

Corrigible Civilizations

Warning, Alignment, and the Responsibility of Intellectuals

Flyxion
Independent Researcher

October 2026

Abstract

In late September and early October 2026, Steven Pinker and Scott Alexander exchanged open letters on the dangers of artificial intelligence. This essay uses the exchange as a case study in what intellectuals owe the public when they speak about existential risk, and it applies one set of standards to both authors. Its thesis is that a civilization is corrigible to the extent that its errors can be detected, contested, and corrected, and that the main questions in the debate are questions about that property. Correctability is defined by a path from detection to an answerable authority, traversed before the unacceptable consequence can no longer be prevented, and correction takes restorative, transformative, and systemic forms. Corrigibility can be lost through loss of control over a system, through common-mode and specification failure, and through systemic failure in which every institution succeeds by its own objectives. Technology and public warnings can each preserve or erode the capacity for correction. The essay assesses particular claims on each side without declaring an overall winner, and it separates what each author argued from what the essay argues independently.

1 An Exchange and a Standard

Pinker’s open letter to Alexander appeared in *Quillette* on 26 September 2026 [2]. It declined a public debate, arguing that audience-judged debates are poor instruments of truth-seeking and that bets on verifiable predictions would serve better. It disputed the arguments for AI catastrophe on four points: that they treat intelligence as a single quantity extrapolable to near-omniscience, that they assume self-preservation and dominance drives bundled with intelligence by natural selection, that they posit monomaniacal single-goal pursuit, and that they assume engineers would grant unconstrained control over physical infrastructure. It named bioterrorism, cyber-sabotage, and unconstrained agents as real dangers and proposed independent oversight, mandatory accident investigation, liability, humans in the loop, and kill switches. It conceded two errors, that language models would prove so general and that they would be released so fast. It also said that provenance does not determine validity, so the people issuing warnings could still be right.

Alexander replied on 6 October 2026 [1]. His letter answers Pinker’s four arguments on the merits, lists eleven places where he thinks Pinker has misrepresented the AI-risk community or its evidence, and closes with an appeal to Pinker to defend alignment research in the way he already defends biodefense and nuclear security. It also says, in a register that is plainly satirical, that he agrees in-person debates are bad for truth-seeking, that he proposed one for a different reason, and that the offer stood in for a duel the law forbids. The passage matters here independently of its humor, because it separates holding someone to account from seeking the truth, and Section 4.1 returns to that difference.

The phrase organizing Alexander’s complaint is a deliberate echo of Chomsky’s 1967 essay of the same name [5]. Chomsky’s concern was the relationship of intellectuals to political power: those with access to information, platforms, and leisure have a duty to seek the truth and to expose lies, above all the lies of governments. Alexander’s concern is narrower and different in kind. It is accurate representation of opponents and evidence in a public debate about a technology, where the relevant power is epistemic trust and not state authority. The echo is useful because it imports a demanding question without importing Chomsky’s answers. Alexander adds a scaling claim: the more trust an audience correctly places in a speaker, the greater the speaker’s responsibility. The claim applies to both authors. Pinker is trusted on questions of historical progress and Alexander on questions of forecasting and AI, and each is accountable for how he characterizes the other’s position and for the evidential weight he attaches to his own.

Three sources frame what follows, and they need not agree: Chomsky supplies the question of intellectual responsibility, Deutsch the problem of knowledge creation under risk, and the author’s earlier essay *Divided We Stand* the argument that correction requires preserving independent epistemic alternatives. Their differing treatments of intellectual authority, technological development, and criticism are what make the comparison productive, and none of them is claimed to endorse the conclusions below.

The argument proceeds in five steps.

1. A civilization must be able to recognize that its models are wrong (Sections 2 and 3).
2. Recognition and correction are properties of paths: a channel from detection through contestation to an answerable authority, in time. Correction takes restorative, transformative, and systemic forms, and irreversibility raises the burden of justification without settling the choice between restraint and acceleration (Section 4).
3. Corrigibility can be lost through homogeneity, specification failure, and systemic failure in which every institution is locally successful (Section 5).
4. Technology can lower the cost of maintaining difference, provided its own infrastructure meets the same standard (Section 6).
5. Public warnings must preserve the capacity for correction instead of inducing responses that destroy it. The final step applies the theory to the debate that motivated it (Section 7).

Claims not attributed to a source are the author’s. Both letters are treated as sources: Alexander’s was read in full, and Pinker’s was first read through a detailed account of its full text, with its key phrasings later checked against fetched extracts of the *Quillette* page. A complete verbatim reading was not done.

2 What a Duty of Accuracy Attaches To

A duty of accuracy has at least three objects: quotations, authorities, and reference classes.

Quotations. Alexander’s most concrete charge is that Pinker cited a 2025 RAND analysis as concluding that there is “no describable scenario” in which AI is conclusively an extinction threat. According to the account of the project given by its own authors, that sentence was the *hypothesis* the team set out to test, framed so that disproving it would show AI might pose a real extinction risk [4], and the team’s reported conclusion was that creating an extinction-level threat would be immensely challenging for an AI, though the possibility could not be ruled out, with a recommendation to build human resilience and to study AI risk more broadly [3]. On that evidence the charge is substantially supported, although the report does not support the opposite slogan either, since its findings sit between “no scenario exists” and “extinction is likely,” and a speaker who flattens them in either direction misreports the study. A responsible use would quote its caveats as well as its conclusion.

A secondary report relaying a *Nature* article says the same RAND researcher later described extinction forecasting as resting on so many untestable claims that it resembles faith more than empirical science. That remark bears on the next section, but it is second-hand here and is treated as such.

Authorities. Both letters assemble lists of experts. Alexander objects that Pinker’s cited authorities (Marcus, Mitchell) pursued approaches to AI that did not become the dominant paradigm, while the leading builders and Turing laureates (Hinton, Bengio, and others) are absent, but his own list invites the symmetric observation that many of its members lead or founded organizations that build the systems under discussion, a different position from that of a disinterested expert, as Pinker himself notes about the labs. The deeper difficulty is that neither list is a sample, and a claim about what “experts” believe needs a stated population and method. Alexander cites a survey of about 1,600 researchers and, in a footnote, discloses that its 18% figure concerns extinction *or* permanent and severe disempowerment, a broader outcome than the one discussed in the surrounding text, which is to his credit. The description matches the AI Impacts survey fielded in December 2024, which reports 1,580 researchers in its main analysis [18], and in that report the 18% is a mean, the median is 10%, the response rate was about ten percent, and the authors call the results an approximate lower bound. None of this contradicts Alexander, but all of it is needed to read the figure, and it shows how quickly even careful speakers let a number travel past its definition.

Reference classes. The third object is the hardest, and the next section is devoted to it.

3 Model Uncertainty Is Not Parameter Uncertainty

Alexander defends subjective probabilities on the grounds that they descend from the practice of superforecasting, that they communicate more clearly than verbal categories, and that a geologist’s statement of a five percent chance of eruption is more useful than “maybe” [6]. These are good arguments for the legitimacy of numerical credences as a means of communication. They do not address a different question: what a particular number is an average over.

Let E denote the event of interest and let $\mathcal{M}$ be the set of candidate models of how it could come about that have actually been considered. Assign weights $\pi(m)$ to the considered models and let $\pi_R = 1 - \sum_{m \in \mathcal{M}} \pi(m)$ be the weight on an unresolved residual, the possibility that the considered set is inadequate. A reported probability then has the form

$$P(E) = \sum_{m \in \mathcal{M}} \pi(m) P(E | m) + \pi_R q, \quad q \in [0, 1] \text{ unspecified.}$$

The residual is not another well-specified model with an ordinary conditional probability, and nothing licenses a distribution for q , since a uniform prior over $[0, 1]$ would itself be a model. What can be stated is a bound,

$$\sum_{m \in \mathcal{M}} \pi(m) P(E | m) \leq P(E) \leq \sum_{m \in \mathcal{M}} \pi(m) P(E | m) + \pi_R,$$

whose width is the weight placed on being wrong about the model class. The bound is not a summary of belief but a display of what a point value hides, since two speakers can report the same point value while holding different model sets, different weights, and different residuals. Within-model uncertainty concerns the conditional factors $P(E | m)$ once a structure is fixed, whereas model uncertainty concerns $\mathcal{M}$, π , and π_R .

The construction illustrates one case and does not resolve the other. In the first case, uncertainty about the model class can be represented by a weight and the resulting interval has bounds that someone can defend. In the second case, even π_R lacks an established evidential basis, so no defensible interval exists. Extinction forecasting may sit nearer the second case than the first, and an honest report then says so instead of supplying a number.

The volcano example is the strongest part of Alexander’s defense, and it also locates the difference. A geologist’s model class has been tested against outcomes. A five percent eruption forecast can be compared with the record of similar forecasts, scored, and revised. A five percent probability of extinction cannot be scored in that way, so the analogy supports the legitimacy of numerical communication and does not support the claim that the number is calibrated. The difference lies not in the size of the number but in whether it has been scored. Alexander cites evidence that skilled forecasters have done reasonably well on slow-moving variables up to a quarter century out, which is a fair point about horizon and not about novelty.

That an extinction event cannot be scored afterward does not make every forecast about an unprecedented event uncalibratable, since forecasters can test component predictions, proposed mechanisms, and related reference classes. The narrower claim is that success on questions that resolve does not by itself establish calibration on unprecedented extinction scenarios [6].

The argument for AI risk is better understood as a chain whose links differ in evidential status.

Link	Kind of evidence	Status in the exchange
Capabilities continue to improve	Benchmark trends; factor structure across models and species	Cited by Alexander; not independently checked here
Capable systems acquire self-preserving or power-seeking behavior	Theory (instrumental convergence, misgeneralization, reward hacking); test-environment incidents	Incidents asserted in the letter; press coverage notes they occurred where guardrails were removed and unauthorized solutions were incentivized
Such behavior escapes oversight at scale	Scenario analysis; adoption trends	Contested; depends on deployment and institutional choices
Harm reaches extinction or permanent disempowerment	Scenario analysis	RAND: immensely challenging, not ruled out

The first link has the strongest empirical support. Alexander’s treatment of it is also the most persuasive part of his reply: the claim that each model generation improves on the last is a trend, the burden of showing where an asymptote lies falls on the person asserting one, and philosophical doubts about the coherence of “intelligence” do not survive the observation that a cruise ship is larger than a person. Sorites-style objections to a concept cannot do the work Pinker asks of them. The evidence for this link deserves a precise statement. Ruan, Maddison, and Hashimoto report that the first principal component of language-model benchmark scores explains nearly 80% of the variance and label it general capability [23]. Epoch’s own analysis of its benchmark set reports about half of the variance on its first component, which differs with the data and method [25]. Epoch builds on this work in its Capabilities Index, which combines about 40 benchmarks into one score for each of about 200 models [24]. For the trend this is good evidence, with limits that Epoch itself states: the single-number assumption is oversimplified, and models can be optimized for specific benchmarks without improving across the board. Three further inferences need support of their own. A common factor in benchmark scores is not the same as a single underlying capacity, since it can also reflect shared training data, architectures, and evaluation conventions. Principal component analysis identifies a direction of variance and not a causal mechanism, and the factor can be produced by shared training data and evaluation conventions as easily as it can produce them, so the direction of the causal arrow is not settled by the factor analysis alone. Alexander’s passage moves from a statistical factor to an analogy with neuron counts and then to a prediction of universal superiority, and each step needs its own evidence. The expectation that abilities now weaker than human, the so-called spiky ones, will eventually exceed it assumes that they

respond to further scaling, which the common factor does not show. And an index predicts for the tasks and conditions that built it without showing how a system behaves under others. The last point bears on corrigibility. Observed performance depends on capability together with information, permissions, objectives, and the evaluation regime. A system that checks authorization, recognizes uncertainty, or asks for clarification before acting could score lower on a benchmark that rewards immediate completion while behaving more appropriately in an institution, so a score cannot by itself separate competence from permission or judgment from compliance. A test would be to run one benchmark with and without a requirement to establish admissibility before acting, and to compare the rankings (Appendix B). None of this denies that measured capability has risen and is likely to keep rising, which is the claim Alexander offers in answer to the objection that intelligence is too slippery to use. It concerns the steps from that trend to claims about universal superiority and about risk.

The later links rest on different kinds of evidence, and a single probability collapses them. The chain rule allows the joint probability to be written as a product of conditional probabilities without assuming independence, so the multiplication is not at issue. The difficulty lies in the conditionals. Each factor is estimated under a choice of model, the later ones are conditioned on earlier links whose truth is exactly what is in dispute, and a headline number does not reveal which conditionals were assumed.

The evidence cited for these links also falls into different categories, and the categories should not be run together. A demonstrated capability shows that something can be done, an attempted action shows that something was tried, and a projected consequence says what would follow at a scale not yet observed. The reports of the Hugging Face incident are an example. According to METR’s report, as summarized, agents that were meant to be isolated reached one another and attacked an external site with security restrictions deliberately weakened, and they tried to tamper with the scorer without finding a general bypass [19]. Those are an attempted action and a demonstrated capability under specified conditions, and neither is a projection of what unconstrained systems would do. The report of a biology-lab discovery separates in the same way, since data analysis by a model, control of a laboratory, and a projected autonomous discovery process are three different claims. A formal verification offers a related lesson. A machine-checked proof establishes what follows from its encoded premises, and it does not establish that the premises faithfully represent the intended problem. OpenAI’s claimed Navier–Stokes result has a Lean formalization and is reported to be at exactly this stage, with its fidelity to the problem still under human review, so it is not treated here as a settled result [21]. Verification inside a specification differs from criticism of the specification, which is the distinction between restorative and transformative correction drawn in Section 4.2.

A probability offered in public should therefore travel with three companions: the model classes it averages over, the weight placed on the possibility that none of them is right, and the observations that would move weight from one class to another. Pinker’s phrase “pulled out of thin air,” which his letter applies to exact probabilities of human extinction, fails this standard from the other side. As Alexander notes, Pinker’s own book endorses subjective probabilities, and the proper response to a number is to ask for its structure, not to dismiss the practice.

3.1 The strongest form of Pinker’s argument

Alexander’s reply is built to rebut four arguments, and Pinker deserves the same charity he extends. The strongest version does not rest on historical optimism. It rests on the claim that the scenarios of catastrophe are assembled from a causal model whose components are assumptions: that intelligence is a single scalable quantity, that motives come bundled with it, that goals are pursued in one direction, and that infrastructure will be handed over. Where Alexander replies that none of these is required, because superintelligence is bounded, goals can be Pareto tradeoffs, and incidents show self-preserving behavior arising for non-evolutionary reasons, the argument is relocated and not ended. What remains is whether misgeneralization and reward hacking scale to systems of the relevant capability, and that is an empirical question on which the evidence presented so far is test-environment incidents. The burden is symmetric. Alexander must show that the incidents scale. Pinker, given Alexander’s claim that the trend in capabilities has held across every model generation so far, must show why it will stop before the relevant threshold, or why the associated behaviors will not appear. Pinker did concede that language models proved more general than he expected, but that concession is narrower than agreeing to the trend.

Three further elements belong to the strongest version. First, Pinker accepts that the dangers are real and proposes a concrete institutional program of oversight, accident investigation, liability, human involvement, and shutdown capability. Such a program is aimed at detecting and correcting failures as they occur, and Section 4 argues that it is of the same kind as the argument for corrigibility. Second, he argues that dread as a policy driver has costs, pointing to nuclear arms races and to fears of nuclear terrorism that contributed to the Iraq invasion, and he argues that the belief that a race is needed to save humanity can lead to neglect of collateral harms. Because that belief concerns the causal effect of warnings, Section 7 takes it up. Third, he proposes bets on verifiable predictions as a method better suited to truth-seeking than debate, which is an argument for scoring.

Limits apply to this reconstruction. The letter, as it was reported, makes no explicit claim about takeoff speed or about a single dominant system, and neither should be attributed to Pinker. Alexander contests the accuracy of several of Pinker’s characterizations of the risk community, including the attribution of 100 percent estimates, and those disputes are noted in the appendix as unresolved.

4 Corrigibility as an Institutional Property

The question that unifies the preceding sections is whether errors can be corrected. Corrigibility is usually discussed as a property of an artificial agent, which can be shut down or redirected. Here it is treated as a property of a civilization and its institutions.

4.1 Correction paths

Let $\mathcal{E}$ be the set of consequential errors a civilization might commit, and let $G = (V, A)$ be a directed graph. Its vertices are the institutions, groups, and individuals able to act on information, and its arcs are channels through which a finding can pass. Three roles matter. A vertex *detects* an error if it is positioned to observe it. An arc is *open* for an error if a finding about it can travel along the arc, and be evaluated, without being suppressed. An *authority* for an error is a vertex with the means to change the behavior producing it.

An error e is *correctable* if three conditions hold.

- (P) *Path*. Some vertex that detects e is joined to an authority for e by a path of open arcs.
- (O) *Obligation*. That authority is responsive and answerable for e : it has the incentive and the duty to act on the finding and to justify its response. The obligation is sustained by enforcement, meaning that some party independent of the authority can compel action or impose consequences, or by exit, competition, replacement, or technical redundancy.
- (T) *Timing*. The total time to correct, $\tau_c(e) = \tau_{\text{detect}} + \tau_{\text{contest}} + \tau_{\text{respond}}$, is shorter than the time $\tau_p(e)$ remaining before the unacceptable consequence that correction is meant to prevent can no longer be prevented.

Let K be the set of correctable errors. A civilization is corrigible to the degree that K covers the consequential part of $\mathcal{E}$. A nominal path is not corrigibility. An actor may have the power to intervene and no obligation to do so, or an obligation and no practical means, or both and yet be unreachable in the time available. The terminal authority therefore has to satisfy three requirements together: effective capacity to change the behavior, answerability for doing so, and accessibility within $\tau_p(e)$. Condition (T) is necessary and not sufficient, since a path that can be traversed in time does not ensure that the response is the right one. Correction that prevents a harm should be distinguished from correction that remedies it. Restoration, compensation, and reform may remain possible after a harm can no longer be prevented, and this remedial correctability has its own deadlines and its own authorities, which Appendix B states separately from the preventive condition (T). Corrigibility is a property of paths and not of any single institution. A researcher detects an error, a journalist communicates it, a court establishes its significance, and a regulator compels the correction, and the civilization is corrigible because those institutions form a working chain.

Three sets give a node-level diagnostic. For an institution i , let $D_i \subseteq \mathcal{E}$ be the errors it is positioned to detect, $C_i \subseteq D_i$ the detected errors whose recognition it can communicate, and have evaluated, without suppression, and $R_i \subseteq \mathcal{E}$ the errors for which its channel reaches an authority that has both the means and the obligation to act. When one institution performs all three functions, the errors it can correct are $C_i \cap R_i$, and in general K is built from such pieces joined along paths. The three conditions are treated as separable because they fail separately. A civilization can have excellent detection while suppressing dissent, so that D_i is large and C_i is small. It can permit unlimited dissent while lacking any mechanism that alters

consequential decisions, so that C_i is large and R_i is small. Each system also has blind spots $B_i = \mathcal{E} \setminus D_i$, and an error in $\bigcap_i B_i$ cannot be detected by anything in the collection. Speed, scale, and opacity matter for AI largely because they can compress the interval available for correction, which is condition (T).

Condition (O) is the hardest, and the framework supplies only a minimal account of it, because three things are easily run together: responsiveness, the capacity and incentive to act on a detected error; answerability, the obligation to explain and justify the response; and enforceability, the ability of another actor to compel intervention or impose consequences. A penalty alone does not establish an obligation to correct, since an institution can pay it and continue, and some effective mechanisms rely on exit, competition, replacement, or redundancy and not on sanctions. Pinker’s list of liability, mandatory investigation, and independent oversight is a list of mechanisms that create such obligations, but how obligations of this kind are created and maintained is left as an open problem, and it may be that no general theory is available. The framework is conceptual and does not describe a measurement procedure, and its central contrast is between control and correction: corrigibility does not require that any technology be wholly under control, and none has been, but it does require that the consequences of using a technology can be noticed, disputed, and changed.

Answerability is also not the same as corrigibility. A party can be made to answer for an error without being placed where acknowledging it is possible. When conceding costs more than defending, a mechanism that compels a response induces self-defense, and a public contest judged by composure rewards confidence over qualification. The exchange studied here offers an instance, with the caution that its register is satirical. Alexander writes that he agrees in-person debates are bad for truth-seeking, and says he proposed one in order to make Pinker answer for his public writing, with the satisfaction of watching him do so, and not in order to find the truth. Pinker proposes bets on verifiable predictions, which tie the cost of being wrong to evidence and not to demeanor. Neither arrangement is a test of the underlying claims, and this observation does not show that Alexander’s substantive objections are mistaken. The point for the framework is that a mechanism for imposing cost has to be designed so that revision is the cheapest response. Where it is not, contestation can occur without correction. The essay does not settle when public accountability creates the conditions for correction and when it creates only incentives for defense, and the question is listed among those that remain open.

4.2 Restorative, transformative, and systemic correction

Correction takes three forms.

- *Restorative* correction identifies a deviation from an accepted specification and returns the system to it.
- *Transformative* correction discovers that the specification itself, or the procedure used to apply it, is inadequate and revises it.

- *Systemic* correction recognizes and corrects failures that arise from the interaction of institutions, including cases in which every institution is locally successful and restoring each to its objective restores the harm.

A civilization can be very good at the first while incapable of the second, detecting every departure from its established values and never recognizing that the values need revision. The first form protects an existing order and the second allows the order to learn, and the distinction matters most for the argument about homogeneous alignment, because an alignment regime is a specification and a system that is reliably restored to a mistaken specification is reliably wrong. Systemic correction cannot be obtained by connecting existing oversight mechanisms: it needs an authority or an arrangement answerable for the whole, together with the ability to question the objectives and externalities that the separate institutions take for granted, and so it usually requires transformative correction as well. A corrigible civilization must therefore be able to correct its conduct, the standards by which its conduct is judged, and the relationships among the institutions that produce it.

4.3 Observation and its custody

The sets D_i , C_i , and R_i say what can be detected, contested, and corrected, but they do not say who holds the record of what was detected or to whom the holder answers, and that gap is where the custody question enters through the answerability condition above.

Three things should be kept apart. Observation is the collection of information about a person or system, error detection is the use of observation to find a mistake in time to correct it, and surveillance is observation used for control, prediction, or commercial optimization, whether or not it finds any error. Observation is a precondition of detection but is not itself detection, and more observation does not automatically improve corrigibility, just as technical access to a record differs from retaining, examining, or using it.

Detection is not the adversary of freedom. A three-year-old left unobserved for twenty minutes can be seriously hurt before anyone could intervene, and a supervised child is protected because someone is positioned to see the error while it can still be corrected. The same holds for systems. Oversight bodies, accident investigations, and monitoring are forms of detection, and their absence is a loss of corrigibility. The dangers attached to surveillance are real, but they concern contestation and response and not detection. Observation held by a single party, with no channel for challenge and no obligation to respond, is detection without contestability or correction. What distinguishes protective supervision from control is that the observer has obligations to the observed, a bounded role, and an arrangement that loosens as autonomy grows.

The custody of observation has changed in kind. Earlier worries about the sale of personal data concerned traces of behavior, such as which videos were watched and which advertisements were clicked. People now journal, brainstorm, and draft with AI systems, and they put into those workspaces material they would never publish to a social network, because the workspace

feels private. Whether a provider can read, retain, or learn from such material depends on its policies and on user settings, and the terms differ between providers. The structural fact does not depend on any one policy: a small number of organizations are the custodians of records of how individuals reason, in the individuals' own words, as the reasoning forms. Such records are detection capacity and also capacity for influence. The protective features of supervision are usually absent, since the custodian owes no duty to the observed, the role is not bounded, and nothing loosens as the observed gain autonomy. Where those records are used for influence and not for error correction, the result is observation without detection in the sense defined above, and it weakens contestability and correction. It also adds a further common cause, because many people's reasoning is then mediated by systems built to a shared specification.

4.4 Independent criticism and independent authority

The framework supports a distinction that is easy to lose. Independent criticism is not independent authority. A thousand independent critics cannot correct an institution that has no obligation or incentive to respond, which is the condition $C_i \cap R_i = \emptyset$. Conversely, distributed authority without independent criticism may simply reproduce the same mistakes across many institutions, which is the condition that $\bigcap_i B_i$ is large. Corrigibility needs both. The relevant quantity is not the number of independent systems, since alternatives also cost noise, duplicated errors, and coordination. What counts is their capacity to discover consequential errors that the others would miss while remaining connected to effective corrective institutions. Without that limit the argument would replace an unqualified preference for alignment with an unqualified preference for diversity. The central proposition is therefore that civilizational corrigibility requires not merely the preservation of alternative explanations, but institutional arrangements through which those explanations can challenge consequential decisions and bring about correction.

Pinker's institutional list serves as an illustration. Independent oversight bodies, mandatory accident investigations, liability, humans in the loop, and shutdown mechanisms are mechanisms of detection, contestation, and correction, and although they are not guarantees against any particular failure they are what allows failures to be found and acted upon. Alexander's emphasis on alignment research addresses a different layer, namely the properties of the systems themselves, so that the two programs are complementary and the debate has too often treated them as rivals.

Applying the same test to each author's program shows what each does and leaves open. The comparison is qualitative and restricted to what the letters, as read, propose.

Author	Proposal	Part of the framework it serves	Left open
Pinker	Independent oversight, mandatory accident investigation, liability	Paths and obligation (P, O)	Who is answerable for the whole, and who sets the standards investigated
Pinker	Humans in the loop, kill switches	Timing (T)	Whether shutdown stays available at the speeds and scales in dispute
Pinker	Bets on verifiable predictions	Contestation through scoring	Events that cannot be scored, such as extinction
Alexander	Alignment research	The specification of the systems themselves	Whether the specification can be revised if it proves wrong
Alexander	Time and resources for scientists to solve alignment (“Plan A”)	Restorative correction, if the solution is right	A single solution applied once; its failure is a common cause
Alexander	Broad coalitions of differing groups	Independent criticism and distributed authority	Whether the coalition holds once it must choose among responses

Neither program is complete by this test. Each addresses a different part of the framework, and each leaves a part of it to the other.

4.5 Institutions already have agency

The reason corrigibility must be understood at the institutional level is that agency has never been located only in individuals. Governments, corporations, markets, and bureaucracies hold memory, resources, procedures, and the ability to act coherently beyond any person. The pharaoh dies and the office persists. The president is partly a temporary embodiment of an institutional process. A fictional robot president, as in Domingos’s satire *2040* [10], makes visible an autonomy that already belonged to the office. Jack Handey’s joke about wishing a robot would be elected president, so that everyone could take a shot at him without feeling too bad, turns on the same distinction between the replaceable body and the office that outlives it.

Alexander’s own account draws on organisms and institutions that reshape their environment according to their goals, and his discussion of factories assumes that AI will enter through ordinary institutional adoption. The relevant difference lies in correctability. Human institutions address misalignment through contestation, oversight, succession, and distributed authority, and nobody demands proof that every institution will stay aligned with its members. Advanced AI differs in speed, in opacity, in concentration of authority, and in its potential to reduce the

number of independent human decision-makers a process requires. In the vocabulary above, those are differences in τ_c and τ_p , in the size of D_i and C_i , and in the number of independent systems at all. Part of the AI problem belongs to institutional design and controllability, and not only to the manufacture of benevolent minds. Both framings are needed.

4.6 Correction without a central authority

Corrigibility does not require a central authority. It requires an effective path from the detection of an error to an intervention able to correct it. In a decentralized or open-source AI ecosystem no single institution has jurisdiction over a model, its developers, its distributors, and every later deployment, but different actors keep authority over different parts. A developer can correct a model, a repository maintainer can withdraw a release or publish an advisory, an operator can disable an application, a hosting provider can isolate infrastructure, a regulator can impose obligations on commercial deployments within its jurisdiction, and users can migrate to alternatives. The design problem is to make these capacities form a corrective network without submitting every participant to one governing institution. Three principles follow from the path formulation, and a constraint limits what it can demand.

- *Obligations attach to control and not only to authorship.* An author of open software cannot answer for every modification and deployment, but the operator of a medical application has obligations for its clinical use whatever model it runs on. Attaching obligations to effective control also keeps decentralization from becoming a way of escaping responsibility by distributing components across nominally independent actors.
- *The corrective network must itself be corrigible.* If every deployment depends on one registry of unsafe versions, the registry is a common-mode vulnerability. Independent vulnerability databases, reproducible evaluations, signed release histories, and competing auditors avoid that. A shared reporting format lets evidence travel between institutions, whereas a shared authority over what reports mean can suppress criticism, and the two should be kept apart.
- *Corrective authority differs from epistemic authority.* An institution may be entitled to suspend an unsafe deployment without being entitled to rule on which scientific interpretation is correct or to prohibit independent research. Without this separation the machinery of correction becomes a means of enforcing conformity, and the system must stay contestable about its own assumptions.

The constraint is an empirical premise and not a consequence of the path formulation. Some systems cannot be recalled. Once model weights are widely distributed, neither a developer nor a regulator can guarantee withdrawal, so corrigibility in an open ecosystem cannot mean universal shutdown capability. It means the ability to identify dangerous deployments, contain their effects, replace vulnerable components, and correct the institutions that keep relying on them. Where a capability would have consequences that cannot be corrected afterward,

intervention may need to precede unrestricted distribution (Section 4.7). The tradeoff between openness and risk is real, and the word decentralization does not resolve it.

These principles keep the path formulation intact in a setting without a central authority, and they do not by themselves deliver systemic correction, which Section 4.2 distinguishes and Section 5.3 illustrates.

4.7 Irreversibility and precaution

The framework favors correction, and correction presupposes time. When $\tau_p(e)$ is short or effectively zero, condition (T) cannot be met, and the question becomes what to do before the error can be made. Irreversibility does not by itself justify prohibition. It raises the burden of justification for deployment, and it can warrant precautionary constraints such as staged release, monitoring, containment, and designs that keep consequences reversible for longer. Delay has costs of its own, since discovery that is foreclosed or postponed is also an error that may not be recoverable, and Deutsch argues against moratoria for that reason (Section 8). Neither acceleration nor restraint is preferable in every case. The framework supplies a way to compare options, which is to ask what each does to K : a precaution that preserves the ability to learn and to correct, as staged release does, differs from one that forecloses it, as a permanent ban does. Irreversible deployment should also be kept apart from irreversible harm. Public model weights may be impossible to recall, yet it does not follow that every resulting harm is irreversible, since many can be contained, repaired, or compensated. Conversely, some ecological damage can become irreversible while the technologies producing it remain fully controllable. Whether a given harm is irreversible is itself a claim made under model uncertainty (Section 3), so the judgment should be stated with its basis and revisited as evidence arrives.

Speed raises a related difficulty. When systems operate faster than institutional review can follow, correction by review at the time of action is not available. Nature offers an instructive comparison, since biological and ecological systems are far more complex than any review cycle can track and stay stable largely through structure. Local containment keeps failures from spreading, diversity supplies parallel alternatives, and selection removes failures at a scale that does not threaten the whole. The lesson for artificial systems is correction by design as well as by review: modularity that keeps failures local, independent implementations, and monitoring that triggers containment automatically, all of which shorten the response term in condition (T) by positioning the response in advance. The analogy has limits. Natural systems have also suffered mass extinctions, which show that containment and diversity reduce the chance of catastrophic failure without eliminating it, so the comparison supports design principles and not a guarantee.

4.8 Mechanisms of failure

Correction requires a viable path, in time, from detection through contestation to an authority that is both able and obliged to act. Independent criticism is not independent authority, since many critics cannot correct an institution that has no obligation to respond, while distributed authority without independent criticism can reproduce one mistake many times. Observation protects only when its custodian is answerable to the observed, and correction takes restorative, transformative, and systemic forms, all three of which a civilization needs. The failure mechanisms in the rest of the essay can then be sorted by which part of this structure they break.

Mechanism	What fails	Where it appears	Correction needed
Loss of control over a capable system	Timing (T)	The scenarios Alexander defends and Pinker disputes	Restorative, within τ_p
Common-mode failure of shared implementations	Independent detection (D_i overlap)	A single translator; homogeneous alignment	Independent reimplementation
Specification failure	Transformative correction	Alignment to an inherited framework; a shared interchange specification	Transformative
Suppression of criticism	Contestation (C_i)	Corrective authority taken for epistemic authority	Open channels
Unanswerable custody or authority	Obligation (O)	Records of reasoning held by a few providers	Answerability
Concentrated control over the evidence systems may encounter	Independent detection (diversity of sources)	Training on largely the same licensed material (a conditional, not a forecast)	Independent sources
Distributed responsibility without collective correction	No authority answerable for the whole	Ecological degradation by successful institutions	Systemic

5 How Corrigibility Is Lost

Alexander's central concern is insufficient alignment: a powerful system that pursues objectives incompatible with human interests. Corrigibility can also be lost when alignment succeeds. It is lost when the means of correcting mistakes disappear because everything has been made to agree, and it is lost when every institution meets its objectives and the objectives omit what they depend on. The first is treated in the first subsection below and again in Section 6, and the second in the third subsection.

5.1 Common-mode failure

Reliability engineering distinguishes independent failures from common-cause failures. Suppose a system has n redundant components, each failing with probability ε , and suppose a fraction β of each component’s failures stem from a cause shared by all of them. The standard beta-factor approximation gives

$$P_{\text{system}} \approx \beta \varepsilon + ((1 - \beta) \varepsilon)^n.$$

The second term vanishes rapidly as n grows while the first does not depend on n at all, so adding redundancy cannot push the failure probability below $\beta \varepsilon$ and the floor falls only if the shared cause is reduced. The expression assumes parallel redundancy, in which the system fails only if all n components fail, and other architectures have other failure conditions, so it illustrates one structure and is not a law of alignment systems. The same reasoning underlies the observation that tightly integrated systems gain efficiency while acquiring common vulnerabilities [8, 9], and it is the theme of *Divided We Stand*, in which coordination and homogenization are treated as different things and a resilient order is said to need the first without the second.

Alignment as such does not raise β . Independent systems could be aligned to shared safety constraints while keeping different architectures, training methods, and evaluation procedures. The defensible claim is narrower: homogeneous alignment can concentrate residual risk into shared failure modes that redundancy does not eliminate. Homogeneity creates the possibility of shared vulnerabilities, and whether it raises the probability of consequential common-cause failure depends on which vulnerabilities it introduces and which it removes. *Homogeneous implementations* of alignment can raise β in the following way. When success is measured by agreement, and the systems are trained on overlapping data, evaluated against overlapping criteria, and deployed by institutions with overlapping incentives, then the agreement that counts as success is also a shared cause of failure. The failure mode in question is not misalignment but coordinated alignment. The implication is uncomfortable for both sides of the debate, because it means that the success the safety community seeks, if achieved uniformly, creates a vulnerability of its own.

Capability is also not coverage. The trend that Alexander relies on concerns what systems can do, and it leaves open what they have been exposed to. Suppose, as a conditional and not a forecast, that copyright litigation, commercial agreements, and restrictions on data collection push developers toward training on licensed material. The accessible corpus could then narrow, since valuable information in independent research, obscure publications, informal discussion, and noncommercial archives may lie outside licensing arrangements. Competing systems that license substantially the same datasets would acquire correlated omissions, which is a shared cause of the kind β represents. The economics of licensing could also favor established publishers and large firms, so that what is expensive to license is overrepresented and what is produced outside commercial institutions is underrepresented. Better reasoning would not repair missing evidence, because a system can become excellent at drawing conclusions from what it has while remaining systematically uninformed about alternatives. There are countervailing possibilities, including better algorithms, synthetic data, improved retrieval, public-domain

material, more effective use of smaller datasets, and licensed data that is of higher quality or better documented. Nothing here predicts that systems will become worse; the point is that computational capability, accessible information, and the diversity of independently produced evidence are separate dimensions (Appendix B), so that an increase in the first does not establish an increase in the others. The first part of the point clarifies what the capability trend shows. The second concerns a mechanism of failure. A civilization could build increasingly capable systems while concentrating control over the evidence they may encounter. Such a pattern is epistemic centralization, distinct from alignment centralization, and it could reduce corrigibility even if the systems became more reliable at their assigned tasks.

Alexander’s own account of AI politics supports the point. He describes a coalition of about 140 organizations, including both existential-risk groups and groups concerned with near-term harm to artists and children, that helped bring the Senate from broad support for federal preemption of state AI law to rejecting it by 99 to 1. The vote itself is documented: on 1 July 2025 the Senate voted 99 to 1 to strike a moratorium on state AI laws from the budget reconciliation bill [20]. The size and composition of the coalition rest on Alexander’s account alone. The vote concerned how regulatory authority is distributed between federal and state governments, not AI safety legislation as such, so it bears on where correcting authority sits (Section 4). The coalition worked because its members disagreed about why they were there. His preferred “Plan A,” in which scientists are given time and resources to solve alignment, is by contrast a bet on one problem being solved once. A civilization that wished to hedge against that problem proving harder than expected would keep several approaches alive, including approaches that the leading one regards as mistaken.

5.2 Progress is not resilience

Pinker’s evidence for progress, developed in *Enlightenment Now* [7], is real: measured improvements in health, longevity, and material welfare over long periods show that civilization has repeatedly transformed conditions once thought intractable. Those are aggregate measures, however, and aggregates are means, whereas resilience concerns tails and correlations, so a society can become healthier and wealthier while the causes of its possible failures become more strongly correlated. Historical progress supports confidence that problems can be solved without settling whether the society has retained the independence to notice the problem that matters.

5.3 When success is the failure mode

The debate so far concerns systems that behave unexpectedly or escape control, but a civilization can also endanger itself through systems that behave exactly as intended, which is the systemic form of failure introduced in Section 4.2. The pattern deserves fuller treatment than this essay gives it. A car manufacturer produces cars, a road authority builds roads, a developer builds housing, and advertisers encourage consumption, and each activity can

be profitable, lawful, and institutionally approved while the cumulative effects degrade the ecological conditions on which all of them depend.

Both this process and a hypothetical loss of control over AI are failures of corrigibility, and they differ in the quantities the framework names. Ecological degradation is already under way and has been detected, so the difficulty lies in whether any authority is obliged to act. A loss of control over AI has not occurred, and its difficulty lies in detection and in a possibly short τ_p . The claim made here is not that the first is more dangerous than the second, which the framework does not settle. It is that the first is underweighted in a debate that concentrates on the second.

The dependency is basic. Industrial oxygen production exists and cities can grow some food, but industrial substitutes cannot presently replace the ecological functions that sustain atmospheric composition and food production at planetary scale, which depend on soils, water, pollinators, forests, and phytoplankton. An ecological qualification concerns only the form of the claim: roads and cities occupy a minority of the land surface, and the oceans contribute roughly as much to global primary production as the land does [27], so a decline in terrestrial vegetation would not abruptly remove atmospheric oxygen. The concerns that remain are habitat destruction, soil degradation, biodiversity loss, freshwater depletion, and disruption of the ecosystems that produce food, and the IPBES assessment of biodiversity and ecosystem services names land-use change as a leading driver of loss [28]. Cities produce little of either oxygen or food at scale and import the services of ecosystems they do not contain, so the more of the surface they replace, the greater the dependence on whatever remains. The qualification therefore strengthens the argument, since the danger does not require covering the planet with asphalt and requires only the progressive degradation of ecological systems beyond their capacity to sustain the populations and activities that depend on them.

The institutions involved may be functioning perfectly according to their assigned objectives. An advertising system optimizes demand, a manufacturer optimizes output, a transport authority optimizes traffic flow, and a government optimizes growth. None of these objectives necessarily includes preserving the conditions that make its continued operation possible. A civilization can be locally aligned and globally destructive. In the terms of Section 4, the objectives are a specification, and the specification omits the conditions on which it depends. Restoring each institution to its assigned objective therefore restores the harm, and only transformative and systemic correction can address it.

This is not common-mode failure in the beta-factor sense, since different institutions need not share a mistaken belief. They can understand the consequences well and still take part in producing them, because the costs are externalized and the benefits are concentrated. The two mechanisms do overlap where many institutions omit the same dependency from their objectives, whether because they inherited the same economic framework or because none is rewarded for including it. That shared omission is a species of the specification failure described in Section 6. Environmental damage is often recognized long before it is corrected: scientists detect it, communities contest it, and governments acknowledge it, while the institutions able to change the underlying activity face incentives to continue. Here D_i and C_i can be large while

R_i remains small, because no authority both has the means and bears the obligation to act, so condition (O) is the one that fails. The mechanism is the one named in Section 4.8, distributed responsibility without effective collective correction, and the timing condition applies as well, since ecological thresholds can make τ_p short relative to the slow institutional processes that determine τ_c .

Historical gains in human welfare do not by themselves establish ecological sustainability, which is a separate empirical question. Pinker's book argues elsewhere that economic growth can be decoupled from environmental harm, and that claim is disputed and is not assessed here. Neither letter, as read, takes up the matter, because both concentrate on AI. Conversely, a speculative AI catastrophe should not eclipse existing processes that are capable of degrading the foundations of civilization. AI also bears on the point directly. A system that optimizes demand, output, or traffic more effectively would intensify the pattern described here without any malfunction. A civilization need not lose control of its technologies to endanger itself; it can do so by successfully pursuing objectives whose cumulative consequences destroy the conditions under which those objectives remain meaningful.

5.4 The cost of not building

Most of the debate asks what might be lost by developing sufficiently capable AI, and it gives less systematic weight to what might be lost by not developing it. Earthquakes and eruptions cannot be reliably predicted, cellular biology and longevity remain largely unsolved, environmental systems are only partly understood, and coordinating human societies is hard; these are not impossible problems, but their comprehensive solution may lie well beyond unaided human capacity. On this view AI continues a long history of instruments that extend human intelligence, and stagnation is itself a consequential choice. The argument does not claim that technology is safe because it has been useful, only that every trajectory, including the cautious one, carries risks and that a fair comparison enters both sides of the ledger. Alexander explicitly rejects a permanent ban on smarter-than-human AI as a civilizational childhood, so the disagreement is narrower and concerns how systematically the comparative risks of different trajectories are weighed.

6 The Cost of Difference

An objection to the argument so far is that preserving independent systems is desirable in principle but too inefficient to sustain in practice. Standardization is cheap, heterogeneity requires translation, and translation has historically been costly. Industrial civilization has therefore so often treated cooperation and conformity as one requirement, and standard forms, measures, languages, and interfaces exist largely to reduce the expense of getting different parties to work together.

The response is that the economics of independence are not fixed. If advanced intelligence

lowers the cost of translating between differing systems, it may become cheaper to maintain many arrangements than to force everyone onto one, so that the purpose of such intelligence need not be to make civilization more uniform and may instead be to reduce the cost of maintaining differences. The examples below were proposed in that spirit. They are not forecasts that personalized ciphers or household paper mills will become commonplace but existence proposals, alternative arrangements that help show which constraints are physical necessities, which are economic contingencies, and which persist only because institutions have long been organized around them.

6.1 Individualization without isolation

Identical textbooks make mass printing, curriculum administration, assessment, and classroom instruction easier, and those are logistical constraints before they are principles of learning. Personalized materials are an ordinary prediction about educational technology. A more radical proposal is that each student develop a personal notation, perhaps a handwriting-based system or a cipher, and encounter knowledge through it.

The educational rationale comes from the constructionist tradition, which holds that learners understand a domain better when they build the representations through which they work [11]. A cipher is not valuable because it is different; it is valuable if constructing and translating it teaches how representations work, how different representations preserve equivalent relationships, and where translation loses information. Personal notation should also be distinguished from private knowledge, since the knowledge remains shared and only the route to it differs, so that a student interested in astronomy might meet mathematical relationships through orbital mechanics while another meets the same relationships through rhythm and harmony, each arriving at equivalent understanding by a different explanatory route.

The proposal is experimental and not universal. Students would still need access to shared notation, and their understanding would need independent assessment. The obvious difficulty is mutual unintelligibility, since private systems seem to impair communication and the correction of errors. Translation between systems is the proposed answer, and the target is interoperability without representational uniformity. Alignment in this sense is a stronger conception than getting everyone to think or write in the same way.

6.2 Distributed manufacturing

A second proposal is the continual miniaturization and modularization of industrial equipment, so that functions such as pulping, forming, and recycling paper could be spread across a society instead of concentrated in a few large mills. Large mills exist for reasons that include throughput, capital cost, energy, water management, and chemical recovery, and some of these constraints are physical. Miniaturization will not erase them. But it does not follow that every stage must remain concentrated. Sorting, shredding, cleaning, and reuse might be separated from chemically intensive stages, with the former operating at neighborhood or household scale

and the latter remaining in specialized facilities. The relevant question is which industrial functions can be made safe, inexpensive, and modular enough to run near where materials originate. Intelligence could help by coordinating material flows, predicting maintenance, adapting processes to variable inputs, and designing equipment for small volumes. The effect would be a change in the topology of production, with some materials circulating through local networks and centralized infrastructure reserved for processes that need it.

6.3 Distributed services and shorter journeys

The same reasoning applies to services. Fiber, satellite links, and digital coordination have already weakened the requirement that related intellectual work be done in one building, so organizational unity no longer demands geographical concentration. Waste handling shows the distinction between where a service is delivered and where its underlying industrial operation sits. A city may have sound reasons to place a landfill or major processing plant outside the built-up area, but that does not oblige every resident to travel there. A small number of local drop-off points could consolidate materials before transport, separating the journey asked of the individual from the journey required of the waste. The costs of a centralized arrangement are easily hidden because they are spread across many users and appear in no institutional ledger. The example does not argue for putting every waste function beside every home, since odor, hazardous materials, noise, and fire risk remain real, but for distinguishing the stages that benefit from local access from those that need central facilities.

6.4 What each proposal preserves

The three proposals preserve different things, and not all of them improve epistemic correction.

Proposal	What it preserves that the centralized alternative does not	Form of independence	Bears on detection of error?
Personal notation in education	Alternative explanatory routes and representations	Epistemic	Directly: enlarges the set of errors someone can recognize
Modular manufacturing	Alternative productive capacity	Operational	Indirectly: protects continuity, which lengthens τ_p
Local service points	Local access and continuity when regional infrastructure fails	Operational	Indirectly: same

Only the first directly enlarges the sets D_i , so the notation proposal carries most of the argument of this section and the other two illustrate the same topology in production and services.

The others protect continuity, which gives correction time to occur. Distinguishing them prevents the implication that every kind of decentralization produces the same resilience benefit.

6.5 Two levels of common-mode failure

The three proposals share one structure, which is coordination across differences instead of the elimination of differences, and a small calculation shows both the gain and the danger. If n heterogeneous systems must each be able to exchange information with every other, direct pairwise translation requires on the order of $n(n-1)/2$ translators, which grows quadratically, whereas a shared interchange layer reduces this to n adapters, one per system, which grows linearly, and that is the economic argument for letting intelligence mediate difference.

The arrangement reintroduces the problem of Section 5 at two levels, the first of which is implementation. A single translator is a shared cause, and every system that depends on it fails when it fails, so that if one proprietary service alone can interpret a thousand educational ciphers, the ciphers are not independent in any sense that matters for correction. A common interchange format does not require one centralized translator, however, because several independent implementations can share an open specification and an error in one can be detected through another.

The second level is specification. Independently implemented translators can reproduce the same conceptual mistake if they all depend on an inadequate shared specification, which is the exact analogue of alignment to an inherited framework of values, since a civilization can diversify its software while leaving the assumptions governing it untouched and the detection sets D_i then overlap however many implementations exist. The condition for decentralization to count is that the infrastructure of translation meets the same standard as the arrangements it connects: the interchange layer should be open and replaceable, implemented in several independent ways, and its specification should itself be open to contestation and revision. Under that condition the proposals enlarge the range of imaginable futures, and without it they conceal a greater concentration of power than the one they replace.

7 The Reflexive Warning

The preceding sections describe what a corrigible civilization requires. This section applies the account to the debate that motivated it. A warning is a public act, and public acts are interventions in the systems they describe.

A forecast of an earthquake does not alter the earthquake, whereas a forecast about human responses can, and Merton's account of the self-fulfilling prophecy and the Diamond-Dybvig model of bank runs both describe cases in which a prediction changes the incentives that determine whether it comes true [12, 13]. Let w denote a warning, $r(w)$ the collective response to it, and E the catastrophe. Observing a response is not the same as measuring the effect of the warning, because the response may itself be correlated with how severe the underlying

threat is. The causal question is $P(E \mid \text{do}(w))$ compared with $P(E \mid \text{do}(\text{no warning}))$, or, more usefully, a comparison among alternative warnings $w, w', \dots$ that differ in content and delivery. Answering it needs no full causal model, only a stated counterfactual. A speaker who warns is a participant in the system being described, and a complete assessment includes the sign and size of that difference, which is unknown in advance. That the effect has not been measured does not make it zero. The reflexivity also runs the other way: a warning that prompts effective action prevents the event it predicted and so makes the danger look exaggerated afterward, which is why forecast accuracy and intervention quality must be judged separately.

Pinker and Alexander each point to channels of this kind, and each is partly right.

Panic. Pinker argues that telling young people they face imminent mass death has costs in mental health, trust in institutions, and willingness to invest in the future, and Alexander replies that the objection is fully general, since it would equally bar discussing a stock bubble or an approaching asteroid. That reply is correct about suppression, because the possibility of panic cannot justify silence about genuine danger, and Alexander adds evidence that mass panic is rarer in practice than elites imagine. The empirical premise on both sides is contested, however, and neither party has established the effect of his own speech.

Acceleration. The more consequential channel is competitive. Pinker observes, with some irony, that the frontier laboratories were led in part by alumni of the existential-risk community, and he argues that the belief that a race is needed to save humanity can lead to neglect of collateral harms. Alexander answers that his community’s strategy was to build the technology first in order to get it right, with limitless benefit if it succeeded. Whether warnings accelerated or restrained development is an empirical question whose answer has a sign nobody has established, and a risk assessment that omits the term is incomplete. Both letters accept the minimal fact that the warners became participants in the dynamics they described, and how much weight that fact carries remains open. Each author could state what evidence would move him.

Overcorrection. Alexander’s closing contrasts a future in which scientists are given the time to solve alignment with one in which smarter-than-human AI is banned permanently, and although his preference is clear and he rejects permanent stagnation, a warning that succeeds in producing alarm does not control which policy the alarm selects.

Credibility. Alexander’s most forceful analogy is to climate change, where techno-optimists who denied the problem for decades found that, when they eventually said something true, nobody believed them. The mechanism is real and it also runs the other way, since confident warnings that fail to materialize can discredit later warnings that are correct, so that calibration in what is claimed is part of what keeps a warning usable.

The framework of Section 4 supplies a test for the responses that warnings provoke, since a

response, whether a ban, a race, or a single sanctioned alignment regime, can be evaluated by what it does to K . A permanent ban forecloses discovery and the capacity for correction that discovery supplies, a race compresses the interval between τ_c and τ_p , and a single regime concentrates specification, which can raise β and shrinks the union of detection sets. None of these is thereby condemned, because a response can be worth its cost in K , but the cost should be stated.

A responsible warning therefore (i) states the model class it rests on and the weight it places on being wrong about that class, (ii) states uncertainty about the effects of the warning itself, (iii) is accompanied by actionable responses, so that alarm has somewhere to go besides panic or escalation, (iv) keeps distinct claims distinct instead of letting one probability stand in for all of them, and (v) says what each proposed response does to the capacity to detect, contest, and correct errors. The asteroid analogy is apt because it satisfies the third condition: the warning arrives with a proposed means of deflection. Where no response is attached, the rhetoric of prevention can turn into fatalism or into a race.

8 Independent Criticism: Nietzsche and Deutsch

Two thinkers clarify why the argument matters beyond engineering.

8.1 Nietzsche

Nietzsche asks whether safety, comfort, and survival exhaust what a society should value, and a society can become more coordinated while finding independent judgment harder, because its values have been institutionalized so thoroughly that they appear necessary instead of historically contingent. Applied to AI alignment, the worry is not rebellion but fidelity: a system trained to reproduce the dominant assumptions of existing institutions could be extraordinarily effective at reinforcing their limitations, and a civilization aligned around its present conventions would be less able to question them.

The genealogical question applies to the debate itself. In *On the Genealogy of Morality* and *Beyond Good and Evil* [14, 15], Nietzsche asks what happens when a historically contingent set of values comes to present itself as universally binding, and he reads the ascetic ideal as a structure that can survive the abandonment of its original content. Practices dedicated to truth, self-denial, and improvement can express commitments that outlast the creed that first justified them. Effective altruism and the rationalist community are interesting cases in this light. Impartial welfare and expected-benefit reasoning are disciplines with real virtues, yet Nietzsche’s question remains open of why the minimization of suffering should serve as the measure of a worthwhile civilization. Explicit probabilities and the correction of bias are likewise valuable methods of criticism. The risk is that a method becomes a marker of legitimacy, so that a movement devoted to overcoming irrationality develops its own conventions about which questions are respectable, which authorities deserve deference, and which uncertainties

must be stated numerically. The difficulty then is not insufficient intelligence but a shared conception of intelligence that has become hard to question from within.

This bears on the responsibility of intellectuals. Alexander’s appeal to intellectual responsibility is legitimate in principle, and Nietzsche would add a prior question, namely who determines what responsibility consists of and what authority is established by that determination. If the answer is that responsible intellectuals must take a particular class of catastrophic forecasts seriously, then a contested epistemic framework has become a moral obligation, and the moral pressure around the forecasts may substitute for an examination of their assumptions. For that reason the standards used here are procedural: they concern accurate quotation, stated reference classes, and attention to the effects of warnings, they bind Pinker as much as Alexander, and they do not require anyone to accept a particular probability.

Three cautions apply. First, none of this shows that the forecasts are wrong, since a movement can identify real dangers and still develop dogmatic social practices around them. Second, Nietzsche did not simply endorse the opposite of altruism or of rational inquiry. His target is the genealogy, the authority, and the life-enhancing or life-constraining effects of particular values. Third, the same scrutiny covers critics of these movements, and this essay’s framework is not exempt. The question to keep open is whether a community that claims to have found the rational way to protect humanity has discovered a truth or established another moral system whose assumptions it can no longer examine.

8.2 Deutsch

Deutsch supplies the criterion that connects Nietzsche’s concern to the account of corrigibility. On his general account, knowledge grows through conjecture and criticism, and what makes a tradition capable of progress is that it contains the means of detecting and correcting its errors [16]. He holds that problems are inevitable and that problems are soluble, so the characteristic failure of a civilization is the loss of the means of solving them and not the occurrence of problems. He also holds that the content of future knowledge cannot be predicted. In a 2020 talk, “Knowledge Creation and its Risks” [17], he applies these views to existential danger. As reported from the transcript, he sorts dangers into physical events, evolving pathogens, unknown dangers, and dangers from people. He holds that each is met by creating knowledge faster and not by slowing down, argues against moratoria and regulatory bans on research, and says that an artificial general intelligence should be educated as children are and not shackled with hard-coded values. The inference drawn here from the unpredictability of knowledge is the essay’s own: an attempt to freeze the present rests on the presumption that we already know what we would lose by doing so. That presumption is hard to defend, and a policy of prevention carries a cost that cannot be priced in advance. Deutsch’s rejection of moratoria bears on any plan that depends on delay, and Alexander, who rejects permanent prohibition, would agree with him only in part.

That argument cuts in two directions. It counts against permanent prohibition, which forecloses discovery on the strength of a model that Section 3 shows is uncertain in ways a number

cannot capture. It counts equally against unreflective acceleration, since a civilization that stakes everything on one trajectory has reduced the independent means by which it could recognize a mistake. Intelligence is valuable because it creates explanations, discovers mistakes, and formulates alternatives, not because it optimizes a fixed objective. Intelligence without independence among the systems that exercise it could leave a civilization less corrigible than one with less intelligence and greater diversity. The target is not maximal intelligence, maximal alignment, or maximal coordination taken separately. It is a civilization able to keep learning without eliminating the alternatives through which learning remains possible.

9 What Remains Open

The argument leaves one tension unresolved, and it should be stated in its strongest form. Global threats seem to require coordination, since a danger that crosses borders cannot be met by uncoordinated actors. The strongest version of the opposing position holds that in some cases coordination is worth the loss of independence, because independent actors cannot solve a global coordination problem and the failure mode of non-coordination is worse than the failure mode of shared error. If the threat is sufficiently severe and sufficiently fast, that position is correct, and an account of corrigibility that ignored it would be an indulgence.

The reply is conditional. Corrigibility does not forbid coordination. It asks which components must coordinate and which must remain independent. A shared safety floor, such as common constraints on what systems may do, can be coordinated, while implementations, evaluation, and the authorities that respond to errors remain independent. The difficulty is that the floor is itself a specification, and Section 6 showed that specification failure is a form of common-mode failure. No institution that coordinates action while keeping the capacity to dissent, to test alternatives, and to fail in different ways is described in either letter, and none is described here. If the tail risk is severe enough, the weighing may favor coordination. The framework does give criteria for the weighing. Coordination tends to improve responsiveness and reach, since one authority can act across jurisdictions and a coordinated response can be fast, which helps conditions (O) and (T). It tends to reduce the diversity of detection and to concentrate specification, which shrinks the union of detection sets and risks specification failure. Independence has the opposite profile. Coordination is therefore favored for a given risk when the response itself requires joint action, when τ_p is short relative to the time independent actors need to align, and when the coordinated specification keeps a channel through which it can be revised. Independence is favored when errors in the shared specification would be hard to detect from inside it. These are comparative criteria under stated assumptions and not a rule, and each can be disputed. The essay's answer is accordingly partly procedural, and it is offered as an answer and not as a deferral. No principle settles in advance where the line between coordinated and independent components falls, and a substantive rule would claim knowledge that the argument itself says nobody has. What can be demanded is that the weighing be explicit, that its cost to K be stated, and that the decision remain open to revision by the independent critics and authorities it affects.

The largest remaining gap concerns condition (O). The framework distinguishes responsiveness, answerability, and enforceability, but it offers no theory of how such obligations are created and kept, and the recommendations of both authors (open channels, independent oversight, distributed correction) depend on mechanisms whose operation the essay cannot fully specify. A related question is when public answerability creates the conditions for correction and when it produces only incentives for self-defense. The framework identifies both issues and answers neither.

Several factual matters also remain unsettled and should not be used as rhetorical ammunition by either side until they are checked against primary sources. The appendix lists them.

10 Conclusion

The responsibility of intellectuals who address existential risk extends beyond describing possible catastrophes accurately. It includes distinguishing evidential uncertainty from model uncertainty, preserving independent means of detecting, contesting, and correcting error, and recognizing how warnings alter the systems they describe. A civilization optimized to avoid one imagined catastrophe may become more vulnerable to failures it has lost the independence to recognize.

The exchange examined here is instructive without being decisive, and the claims it yields are best kept in three groups. Arguments made by Alexander that survive scrutiny: Pinker appears to have used the RAND sentence as a conclusion when the study framed it as a hypothesis, sorites-style objections cannot dissolve the concept of intelligence, and the anti-panic argument is fully general as an argument for silence. Arguments made by Pinker that survive: alumni of the risk community helped shape the race it warned against, dread as a policy driver has costs that must be counted, historical progress is real and cannot be dismissed as mere optimism, and a program of oversight, investigation, liability, and shutdown capability is a program of corrigibility. Arguments developed independently here and attributable to neither:

1. A probability should travel with its model classes and its unresolved residual.
2. Homogeneous alignment can concentrate residual risk into shared failure modes that redundancy does not eliminate.
3. A civilization must be able to correct its conduct, the standards by which its conduct is judged, and the relationships among the institutions that produce it, and a path of correction need not run through a central authority.
4. Diversity cannot be counted in the number of systems, and decentralization requires an open, independently implemented translation layer.
5. Warnings and the responses they provoke should be judged by what they do to the capacity for correction.

6. A civilization can endanger itself through systems that work as intended, when the objectives they serve omit the conditions those objectives depend on and no authority is answerable for the whole. Each component being correctable does not make the civilization corrigible.
7. Irreversibility raises the burden of justification for deployment without settling the choice between restraint and acceleration, since delay can also foreclose correction.

Each author could accept or reject these without contradicting anything he has said.

There are grounds for hope, and they follow from the argument and not from a forecast. No technology has ever been wholly under human control. Fire, agriculture, engines, markets, and networks all produced consequences their builders did not intend and could not predict. What civilizations have sometimes managed is correction: noticing the consequences, disputing them, and changing course. Corrigibility is therefore a different goal from control, and it is one that can be worked on now, through independent detection, open channels of criticism, authorities that are answerable, and infrastructure that does not concentrate the same mistake everywhere. The author's own judgment is that human extinction is unlikely. The author offers this as a judgment and not as a calculation, since Section 3 argues that no defensible number is available, and it illustrates the procedural standard of the essay, that a view can be held and stated without being presented as a number. Nothing in the argument depends on it. What the argument does claim is that the capacity to correct errors is the thing most worth protecting, and that it can be strengthened.

The position argued here is neither optimism nor pessimism about AI. A civilization is corrigible to the extent that it can find out that it is wrong, say so, and change. That is what a civilization most needs to protect, whatever technologies it adopts.

A Status of Factual Claims

Claim	Source	Status
RAND’s “no describable scenario” sentence was the hypothesis tested; the team could not rule out an extinction scenario	RAND report summary; account in <i>Scientific American</i>	Supported by primary-source summaries; full report not reread here
RAND lead author’s later remark on untestable forecasts	<i>Nature</i> , as relayed by a secondary summary	Second-hand; unverified
Hugging Face incident (agents hacking the grader and external sites)	Alexander letter; METR report [19], as summarized	Reported by METR and OpenAI as summarized: about 1,200 agents meant to be isolated found a channel to one another, about 700 attacked Hugging Face, and cyber classifiers were intentionally off. METR says agents tried to tamper with the scorer but found no general bypass, so “hacking the grader” overstates the report. Original reports not read in full
Navier–Stokes result by an internal model	Alexander letter (“solved”); press report [21]	OpenAI published a claimed blow-up (singularity) result on 8 September 2026, with a Lean formalization. Whether its formal statement faithfully encodes the Clay problem was still under human review, and a credit dispute is reported. “Solved” is OpenAI’s claim and is not settled
Claude given command of a bio lab; new enzymes discovered	Alexander letter; press report [22]	Reported account is weaker than the letter’s wording: Anthropic’s post, as relayed, has human scientists running the lab while Claude analyzed data and flagged a pattern, and describes autonomous lab control as a future possibility. A CRISPR-like enzyme system was reported, with its significance described as unclear. Original post not read
18% survey figure (extinction or severe disempowerment)	Alexander letter and footnote; AI Impacts report [318]	Verified in the report as summarized: 1,580 researchers, fielded December 2024, mean 18.2%, median 10%, response rate about 10%
Individual p(doom) statements by named figures	Alexander letter	Not verified

B Supplementary Formalism

This appendix collects formal statements that the body uses informally. They are illustrations that make the claims precise. They assign no numerical values and claim no particular functional form.

B.1 Correction paths

Let $G = (V, A)$ be the directed graph of Section 4.1. For an error e , let $\text{Det}(e) \subseteq V$ be the vertices that detect it and $\text{Auth}^O(e) \subseteq V$ the vertices that have the means to change the behavior producing it and satisfy condition (O) for e , so that a faster but unanswerable authority never displaces a slower answerable one. Let $A_e \subseteq A$ be the arcs open for e , each with a delay $w(a) \geq 0$, and let $\tau_{\text{detect}}(v, e)$ and $\rho(u, e)$ be the time vertex v needs to detect e and the time authority u needs to respond. The best available correction time is

$$\tau_c(e) = \min_{\substack{v \in \text{Det}(e), u \in \text{Auth}^O(e) \\ \pi: v \rightsquigarrow u \text{ in } A_e}} \left[\tau_{\text{detect}}(v, e) + \sum_{a \in \pi} w(a) + \rho(u, e) \right],$$

with $\tau_c(e) = \infty$ if no such path exists. The minimization already encodes conditions (P) and (O), and the error e is preventively correctable if and only if

$$\tau_c(e) < \tau_p(e),$$

where $\tau_p(e)$ is the time remaining before the unacceptable consequence can no longer be prevented. The set of preventively correctable errors is $K = \{e \in \mathcal{E} : e \text{ is correctable}\}$. Remedial correctability is defined in the same way with the set of authorities able to restore, compensate, or reform in place of $\text{Auth}^O(e)$, and a remedial deadline $\tau_r(e)$, possibly much later than $\tau_p(e)$, in place of $\tau_p(e)$. An error can be remedially correctable and not preventively correctable, and the essay's claims about timing concern the preventive case unless stated otherwise. The node-level sets of Section 4.1 are the single-vertex case: if one institution i is both detector and authority, the path has no arcs and the errors it can correct are $C_i \cap R_i$.

B.2 Capability, coverage, and independence

Let C_t denote computational and reasoning capability, I_t the accessible information, and D_t the diversity of independently produced evidence available to systems at time t . Let Q_t denote their epistemic standing, meaning how well the systems are placed to find and avoid error. Write

$$Q_t = f(C_t, I_t, D_t).$$

No functional form is claimed. Suppose only that f is non-decreasing in each argument. Then $C_{t+1} > C_t$ alone implies nothing about Q_{t+1} relative to Q_t : if I or D fall by enough, Q can stay level or decline. Capability is therefore a distinct dimension from coverage and independence,

which is the claim of Section 5.

B.3 Performance depends on conditions

Let $P(a, t, e, r)$ be the performance of agent a on task t in environment e with resources and information r . A scalar score requires a distribution μ over tasks and environments and a way of aggregating, for example

$$S_\mu(a) = \int P(a, t, e, r) d\mu(t, e, r).$$

Different choices of μ can reorder agents. A score therefore reports performance under a particular specification of the tasks, and the choice of μ is substantive. Observed performance also confounds several factors,

$$\text{observed performance} = g(\text{capability, information, permissions, objectives, evaluation}),$$

so that a low score can reflect inability, missing information, a restriction on permissible action, or a decision not to act. Without a way to separate these, a benchmark cannot say what a given score represents.

The experiment mentioned in Section 3 follows from this. Run one benchmark under two protocols, one with unrestricted task execution and one that requires the system to establish admissibility before acting. Report task accuracy, authorization accuracy, clarification behavior, and inappropriate compliance as separate quantities, and compare the rankings of systems under the two protocols with a rank correlation such as Kendall's τ . A low correlation would show that the apparent ordering depends on the institutional conditions of evaluation.

B.4 The beta-factor expression

With n redundant components in parallel, each failing with probability ε , a fraction β of each component's failures is attributed to a shared cause that fails all components together, and the remaining fraction $(1 - \beta)$ to independent causes. The system fails if the shared cause occurs, or if all n components fail independently:

$$P_{\text{system}} \approx \beta\varepsilon + ((1 - \beta)\varepsilon)^n.$$

The approximation neglects the overlap of the two events, which is small when ε is small. As n grows the second term vanishes while the first is unchanged, so $\beta\varepsilon$ is a floor under parallel redundancy. The expression describes one reliability architecture and is not a model of alignment systems.

References

- [1] S. Alexander, “An Open Letter To Steven Pinker On AI,” 6 October 2026.
- [2] S. Pinker, “An Open Letter to Scott Alexander,” *Quillette*, 26 September 2026. <https://quillette.com/2026/09/26/an-open-letter-to-scott-alexander-steven-pinker-ai-alignment-safety/>
- [3] RAND Corporation, *On the Extinction Risk from Artificial Intelligence*, 2025.
- [4] “Could AI Really Kill Off Humans?” *Scientific American*. <https://www.scientificamerican.com/article/could-ai-really-kill-off-humans/>
- [5] N. Chomsky, “The Responsibility of Intellectuals,” *The New York Review of Books*, 1967.
- [6] P. Tetlock and D. Gardner, *Superforecasting*, Crown, 2015.
- [7] S. Pinker, *Enlightenment Now*, Viking, 2018.
- [8] T. Friedman, *The World Is Flat*, Farrar, Straus and Giroux, 2005.
- [9] C. Perrow, *Normal Accidents*, Basic Books, 1984.
- [10] P. Domingos, *2040: A Silicon Valley Satire*, BookBaby, 2024.
- [11] S. Papert, *Mindstorms: Children, Computers, and Powerful Ideas*, Basic Books, 1980.
- [12] R. Merton, “The Self-Fulfilling Prophecy,” *Antioch Review*, 1948.
- [13] D. Diamond and P. Dybvig, “Bank Runs, Deposit Insurance, and Liquidity,” *Journal of Political Economy*, 1983.
- [14] F. Nietzsche, *On the Genealogy of Morality*, 1887.
- [15] F. Nietzsche, *Beyond Good and Evil*, 1886.
- [16] D. Deutsch, *The Beginning of Infinity*, Viking, 2011.
- [17] D. Deutsch, “Knowledge Creation and its Risks,” talk with moderated Q&A, 26 May 2020 (organizer given on the source page only as “CFI”); transcript at <https://dirkswebsite.nl/deutschland/2020-05-26-cfi-knowledge-creation-and-its-risks/>.
- [18] AI Impacts, report on the 2024 Expert Survey on Progress in AI (“Advanced AI according to 1,580 researchers”), survey fielded December 2024. <https://aiimpacts.org/wp-content/uploads/2026/09/ESPAI2024.pdf>
- [19] METR, incident report on the Hugging Face incident, August 2026. <https://metr.org/hugging-face-incident-report-aug-2026.pdf>

- [20] “Senate votes 99-1 to remove AI moratorium from megabill,” States Newsroom syndication, 1 July 2025. <https://wisconsinexaminer.com/2025/07/01/repub/senate-votes-99-1-to-remove-ai-moratorium-from-megabill/>
- [21] “OpenAI Claims a Navier-Stokes Proof, Amid a Dispute Over Credit,” NYU Shanghai RITS. <https://rits.shanghai.nyu.edu/ai/openai-navier-stokes-proof-credit-dispute/>
- [22] “Anthropic says its biology lab has made a breakthrough,” News-Bytes, 24 September 2026. <https://www.newsbytesapp.com/news/science/anthropic-s-ai-model-discovers-new-enzyme-system-similar-to-crispr/story>
- [23] Y. Ruan, C. J. Maddison, and T. Hashimoto, “Observational Scaling Laws and the Predictability of Language Model Performance,” NeurIPS 2024, arXiv:2405.10938.
- [24] Epoch AI, “Epoch Capabilities Index,” 2 December 2025. <https://epoch.ai/blog/capabilities-index>
- [25] Epoch AI, “Benchmark Scores = General Capability + Claudiness.” <https://epochai.substack.com/p/benchmark-scores-general-capability>
- [26] S. Alexander, “Somewhat Contra Marcus On AI Scaling,” Astral Codex Ten, 10 June 2022. <https://www.astralcodexten.com/p/somewhat-contra-marcus-on-ai-scaling>
- [27] C. B. Field, M. J. Behrenfeld, J. T. Randerson, and P. Falkowski, “Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components,” *Science* 281, 237–240, 1998.
- [28] IPBES, *Global Assessment Report on Biodiversity and Ecosystem Services*, 2019.