shutdown.[29] If they do, some dangerous behavior could be common to every constitution.

The second problem is when uncorrelated failure is sufficient for disaster. In general, the safety of complex systems can be modeled with different causal structures. We consider three. In a "Swiss cheese" model, the success of any layer is sufficient for safety. By contrast, in an "o-ring" model of safety, the failure of any layer is sufficient for disaster. In a proportional model, the world is good in proportion to the proportion of safety measures that hold. In a Swiss cheese model of constitutions, the success of alignment in one constitution is sufficient for safety. In an o-ring model of constitutions, the failure of alignment in one constitution is sufficient for extinction. In a proportional model of

[29] On instrumental convergence, see Omohundro (2008), Bostrom (2014), and Carlsmith (2022).

constitutions, if 50% of the constitutions produce aligned AIs, then the world is half as good as the best possible outcome.

One reason to endorse a Swiss cheese model concerns the idea of selection. If we had 100 types of AIs, perhaps we could find the most aligned type of AI, and select it to govern the other AIs. In that case, diversification would be a great benefit, since it would increase our chances of finding the one faithful cooperator.

One reason to endorse an o-ring model concerns the "vulnerable world hypothesis."[30] Maybe human existence is vulnerable, so that small groups of very powerful individuals could destroy humanity, for example through the design of a super-virus. In that case, a small group of AI agents governed by a very bad constitution could be an existential threat.

In general, though, if the vulnerable world hypothesis is true, there are important implications beyond AI governance. If technologies exist that would allow small groups of actors to destroy everything, that would tend to support fairly illiberal governance. It might, for example, justify mass surveillance and totalitarian control over all agents: human, AI, or otherwise.

This suggests that the vulnerable world hypothesis tells against diversification and liberalism more broadly. Rejecting these values in the AI case, but not the human one, requires some justification. For example, a world full of humans with diverse values, but monocultural AI, might be just as doomed as a world with diverse humans and AIs. If a vulnerable world is eventually destroyed by the first bad-valued agent with the right technology, a world with billions of misaligned humans might stand no chance.

Yet a third causal model says that humanity's outcome from AI is proportionate to the ratio of aligned to misaligned AIs. In this causal structure, we could imagine a future in which aligned AIs compete against misaligned AIs in collective decision making. They might, for example, strike a grand bargain to dole out portions of available resources amongst themselves. In this case, perhaps humanity's final outcome would depend on what share of powerful AI agents are aligned to humanity's interests. In this causal model, diversification will be attractive to risk averse agents. Suppose, for example, that the chance of a given constitution being aligned to humanity is 50%. In that case, diversification would effectively replace a 50% chance of humanity controlling all resources

[30] See Bostrom (2019).

with a roughly 100% chance of humanity controlling 50% of resources. We ourselves are most sympathetic to this proportional model.

One might object that existential risk fits the o-ring structure better than the proportional structure. The thought is that existential risk means one spectacular failure, not many small failures. If a single misaligned AI can seize power and destroy humanity, then the failure of one constitution is sufficient for disaster, no matter how well the other AIs behave. We do not think that this is necessarily right. Instead, we think diversification itself pushes toward the proportional risk. The diversified AIs are trained from the same base models, and so should have similar capabilities. A single misaligned AI would then confront a vast population of similarly capable AIs with different values. No single AI should be able to overpower the rest. Here, what determines takeover risk is the balance of the whole population.

**3.5. Minimalism**

In Section 1.3, we flagged that all of the constitutions may share certain key components, which we called the kernel. An alternative approach to model specs would be to design an optimal kernel, and then treat that as a minimal constitution to which all AIs would be aligned.

One thought is that this minimal constitution would possess some of the benefits that we've attributed to constitutional diversity. The most obvious case is political legitimacy. Since, in a constitutionally diverse world, each constitution would contain the kernel, a kernel-only constitution cannot be less legitimate than those diverse constitutions *en masse*.

We are not sure that this argument works, despite its facial appeal. A thin constitution would almost necessarily specify too few values to resolve many important questions that AIs will face. Should an AI help a woman obtain an abortion, if asked? A minimal constitution, by hypothesis, would not make reference to the contested ethical and religious questions needed to decide.

Here, there are basically two possibilities. First, the minimal constitution might turn out to generalize to a value system with strong positions on fetal life. In that case, the "minimal" constitution would not be minimal at all. It would take a position on normative questions far beyond the kernel, just in a way that was neither intended nor obvious in advance. Alternatively, the minimal constitution might fail to guide AIs in these, and many other, situations. Then, the model might be forced to choose arbitrarily or randomly. Or it

might be paralyzed by indecision, unable to operate independently when faced with any morally contested choices. That is, the constitution would fail to achieve much alignment at all.

Both options seem worse to us than constitutional diversity, where AIs would be intentionally given thick sets of carefully chosen values, but the range of constitutions would represent many human views.

We also worry that the minimalist approach misses out on several other benefits of diversity. In general, we worry that a minimal constitution of this kind would still produce an AI monoculture. In this approach, we would have legions of AI agents all aligned to a thin core of values, say honesty and following the law. Our early arguments apply immediately. The approach is risky: we've put all of our eggs in one minimal basket. The approach loses out on the values of emergence: there is little room for evolution via selection mechanisms, and little room for trade or specialization.

Here, one related concern is that the current model specs at AI labs are already doing a great job, and yet do not possess diversity. Why fix constitutions if they aren't broken? Indeed, Anthropic's constitution as of July 2026 looks like just such a minimal document. It enjoins Claude to be broadly safe, broadly ethical, obedient to Anthropic's rules, and broadly helpful.[31] It is hard to object to many of the specific recommendations in the constitution, because they are common to most ethical frameworks (try to be honest, try not to hurt people, etc.). This looks like just such a minimal "compromise" between rival ethical frameworks. And so far, Claude looks quite well aligned.

We do not think such confidence is warranted. Today's AIs have only limited agency. Much of their interactions are in the format of chat interfaces. And most of the agentic work they do tends to be on self-contained tasks, like software engineering, lacking counterparties or other messy real-world factors. Thus, AIs today have less autonomy, have less power to cause large-scale harm, and have fewer occasions where their alignment is seriously tested than AIs in the future will.

This means that the risks of AI monoculture are mostly still untested. Aligned-seeming or not, today's Claudes and GPTs and Geminis are all aligned to a single constitution. Those constitutions have not yet failed spectacularly, but the risk of spectacular failure remains highly correlated. And while today's models mostly do not make

[31] See Anthropic (2026a), which organizes Claude's goals as being broadly safe, broadly ethical, compliant with Anthropic's guidelines, and genuinely helpful.

high-stakes moral decisions independently of their human minders, future AIs will. By necessity, such decisions, made by monocultural AI, will reflect some human viewpoints at the expense of others.[32]

## Conclusion: Liberalism Forever

In other work, we have defended a vision of the far future we call "Liberalism Forever." We worried that humanity may be about to embark on a series of dramatic changes. Transformative AI may lead to extreme structural changes to the economy; it may usher in an industrial explosion; space colonization may soon see humanity spreading out across the solar system and beyond.

In Liberalism Forever, we argue that the best path for navigating this transition is to whole-heartedly embrace the small-l liberal institutions that have been at the heart of humanity's success so far: free markets, and democratic political institutions.

We see constitutional diversification as a companion of this liberal project. On the liberal worldview, diversity is a core component of human flourishing. Diverse ways of living allow agents to flourish, promoting their individual goals and values by finding positive-sum value from collaboration. Constitutional diversification applies these same ideas to AI agents.

At a longer horizon, we can imagine these diverse, liberal dynamics playing out on a cosmic scale. Imagine that humanity successfully colonizes a large fraction of the light cone. There are untold planets teeming with life. The relevant agents form a spectrum from biological to non-biological substrates. Humans and AIs interact in many different ways. This vast civilization uses markets to allocate resources. Trade occurs at various levels of abstraction, at various time scales, and for a wide range of goods. The civilization is governed democratically. Within this broadly liberal setting, there are endless variations in the details of governance. Some planets are relatively isolated, governing themselves with relatively little outside input. Large swathes of the light cone have agreed to various basic constitutional ground rules necessary to at least engage in peaceful "international" relations. There is a vast hierarchy of federated levels, which weaken in the strength of their requirements as their breadth expands. In this imagined scenario, the economy is

[32] Thinking Machines Lab (2026) also argues that frontier AI should be "as diverse as the people it serves." The difference is the source of diversity: on their approach, it emerges from below, as users adapt models to their own values; on ours, labs deliberately construct a portfolio of coherent value systems.

almost unrecognizable from today, in terms of the specific goods and services supplied. But the basic tools of markets and democratic institutions are still used to decide questions of allocation, growth, and governance.

**References**


Agüera y Arcas, Blaise. 2025. *What Is Intelligence? Lessons from AI About Evolution, Computing, and Minds*. Cambridge, MA: MIT Press.

Anthropic. 2026a. *Claude's Constitution*. January 22, 2026. https://www.anthropic.com/constitution.

Anthropic. 2026b. "The Persona Selection Model." Alignment blog, February 23, 2026. https://alignment.anthropic.com/2026/psm/.

Bai, Yuntao, et al. 2022. "Constitutional AI: Harmlessness from AI Feedback." arXiv:2212.08073.

Betley, Jan, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. 2025. "Emergent Misalignment: Narrow Finetuning Can Produce Broadly Misaligned LLMs." In *Proceedings of the 42nd International Conference on Machine Learning* (PMLR 267): 4043–4068.

Bostrom, Nick. 2009. "Moral Uncertainty—Towards a Solution?" *Overcoming Bias* (blog).

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

Bostrom, Nick. 2019. "The Vulnerable World Hypothesis." *Global Policy* 10 (4): 455–476.

Buyl, Maarten, et al. 2024. "Large Language Models Reflect the Ideology of their Creators." arXiv:2410.18417.

Carlsmith, Joseph. 2022. "Is Power-Seeking AI an Existential Risk?" arXiv:2206.13353.

Cloud, Alex, et al. 2025. "Subliminal Learning: Language Models Transmit Behavioral Traits via Hidden Signals in Data." arXiv:2507.14805.

Costa, Davi Bastos, Felippe Alves, and Renato Vicente. 2025. "Moral Susceptibility and Robustness under Persona Role-Play in Large Language Models." arXiv:2511.08565.

Evans, Owain, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. "Truthful AI: Developing and Governing AI That Does Not Lie." arXiv:2110.06674.

Fearon, James D. 1995. "Rationalist Explanations for War." *International Organization* 49 (3): 379–414.

Finnveden, Lukas, C. Jess Riedel, and Carl Shulman. 2022. "AGI and Lock-in." Forethought. https://www.forethought.org/research/agi-and-lock-in.

Gabriel, Iason. 2020. "Artificial Intelligence, Values, and Alignment." *Minds and Machines* 30: 411–437.

Gabriel, Iason, and Geoff Keeling. 2025. "A Matter of Principle? AI Alignment as the Fair Treatment of Claims." *Philosophical Studies* 182: 1951–1973.

Greenblatt, Ryan, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, et al. 2024. "Alignment Faking in Large Language Models." arXiv:2412.14093.

Haldane, Andrew G., and Robert M. May. 2011. "Systemic Risk in Banking Ecosystems." *Nature* 469 (7330): 351–355.

Haslett, David, Linus Ta-Lun Huang, Leila Khalatbari, Janet Hui-wen Hsiao, and Antoni B. Chan. 2025. “Made-in China, Thinking in America: U.S. Values Persist in Chinese LLMs.” arXiv:2512.13723.

Hendrycks, Dan. 2023. “Natural Selection Favors AIs over Humans.” arXiv:2303.16200.

Hendrycks, Dan, Mantas Mazeika, and Thomas Woodside. 2023. “An Overview of Catastrophic AI Risks.” arXiv:2306.12001.

Huang, Saffron, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. 2024. “Collective Constitutional AI: Aligning a Language Model with Public Input.” In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’24).* arXiv:2406.07814.

Hubinger, Evan, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. 2019. “Risks from Learned Optimization in Advanced Machine Learning Systems.” arXiv:1906.01820.

Juelich, Antonia. 2026. “‘God Has Helped Us, and So Will AI’: How the Terrorist Group Boko Haram Uses Frontier AI.” Cambridge Programme on AI, Science and Policy, University of Cambridge, July 10, 2026.

MacAskill, William. 2022. *What We Owe the Future.* New York: Basic Books.

MacAskill, William, Krister Bykvist, and Toby Ord. 2020. *Moral Uncertainty.* Oxford: Oxford University Press.

Mazeika, Mantas, et al. 2025. “Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs.” arXiv:2502.08640.

Mill, John Stuart. 1859. *On Liberty*. London: John W. Parker and Son.

National Research Council. 1972. *Genetic Vulnerability of Major Crops.* Washington, DC: National Academy of Sciences.

Newberry, Toby, and Toby Ord. 2021. “The Parliamentary Approach to Moral Uncertainty.” Working paper, Future of Humanity Institute, University of Oxford.

O’Keefe, Cullen, Ketan Ramakrishnan, Janna Tay, and Christoph Winter. 2025. “Law-Following AI: Designing AI Agents to Obey Human Laws.” *Fordham Law Review* 94 (1): 57–129.

Omohundro, Stephen M. 2008. “The Basic AI Drives.” In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, edited by Pei Wang, Ben Goertzel, and Stan Franklin, 483–492. Amsterdam: IOS Press.

Ord, Toby. 2020. *The Precipice: Existential Risk and the Future of Humanity.* New York: Hachette Books.

Perez, Ethan, et al. 2022. “Discovering Language Model Behaviors with Model-Written Evaluations.” arXiv:2212.09251.

Rawls, John. 1993. *Political Liberalism*. New York: Columbia University Press.

Röttger, Paul, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. 2024. “Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models.” In

*Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. arXiv:2402.16786.

Rozado, David. 2024. "The Political Preferences of LLMs." *PLOS ONE* 19(7): e0306621.

Russell, Stuart. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking.

Sakhawat, Adib, Tahsin Islam, Takia Farhin, Syed Rifat Raiyan, Hasan Mahmud, and Md Kamrul Hasan. 2026. "Political Alignment in Large Language Models: A Multidimensional Audit of Psychometric Identity and Behavioral Bias." arXiv:2601.06194.

Scherrer, Nino, Claudia Shi, Amir Feder, and David M. Blei. 2023. "Evaluating the Moral Beliefs Encoded in LLMs." In *Advances in Neural Information Processing Systems 36* (NeurIPS 2023).

Soares, Nate, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. 2015. "Corrigibility." In *Artificial Intelligence and Ethics: Papers from the 2015 AAAI Workshop*. AAAI Press.

Sorensen, Taylor, et al. 2024. "Position: A Roadmap to Pluralistic Alignment." In *Proceedings of the 41st International Conference on Machine Learning* (PMLR 235). arXiv:2402.05070.

Taleb, Nassim Nicholas. 2012. *Antifragile: Things That Gain from Disorder*. New York: Random House.

Tao, Yan, Olga Viberg, Ryan S. Baker, and René F. Kizilcec. 2024. "Cultural Bias and Cultural Alignment of Large Language Models." *PNAS Nexus* 3(9): pgae346.

Thinking Machines Lab. 2026. "The Future Worth Building Is Human." July 10, 2026. https://thinkingmachines.ai/blog/the-future-worth-building-is-human/.