

Adverse Conditions

Why We Must Think Together
Beyond the Scale of Our Darkness

Forrest Landry

UVSM PUBLICATIONS
Wytheville, Virginia

© 2026 Forrest Landry. Some rights reserved.

Published by UVSM Publications LLC
965 Reed Creek Mill Rd
Wytheville, VA 24382-8287
USA

uvsmpublications.com
contact@uvsmpublications.com

ISBN 979-8-90710-220-0 (paperback)
ISBN 979-8-90710-221-7 (PDF)

DOI: https://doi.org/10.67291/YUGQ2685

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

You may share this work for non-commercial purposes with proper attribution, but you may not adapt, transform, or build upon it.

Full license text: https://creativecommons.org/licenses/by-nc-nd/4.0/

Author ORCID: 0009-0005-6275-0362

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), except as permitted under the license above, without prior written permission of the publisher.

Contents

Introduction **1**

I. The Question These Essays Share *1*

II. What the First Essay Establishes *3*

III. What Stands Between the Requirement and Its Statement *5*

IV. What Becomes of the Capacity Once It Exists *6*

V. Where the Volume Stops *8*

VI. The One Thing They Agree On *9*

Essay 1 · The Structural Dynamics of New Technology **11**

I. The Shape of the Question *12*

II. The Asymmetry of Advocacy *13*

III. The Compounding of Biases *15*

IV. The Reversal of the Defensive Baseline *17*

V. Where the Costs Come to Rest *18*

The Burden of Proof *20*

VI. What Ought To Be Done *23*

VII. The Three Disabling Factors *24*

VIII. The Requirement *30*

What Technology Is, and Whom It Serves *34*

IX. Summary *37*

X. Conclusion *40*

XI. What This Essay Adds to the Volume *41*

Essay 2 · The Simulation of Precaution **45**

I. What Precaution Is For *46*

II. The Artifacts and the Substance *47*

III. The Filter *49*

IV. Sincere Simulation *50*

V. What an Observer Could Look At *52*

VI. Why Each Observable Is Capturable 54
VII. The Cost of Verification 56
VIII. What Capture Would Have to Cost 57
IX. Summary 59
X. Conclusion 60
XI. What This Essay Adds to the Volume 61

Companion Essay · Three Problems with AI Self-Regulation **64**
I. The Belief in Internal Self-Regulation 65
II. The Principal-Agent Problem 66
III. The Race to the Bottom 67
IV. Rules for Rulers 68
V. One Structure, Seen in Three Places 69
VI. Whether a Company Can Bind Itself 70
VII. Summary 71
VIII. What This Essay Adds to the Volume 72

Companion Essay · Anti-Democratic Risk **74**
I. The Paper and the Agenda 75
II. How the Agenda Increases Real Risk 77
III. The Three Beliefs, Read as a Sales Pitch 79
IV. The Three Pillars 81
V. Real Value Is in the Present 84
VI. Existential Risk, Redefined as a Failure to Buy 87
VII. Two Wrong Claims: Inevitability and Alignment 89
VIII. Emergency Powers and the Use of Weakness 91
IX. What Democracy Means Here 93
X. Conflict and Mistake in the Conversation 95
XI. Makers, Owners, and Users 103
XII. What Is at Stake 104
XIII. Summary 105
XIV. What This Essay Adds to the Volume 106

Essay 3 · Proof in Hostile Territory 109
I. The Question That Prompts This 110
II. The Case of One and One 110
III. What an Adverse Proof Environment Is 111
IV. The Historical Case of the Loopholes 112
V. Hostility Is Proportional to Interest 114
Why an Impossibility Proof With Billions at Stake Will Face a Doubt Campaign 115
VI. The Double Bind of Definition 119
VII. The Challenge Raised Before the Definition 120
VIII. The Response: Narrowing the Assertion 121
IX. Soundness, Validity, and the Preservation of Relevance 124
From Idiosyncratic View to Relevant Proof 125
X. Summary 130
XI. Conclusion 132
XII. What This Essay Adds to the Volume 133

Essay 4 · The Honeypot Problem 136
I. Power Arises Wherever People Gather 137
II. Success as Target 137
III. The Ranking System 139
IV. The Centralized Intelligence Agency as Attractor 140
V. The Convening That Could Not Be Held 141
VI. Decentralization as a Structural Requirement 143
VII. Making Secrets Not Matter 145
VIII. Data, Intelligence, and Meaning 147
IX. Summary 148
X. Conclusion 149
XI. What This Essay Adds to the Volume 150

Essay 5 · Questions as Instruments of Discernment 153
I. The Evaluator's Problem 154
II. Zero-Knowledge Verification 155

III. The Diagnostic Is Novelty, Not Answer *156*
IV. Upgrading the Questions Upgrades the Organization *158*
V. The Pattern, and the Damage Done by Examples *159*
VI. The Inversion: Solicitation as Education *161*
VII. Leading with Risk *162*
VIII. The Arms Race and the Problem of Distribution *163*
IX. Triangulating the Knowledge State *164*
X. Two Questions Others Have Raised *165*
XI. Summary *166*
XII. Conclusion *167*
XIII. What This Essay Adds to the Volume *168*

Afterword · On What Would Have To Be Built **171**
I. The Standing of This Statement *171*
II. The First Property: Sense-Making, Choice-Making, and Agency *172*
III. The Second Property: Fractal Scalability *173*
IV. The Third Property: Robust Anti-Corruptibility *174*
V. The Fourth Property: Resolution of the Principal-Agent Problem *177*
VI. Why Three Are Not Enough *180*
VII. One Line of Work *182*
VIII. What This Afterword Adds to the Volume *185*

Bibliography **187**

Terms **196**

Index **214**

Introduction

On what the five essays share,
and why they are collected

I. The Question These Essays Share

Groups of people can think together. This is an ordinary fact, available to anyone who has sat in a room where a difficult matter was worked through by people who arrived without the answer and left with one that none of them had brought. What happens in such a room is not the aggregation of opinions. It is the formation of an understanding that did not previously exist, held by the configuration of people rather than by any of them individually, and better than what any of them could have reached alone.

The capacity is real, it is well attested, and it is the thing on which every collective decision of consequence depends.

It also has conditions. It works at some scales and not others, under some arrangements and not others, and among parties standing in some relations to one another and not in others. Where the conditions hold, the capacity operates. Where they fail, what remains is a set of procedures that resemble collective judgment and do not perform it: the transmission of preference, the aggregation of votes, the negotiation of positions arrived at beforehand. These procedures scale, which is why they are the ones in general use, and they are not the same thing.

The five essays in this volume take up different subjects, each from its own starting point. The question they share is the one just stated. Under what conditions does the capacity for collective judgment fail, and what would be required for it to succeed at the scale on which the decisions that matter are now being made?

Each essay answers a part of it by examining one structural condition under which collective judgment fails even where every participant is competent and sincere. That qualification is the load-bearing one. None of the failures described here requires anybody to be foolish or dishonest. Each is reproduced by the ordinary and expected conduct of ordinarily well-intentioned people, which is why exhortation does not address any of them and why the analysis is structural throughout. The answer the five essays reach is that the missing thing is the capacity to think together at scale.

The essays are ordered so that the requirement is stated first and the obstacles to stating it are taken up immediately after, before the two essays that examine what becomes of the capacity once it exists. Each was written to stand alone and each can be read alone. A reader who takes them in the order given will find that the later essays answer questions the earlier ones raise and pose questions the earlier ones did not see, and the cross-references throughout are placed to make that traffic explicit rather than to oblige the reader to follow it. Where a passage names another of my books, that book develops the point at length; everything the argument of this volume needs is stated here, apart from the control argument, which Section V names as carried elsewhere. Questions raised by other people are kept as questions, and credited as theirs.

Companion essays stand beside the five numbered essays. Each companion essay is placed after the numbered essay whose subject it carries into a further case, each is unnumbered, and each can be read alone. Four further companion essays on how a proof is received, "The Expert Judgmental", "Bidirectional Crank Filters", "Three Questions of Reason" and "Math Expectations", are Chapters 8 to 11 of *A Paradox of Presentation; Why the Field of AI Safety Does Not Hear the Impossibility Argument, and How It Can Be Presented*, where they serve that book's subject.

Private persons are not named in this book.

II. What the First Essay Establishes

The next three sections give a map of the essays; a reader who prefers to meet the arguments first can go to Section VI, which states the claim they converge on, and return here afterward.

Essay 1, *The Structural Dynamics of New Technology*, identifies the requirement.

Its argument is that the introduction of a new technology into a society is governed by forces that do not track the merits of that technology. Advocacy for a technology has a reliable source of funding in the technology's future profits; resistance protects goods held in common from which no party can extract a proportionate private return, and is funded accordingly. Optimism, action, and commercialization biases compound. The ordinary advantage of the defender reverses wherever a technology confers asymmetric advantage. And cost and risk come to rest on the commons while benefit accrues locally, which is self-reinforcing, because each successful externalization funds the next round of advocacy.

The remedies that follow from this diagnosis are neither obscure nor contested, and they have been recommended, in substantially the same form, for decades. Essay 1 sets them out and then asks why they are not implemented. Its answer is that they are addressed to a decision-making capacity that does not exist: they presuppose an actor who could weigh benefit against systemic risk on behalf of all affected parties and then act on the weighing, and no such actor is present at any scale at which the relevant technologies operate.

Three factors account for the absence, and each is independently sufficient to produce it. Genuine precaution is expensive and simulating it is cheap, and the two are indistinguishable by their visible artifacts. Methods by which a group can think together exist and work at approximately the scale at which each participant can hold a working model of every other, and above that scale no such method presently exists; what replaces

it is representation and voting, which transmit preference and cannot form understanding, and which can select among the options placed before them but cannot notice that the right option is absent from the ballot. And within any organization that develops a technology, authority increases as understanding decreases, so that a risk identified at the face of the work must be escalated through a chain of people each of whom understands it less well than the last and each of whom has a stronger reason to discount it.

The first of these is the subject of Essay 2. The second is the subject of the volume as a whole. The third recurs, in a more general form, as the opening problem of Essay 5. That problem, the evaluator's problem, is not a further disabling factor: it is the third factor seen between organizations rather than within one, the same gradient along which authority increases as understanding decreases.

From this the essay derives a specification. Any system adequate to problems of this class must be genuinely effective at collective sense-making (the forming of a shared understanding), choice-making, and the enablement of agency; must be fractal and scalable, operating at world scale by the principles on which it operates among a few; must be robustly anti-corruptible, such that the attempt to capture it is ineffective and not merely forbidden; and must genuinely resolve the principal-agent problem, aligning the interests of those who decide with the whole of what their decisions affect.

The specification does not admit of partial satisfaction. A system that achieves three of the four properties is taken through the one it lacks. That claim is the hinge on which the rest of the volume turns, and the four essays that follow it can be read as an examination of the several ways in which one or another of the four properties is lost.

III. What Stands Between the Requirement and Its Statement

Two of the essays concern obstacles that arise before any such system could be built: obstacles to establishing, in public, that it is needed.

Essay 2, *The Simulation of Precaution*, develops the first of the three disabling factors that Essay 1 identifies. The substance of precautionary governance is that it changes decisions, and the value of the decisions it changes is what it costs. Its visible artifacts (the committee, the charter, the register, the review) are producible without that substance, for a small fraction of the price, and are then the same artifacts. The essay traces what follows: that the cheaper form is selected by an environment that can observe only artifacts (the precaution filter); that most of the resulting simulation is sincere (sincere simulation), since no role within an organization is assigned to hold the comparison between what happened and what would have happened; that there are observables which would in fact discriminate; and that publishing any of them supplies the specification by which it can be produced without the substance (specification by publication). The essay closes on what survives that result, which is a condition on arrangements rather than a better choice of measurement. Its argument can be put in one line: care shows in what it gives up, and what is given up leaves no object behind.

Essay 3, *Proof in Hostile Territory*, concerns the cost of saying any of this to the parties who most need to hear it. Its central observation is that the difficulty of establishing a claim is not a property of the claim. It is a relation between the claim and its audience, and where the audience has a material interest in the claim being false, the number of steps that must be laid down is set by the audience's willingness to generate challenges rather than by the logical structure of the argument. The essay names the resulting condition an adverse proof environment, identifies the double bind in which such an environment places every attempt at definition,

and describes the response it recommends: narrowing the assertion until its surface is small enough to defend. What is narrowed is the surface of each step, not the conclusion: the full claim is reached by composing narrow results, each of which can be defended alone.

Essays 2 and 3 stand in a particular relation to the volume. They are the reason a book of this kind is difficult to write and difficult for its readers to believe, and they are placed early so that a reader may hold them in view while reading the rest.

Two companion essays stand between Essay 2 and Essay 3, and carry the result of Essay 2 into two further cases. Three Problems with AI Self-Regulation takes up the belief that the companies building AI can supply the needed judgment themselves, and shows that the principal-agent problem, the race to the bottom, and the way leaders keep power each defeat that belief, and that a company can be bound only by a party outside it. Anti-Democratic Risk takes up the study of existential risk, finds it shaped by an agenda that increases inequality and so increases the real risk, and concludes that which risks to take must be chosen by the parties those risks affect.

IV. What Becomes of the Capacity Once It Exists

Two further essays take up the case in which the capacity Essay 1 specifies has been built, or is being built, and ask what then happens to it.

Essay 4, *The Honeypot Problem*, observes that a collective sense-making capacity becomes an object of capture in exact proportion to how well it works. A group whose judgment is unreliable is of no use to a party wishing to direct judgment. A group whose judgment is genuinely good is valuable to any party that would like correct conclusions in advance, or the appearance of them attached to a preferred position. The essay distinguishes this from ordinary corruption (it is a consequence of success, not a failure), follows the drift by which any successful convening acquires the structure of a centralized intelligence agency without anyone intend-

ing it (the centralized intelligence attractor, Essay 4, Section IV), and argues that decentralization is therefore a structural requirement rather than a preference. Its design principle is not the better keeping of secrets but the construction of processes in which secrets do not matter.

Essay 5, *Questions as Instruments of Discernment*, takes the same problem at the scale of a single judgment. A party who must decide which of several claimants is doing real work cannot evaluate the answers, since evaluating them requires the expertise that made the decision hard. The essay develops an alternative: that a question lying necessarily in the path of the claimed work discriminates by whether it is new to the party asked, since answers can be prepared but the shape a hard question leaves in an organization that has lived with it cannot be. It then treats the vulnerability of that instrument, which is that a published question is thereafter something to have an answer ready for, and that its author does not learn when this has happened (the undetectable defeat, Essay 5, Section VIII).

Read together, the two essays make the same point from opposite ends of the scale. A capacity for discernment is valuable, and what is valuable is acquired, and the acquisition is not visible to the party whose discernment has been acquired.

Both essays are therefore addressed to the third property of the specification, the requirement of robust anti-corruptibility, and both arrive at the same reason for its difficulty. An arrangement resists capture only to the degree that capturing it is expensive, and expense is the only quantity available to a designer, since no arrangement is beyond acquisition by a party willing to pay enough. Essay 2 reaches the same conclusion from the direction of measurement, and shows that the conclusion depends on the fourth property (that those who decide bear what their decisions cause), since an observation means nothing when the party observed supplies it. Three essays, each argued from its own starting point, therefore come to recommend the same thing: that the aim is not an arrangement that cannot be bought, but one whose purchase costs what the thing being purchased was supposed to be worth (capture-cost parity, Essay 2, Section

VIII). The question to put to any arrangement is therefore never "can this be captured?", since everything can, but "what does capture cost, against what it buys?".

V. Where the Volume Stops

The limits of the volume are as definite as its claims, and stating them is not a disclaimer but part of the argument.

The volume gives a specification, in Essay 1, and, in the Afterword, a statement of what each of its four properties demands of a design and why three of the four are insufficient. A specification is a statement of requirement, and it is not a design. The volume does not design the governance systems it argues are required, and no one has yet designed them.

Essay 3 describes the control-theoretic strategy as a strategy; the argument itself is carried in other books. It does not develop the control-theoretic argument that Essay 3 describes in its Section VIII. That essay sets out a strategy (enumerate the necessary and sufficient elements of a control relation, bound each separately, and compose the bounds according to the control algorithm) and describes it as a strategy. The argument itself is carried out in other books and is not presented here. *Ape Trap; The Runaway Artificialization of Our Planet* examines each of the six steps of control in turn. *The Ceiling Below the Floor; The Containment Fallacy and AGI Non-Safety* states the result as one bound, on the rate of information that any controller can use. The description is deliberately not developed into the thing described.

A reader may ask whether accepting any of this requires accepting *An Immanent Metaphysics*. The essays argue from incentives, from observation, and from logic. The two places that use the terms of *An Immanent Metaphysics*, Essay 3, Section IX, and Essay 4, Section VIII, name it where they use it, and Essay 3, Section IX, says that the composite proof which its Section VIII describes can be checked step by step without those terms.

The volume's claims about the outcome of unresolved traps describe a default trajectory. It offers no timeline. Nothing in these pages estimates how much time there is, and Essay 4 records the reasons why a serious attempt to produce such an estimate was judged to be more dangerous than useful. Claims about the expected outcome of an unresolved trap (a multi-polar trap: a situation in which each party's sensible action produces an outcome every party would reject) are claims about a default trajectory and not claims about fate.

VI. The One Thing They Agree On

Five essays, about incentive geometry (how the gains and losses of each party are arranged, and so which way each is pushed), about measurement, about the price of an argument, about capture, and about how to ask a question, converge on a single claim.

The missing thing is the capacity to think together at scale.

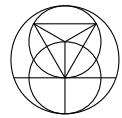
Not better regulation, though that would be an improvement. Not better norms among researchers, nor better disclosure by developers, nor better instruments of oversight, each of which would also be an improvement and none of which is prior. What is missing is the capacity to form a shared understanding among the parties a decision affects, to choose on the basis of that understanding, and to act on the choice in a manner that cannot be acquired by a party who would prefer a different action. Every other item on the list of remedies presupposes that capacity and is inert without it.

This is why the essays belong together though each was argued on its own. Essay 1 states the requirement and shows that nothing satisfies it. Essay 2 shows why the appearance of satisfying it is cheaper than satisfying it, and is therefore what an observer will find. Essay 3 shows what it costs to establish any of this before an audience that would prefer it were false. Essays 4 and 5 show that the capacity, once built, becomes the most valuable thing in the environment and is therefore the thing most worth

acquiring, at the scale of an institution and at the scale of a conversation. The two companion essays find the same structure inside the industry and inside the study of risk, and the Afterword turns the four properties into what a design must supply.

A reader who wants to begin at once can learn the necessary questions of their own field, ask them of themselves first, and give them to those who allocate resources; Essay 5 says how, and the Afterword, Section VII, says why this can begin before anything else is built.

The volume is called *Adverse Conditions* for the plain reason that these are conditions, they are adverse, and they are the ones actually obtaining. Naming them exactly is not the same as remedying them. It is, however, what a remedy would have to be addressed to, and the essays are offered to those who intend to address it.



Essay 1

The Structural Dynamics of New Technology

On why precaution is recommended, and why it does not occur

ABSTRACT

The introduction of a new technology into a society is governed by structural forces that operate independently of the merits of that technology. Four such forces are identified here: an asymmetry in the funding available to advocacy as against resistance; a compounding of optimism, action, and commercialization biases; a reversal of the ordinary defensive baseline wherever a technology confers asymmetric advantage; and a systematic externalization of cost and risk onto the commons while benefit accrues locally. Taken together these forces guarantee that the harms of a new technology receive structurally insufficient attention. The remedies that follow from this diagnosis are well known and are widely recommended. They are nevertheless not implemented, and three disabling factors are identified that account for this failure: the cost of genuine precaution as against the cost of simulating it; the absence of any effective method of group sense-making at scales larger than about one Dunbar number (some one hundred and fifty people); and the systematic disempowerment of technical knowledge relative to decision-making power. The conclusion follows directly. A multi-polar trap is a situation in which each party's sensible action produces an outcome every party would reject. No class of chronic systemic multi-polar trap problem, existential risk (a risk that could end humanity, or all of life) from artificial intelligence among them, can be resolved in the absence of new large-scale governance and economic systems that are genuinely effective at group

> sense-making, choice-making, and the enablement of agency; fractal and scalable; robustly anti-corruptible; and genuinely resolving of the principal-agent problem. Without such systems, inexorable civilizational collapse is the expected outcome. It is inexorable in this sense: while the traps stay unresolved, nothing in the process stops it.

I. The Shape of the Question

When a new technology is introduced into a society, a set of questions arises that appear at first to be questions about that technology. Is it safe? What can it do? Who will benefit, and who will bear the cost?

These are the right questions to ask. They are not, however, the questions that determine the outcome.

The outcome is determined by a prior set of facts about the situation into which the technology arrives: who is funded to speak about it, what biases operate on those who assess it, where its costs come to rest, and what capacity for collective judgment the receiving society actually possesses. These facts are structural. They are the same for every new technology, and they are the same whether the technology in question is beneficial or catastrophic.

This is the essential observation. The forces that govern the introduction of new technology do not track the merits of that technology. They track the incentive geometry of the actors involved. An analysis that begins from the merits will therefore mispredict, and will mispredict in a consistent direction: it will expect deliberation where there is none, and it will attribute failure to ignorance or bad faith where the failure is in fact structural and would have occurred even had every participant been well-informed and well-intentioned.

> **On the level of the analysis**
>
> Every claim in what follows is a claim about structure, not about persons. The argument does not require that any developer, investor, regulator, or advocate act in bad faith. It requires only that each act in accordance with the incentives that bear upon their position. This is the sense in which the failure is structural: it is reproduced by the ordinary and expected behavior of ordinarily well-intentioned participants. Where the argument is strongest, it is strongest because it does not depend on any allegation of malice.

II. The Asymmetry of Advocacy

Consider first who is funded to speak.

A group that advocates *for* a new technology has a source of funding available to it that is structurally reliable: the future profits of that technology. An investor who expects a return has a direct and calculable interest in the technology being adopted, adopted quickly, and adopted without conditions that reduce margin. Advocacy is therefore an investment with an expected return, and it is funded as such.

A group that protests *against* a new technology has no equivalent source. Resistance produces no product, captures no market, and returns no multiple on capital. Whatever it protects (public health, ecological integrity, social cohesion, the future) is a good that is held in common and from which no particular party can extract a proportionate private return. Resistance is therefore funded, when it is funded at all, out of donation, out of volunteer labor, and out of the residual attention of people who have other obligations.

The claim that advocacy makes has a shape that deepens the asymmetry. For any new technology or material X, there will always be a claim of the form: X has many widely beneficial applications. This claim is very often true. That is what makes it dangerous. A false claim can be refuted.

A true claim that does not tell the whole story cannot be refuted; it can only be supplemented, and supplementation is expensive. The cost of supplementing falls on resistance, which is the side with no reliable funding (*The Instruments of Choice*, Chapter Three, Section II).

One might expect a countervailing force from incumbent interests whose position the new technology threatens. Such interests do sometimes fund resistance. But this is an unreliable corrective and a partial one. It is unreliable because an incumbent that has attained market dominance has typically exhausted its investment capital in attaining that dominance, and has committed its remaining capital to defending the existing position rather than to contesting a new one. It is partial because incumbent resistance opposes the new technology only where the new technology threatens the incumbent's revenue, which is not where the technology threatens the commons. An incumbent will fund resistance to a competing product. It will not fund resistance to a shared externality from which it also benefits. The second reason is enough by itself: what the commons needs defended is exactly what the incumbent will not pay to defend. In artificial intelligence the largest incumbents are themselves the developers, and so fund advocacy, not resistance. A reader will ask how resistance can be the side with no money, when large sums are spent on the safety of artificial intelligence. Money for safety work that comes from those who build the technology, or who profit from it, is paid out of the technology's expected profits, and funding that comes from those who profit from a technology is the advocacy asymmetry itself, as "Anti-Democratic Risk", Section XIV, finds for the study of existential risk; "Three Problems with AI Self-Regulation", Section II, shows what becomes of the safety work such money pays for once the founder holds the capital, the people, and the product.

This *structural advocacy asymmetry* is therefore not incidental. It is a direct consequence of the fact that benefit is appropriable and harm to the commons is not. Wherever that condition holds (and it holds for essentially every technology whose risks are systemic), the resources available to advocacy will exceed the resources available to resistance, and will do

so by a wide margin, regardless of the actual risk-benefit profile of the technology. The asymmetry, and the larger argument this essay builds on it, is also set out in *The Instruments of Choice; Five Essays on Why Our Tools for Choosing Together No Longer Fit*, Chapter Three, "The Advocacy Asymmetry", which works it through one case, the history of leaded gasoline. That case follows.

> **One case**
>
> Leaded gasoline shows each step of the argument. Gasoline with tetraethyl lead in it was on sale from 1923. Charles Kettering opened the Kettering Laboratory with support from the Ethyl Corporation, DuPont and others, and named Robert Kehoe its director. For the next fifty years virtually the only money for research on lead came from industry, and most of it went to Kehoe (Needleman, 1998). That is advocacy funded as an investment. Representatives of the Ethyl Corporation visited the geochemist Clair Patterson, who said they had tried to "buy me out through research support that would yield results favorable to their cause". He answered with a lecture predicting that Ethyl's activities would be shown to be poisoning the environment and people. Afterward his long-standing contract with the Public Health Service, and his contract with the American Petroleum Institute, were not renewed (Needleman, 1998). That is resistance resting on contracts that could be withdrawn, and were. Algeria, the last country whose service stations sold leaded gasoline, stopped in July 2021, and the United Nations Environment Programme announced the end of its use worldwide on 30 August 2021 (United Nations Environment Programme, 2021): nearly a century between deployment and the end of sale.

III. The Compounding of Biases

To this asymmetry of resources is added an asymmetry of disposition.

Optimism bias operates on everyone involved. The developer believes the technology will work as intended. The investor believes the market will materialize. The user believes the benefit will be personal and the harm, if any, distant. Each of these beliefs is individually defensible and collectively systematic.

Action bias operates on institutions. To act is legible; to refrain is not. A regulator who permits a technology that later causes harm is judged on a harm that is visible and attributable. A regulator who blocks a technology that would have produced benefit is judged on nothing at all, because the foregone benefit never appears in any ledger. This asymmetry in accountability should produce caution. In practice it produces the opposite, because the constituency for action is organized and funded, and the constituency for restraint is not.

Commercialization bias operates on the framing. Once a technology has been capitalized, the vocabulary in which it is discussed becomes the vocabulary of markets: adoption, growth, scale, penetration, share. This vocabulary has no term for harm to the commons except as an externality to be managed, which is to say, as something that has already been categorized as outside the accounting.

To these must be added the network dynamic sometimes named for Metcalfe: that the value of a networked technology to each participant increases with the number of participants. Where this holds, early adoption produces pressure toward further adoption, and the pressure is felt as advantage by the adopter and as disadvantage by the abstainer. The result is that the decision to adopt ceases to be a free decision. It becomes a decision made under coercion by the accumulated adoptions of others.

The *compounding of biases* has a specific consequence. It is not that risks are assessed and then discounted. It is that risks are not assessed at all, because there is no actor positioned to bear the cost of assessing them and no institution positioned to act on the assessment if it were made.

IV. The Reversal of the Defensive Baseline

In the ordinary case, where the technological capabilities of attacker and defender are approximately equal, the defender holds the advantage. The defender chooses the ground, knows the terrain, has prepared in advance, and needs only to hold a position that the attacker must take. This is why the ordinary state of a society is one of relative stability rather than one of continuous successful predation.

This baseline holds only under the condition of approximate parity. Where a new technology confers asymmetric advantage, enabling an action whose cost to the actor is far below the cost of defending against it, the *reversal of the defensive baseline* occurs, and the attacker holds the advantage.

Artificial intelligence is the presently salient example. The relevant asymmetry is not merely that the technology is powerful. It is that the technology exceeds the cognitive capacity of most defenders to comprehend, monitor, or manage. A defense that cannot be understood by those who deploy it is not a defense. It is a dependency, and dependencies are themselves attack surfaces. Bruce Schneier wrote on this asymmetry between attacker and defender in networked systems (Schneier, 2000), and Nick Bostrom and Stuart Russell on the asymmetry of capability between such a system and those who oversee it, and on what it means for the problem of alignment (making a system's aims and conduct match what people need of it) (Bostrom, 2014; Russell, 2019).

The situation this produces has the structure of a *multi-polar trap*. Each participant, reasoning correctly from their own position, concludes that they must develop the capability, because the capability will exist whether or not they develop it, and the position of one who has it is preferable to the position of one who does not. Each participant is right. The aggregate of their correct reasoning is a race that no participant wanted and from which no participant can unilaterally withdraw.

On the word "trap"

The term multi-polar trap is used here in its precise sense and not as a synonym for "difficult coordination problem". A trap is a situation in which the locally rational action of each party produces an outcome that every party disprefers, and in which no party can improve the outcome by changing their own action alone. The distinguishing feature is that exhortation is useless. Telling participants to behave better does not address the structure that makes worse behavior rational, and a participant who complies unilaterally is simply removed from the field in favor of one who does not. Only a change in the structure changes the outcome.

What makes this a trap, rather than merely a collective action problem, can be stated exactly. A collective action problem can in principle be solved by agreement: if the parties can communicate, bind themselves, and verify compliance, the trap dissolves. The multi-polar trap persists because those three capacities are absent at the scale at which the problem exists. The trap is thus not a failure of goodwill. It is a deficiency of governance instrument (*The Instruments of Choice*, Chapter Three, Section VI), and the governance specification of Section VIII is a statement of the instrument that is missing.

V. Where the Costs Come to Rest

The fourth structural force concerns the distribution of consequence.

The benefits of a new technology accrue locally, promptly, and to identifiable parties: to the developer, to the deployer, to the early user. The costs and risks accrue diffusely, slowly, and to the commons: to the ecosystem, to the information environment, to social trust, to future generations, to those who did not adopt and had no say in the adoption.

This distribution is not a market failure in the sense of an aberration to be corrected. It is the expected result of an accounting system in which appropriable value is measured and non-appropriable value is not. A cost that appears in no ledger is, for the purposes of every decision made by every actor, a cost that does not exist.

Economics has a standard name for this: externality, a cost not reflected in the price of the transaction that produces it. The standard treatment holds that externalities can be internalized by taxation, liability, or the assignment of property rights. Each of these remedies presupposes a functioning apparatus of collective identification and enforcement, operating on a timescale comparable to that of the harm. Where the harm is slow, distributed, and contested, and where the apparatus of enforcement is itself subject to the funding asymmetry of Section II, internalization does not occur. The externality is a consequence of the boundary of the market having been drawn in a particular place, and it can only be addressed by redrawing that boundary, which is a governance act, not a market act (*The Instruments of Choice*, Chapter Three, Section IV).

The *externalization* is therefore self-reinforcing. Each successful externalization increases the returns to the externalizing party, which increases the capital available to that party, which increases the resources it can direct toward advocacy, which returns us to the asymmetry described in Section II. The four forces are not four independent problems. They are one structure viewed from four positions.

The pace of deployment bears on where the costs come to rest. The use of new technologies, and the rate at which technologies of every kind are deployed, is growing at a rate without precedent in history or in biology. The growth runs roughly parallel to the deployment of new materials, metals, chemicals, pesticides and plastics into the whole living environment, where over time they affect plants, animals and humans, insects, birds, fish and microbes, and the long-term cycles of all of these, their dependence on one another, and their evolution. The pace matters because harms that are delayed and diffuse are found only after a technology has

been in use for some time. The faster new technologies arrive, the more of them are in use before the harms of the earlier ones have been found, and by the time a harm is found, the technology that causes it is depended upon, so that withdrawing it is harder than declining it would have been.

The same distribution of benefit and cost holds across time. From the position of any one actor deciding for itself, and given inevitable and inherent spending inefficiencies, it always pays better to borrow significantly against costs in the future, for local benefits in the present that increase inequality, than it does to be conservative and hold and store value for the (even one's own) future. So the costs are carried forward in time, onto those who come later, as they are carried outward onto the commons, and for the same reason: the party who decides is not the party who pays.

The Burden of Proof

Where a technology gives the attacker the advantage, it matters who must show that an AI system is safe before the system is built.

The reason it matters is given by Section IV. Where the technology exceeds the capacity of defenders to comprehend, monitor, or manage it, the defenders are the parties least able to show whether it is safe or not. A rule that left the showing to them would leave it to the parties who cannot do it, and the system would be built because nobody had shown that it should not be.

Where the Burden Lies

The burden of proof is on the engineers who are building AI systems, and on those who are suggesting designs for AI systems. They must *themselves* investigate, rigorously, whether the systems and methods they propose fall short of being robust enough against well-known destabilizing dynamics. This holds for any potentially human-lethal system that they are involved in building or designing, at *whatever* level of practice.

It follows that those who question a proof of impossibility should be especially careful not to implicitly shift the burden of proof.

The shift is easy to make without noticing. A person who answers a proof of impossibility by finding a step that has not yet been shown for every case treats the author of the proof as the party who must establish that the system is unsafe. The builder of the system, meanwhile, has established nothing about its safety. A claim that one step of an impossibility proof fails, even a correct claim, does not move the burden: it leaves the burden where it was, on the party proposing to build.

What the Burden Requires

The burden is to establish (to prove) that the system or method being suggested is actually robust enough: to show that the system can fully account, in known, established and accepted ways, for the total class of all known issues, and to show it *before* recommending the scheme.

The alternative is to overlook all known impossibility issues and to try to make AGI (artificial general intelligence) robust anyway. That action is one of either delusion or deception, and *neither* is acceptable. It is delusion where the builder does not see the known issues, and deception where the builder sees them and presents the system as robust anyway.

A reader will ask whether anyone could meet this standard: to account for the total class of all known issues. The standard applies to any potentially human-lethal system, and where it cannot be met, the consequence is the one Section VIII names: the choice not to build what cannot be made safe.

Why the Burden Is Not on Others

People *other than* the engineers attempting to build AI are not the ones who should be implicitly or tacitly "required" to produce many pages of safety documentation, proofs, and the like. It is not for them to verify or establish whether a type of system that *the engineer* is making or designing will be non-robust, non-safe, or non-aligned.

The person who performs the action of making is responsible for the outcome of that making.

All the safety engineers involved are responsible for seeing that this work is done from the beginning. That is an ethical responsibility, not just a social or moral one. It is ethical in the sense that it comes from the act of making and from the effects of that act on others, and does not depend on what others expect or on the customs of the field.

Why the Burden Follows from the Structure

The four structural forces of this essay give the same answer from their own side.

Only the builder holds what the showing requires. The design, the training record, and the record of decisions are held by the organization that builds the system, and Essay 2, Section VII, shows that this is the material an organization has the strongest reasons to withhold. A party outside the organization can be asked to prove the system unsafe only from the outside, without the evidence that would settle the question.

The parties who would carry a shifted burden are the parties least funded to carry it. Section II shows that advocacy for a technology is funded out of its expected profits, and that resistance is funded out of donation and spare attention. To require those who question the system to prove it unsafe is to assign the most expensive work to the party with the fewest resources to do it, and work assigned that way is seldom done.

And a shifted burden is itself an externalization, in the sense Section V gives the word. The cost of establishing safety is a cost of the building, and its benefit goes to the builder; placing that cost on others places it on the commons, while the benefit stays where it was. The rule that the burden lies with the builder is the rule that the builder bears the costs of the building, applied to the cost of finding out whether it is safe to build.

VI. What Ought To Be Done

The remedies that follow from this diagnosis are neither obscure nor controversial. They have been articulated repeatedly, in substantially similar form, by many parties over many decades. They are set out here in their standard form, so that the argument of the following section may be seen clearly for what it is. The form is that of the four high-level recommendations of the 2018 report *The Malicious Use of Artificial Intelligence* (Brundage and others, 2018), generalized here from artificial intelligence to new technologies, with one clause added to the second and one at the end of the last.

Policymakers should collaborate closely with technical researchers to investigate, prevent, and mitigate the potential malicious uses of new technologies. Researchers and engineers should take seriously the dual-use character of what they build (that the same capability serves harm as readily as benefit), should permit misuse-related considerations to shape research priorities and professional norms, should engage the relevant agencies proactively rather than reactively, and should ensure that mitigation capabilities are developed alongside the capabilities that require mitigating. Best practices should be imported from those research domains in which dual-use risk management is already mature. And all parties should expand the range of stakeholders and domain experts involved in the discussion, while enabling those groups with sufficient agency to implement whatever policy the consensus identifies as balancing benefit, risk, harm, and cost across all affected parties.

Each of these recommendations is correct. None of them is new. And the fact that they are correct, and not new, and nevertheless not implemented, is the fact that requires explanation.

VII. The Three Disabling Factors

Three factors account for the gap between what is recommended and what occurs. Each is independently sufficient to produce the gap. Together they make it structural.

The first is the cost of precaution as against the cost of simulating precaution. Genuine precautionary governance is expensive: socially, procedurally, and economically. It requires that competent people spend substantial time on the analysis of consequences that may never materialize, that decisions be delayed while that analysis is conducted, and that some decisions be made against the immediate commercial interest of the party making them. *Simulated precaution* is cheap. It requires a committee, a report, a set of published principles, and a communications function capable of describing these as a process. Both produce the same visible artifacts. Only one produces the substance. Where the two are indistinguishable to the observer and differ by orders of magnitude (many factors of ten) in cost, the cheaper will be selected by every actor subject to competitive pressure, and no actor will be able to afford the more expensive unless all are compelled to it simultaneously. What an observer would have to look at in order to tell the two apart, and why every such observable is itself capturable, is the subject of Essay 2.

The second is the absence of effective group sense-making at scale. Human beings possess methods by which a group can think together: methods by which the members of a group can arrive at a shared understanding that is better than the understanding any member brought to the conversation. These methods are real and they work. They work at approximately the scale of one Dunbar number (named for the anthropologist Robin Dunbar). The Dunbar numbers are a set of approximate group sizes, each marking a transition in how the members of a group can relate; the one meant when a single number is named is about one hundred and fifty persons, the size of group within which each participant can hold a

working model of every other. The exact figure is debated (Dunbar, 1992; Lindenfors, Wartel, and Lind, 2021); the argument needs only that the bound is set by the participants and is far smaller than the populations a technology affects. Above that scale, no such method presently exists: the bound belongs to the participants, so a method for larger numbers would have to compose small groups, not enlarge them. This is the *Dunbar limit on sense-making*.

A reader will ask why many small groups cannot simply decide and have their decisions added up. Adding up decisions is not composition: a tally passes on what each group chose and not why, and a group that receives a bare conclusion can assess it only by trusting its source. Composition requires that whatever passes between groups be assessable on its content, the condition the Afterword, Section III, sets out.

The familiar case shows the difference. A workshop of a hundred and twenty people divides into ten teams of twelve, each of which governs itself well, and each of which sends one member to a council of ten. The council reaches a decision the council believes in. The decision returns to the teams as a conclusion with a summary of reasons attached, and in most of the teams the summary does not persuade, because what persuaded the council was six hours of argument that the summary is not. The teams then comply or they do not. Where they comply, the process is no longer consensual; where they do not, the council was decorative. A representative can carry the conclusion. The representative cannot carry the relationship that made the conclusion binding. Nor can the loss be recovered by asking a better question from above, because the information required to formulate the right question is precisely what was discarded: the upper level does not know what it is missing, and it therefore experiences a compressed picture as a complete one (*Governance Beyond the Dunbar Number*, Chapters Six and Ten).

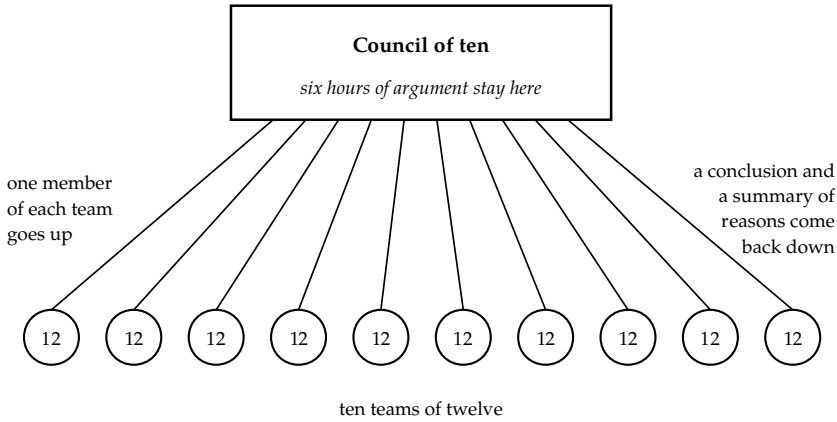


Figure 1. The council of ten (*Essay 1, Section VII*). *Each team of twelve sends one member to the council; the six hours of argument stay in the council, and what returns to the teams is a conclusion with a summary of reasons.*

That book names *accretion* any strategy for governing a larger body that keeps the architecture of the small unit intact and reaches size by combining units: nesting them, layering them, federating them, delegating between them, or overlapping their memberships; and it reports that whole family as failing. Composition, in the sense used here, is a different thing. It does not carry a small unit's position upward intact; it requires that what passes between units be output, the reasoning and not only the conclusion, that a receiving unit can judge on its content. What decides between the two is what crosses the boundary: each of those shapes is accretion where what crosses is a position carried by someone who stands for the unit, and can be composition where what crosses is reasoning that the receiving unit judges for itself.

Different types of collective decision-making have different scaling properties. The limit applies to the formation of shared understanding; the aggregation of expressed preference does scale, and it is what is inadequate to the problem. What exists above that scale, in place of shared understanding, is representation, aggregation, and voting, which are mech-

anisms for the transmission of preference, not mechanisms for the formation of understanding. A vote can select among options that have been placed before it. It cannot notice that the correct option is not among those on the ballot.

That holds of every voting rule, and not of some particular rule. A voting rule is a function from a profile of preferences over alternatives to a selected alternative. In plain words, a voting rule takes what the voters prefer among the choices offered, and returns one of those choices. The alternatives are an input; where they came from is outside the function's domain. Whoever designs the option set therefore holds the substantive power, and holds it at one remove from any accounting, and ranked, approval, and other refined ballots all accept the option set as given (*Governance Beyond the Dunbar Number*, Chapter Ten, Section V). Kenneth Arrow's impossibility theorem (Arrow, 1951) establishes that no ranked voting rule can meet a small set of reasonable conditions of fairness all at once, so that the selection step cannot be made flawless; this argument establishes that flawless selection would not be enough, because the problem lies before the selection step rather than within it.

Counting also conceals its own inadequacy. A procedure that maps each judgment onto a mark, and lets the larger count prevail, always terminates: whatever is fed into it, something comes out, and the something that comes out has the form of a decision. A procedure which always terminates gives no signal when it has been applied to something it does not fit. It reports a number, and the number looks the same whether the mapping was faithful or arbitrary (*The Instruments of Choice*, Chapter Four, Section II). Counting scales in the sense that the arithmetic remains valid. It does not scale in the sense that the answer remains true. The governance of a global technology is therefore attempted with instruments that were never designed for the formation of collective judgment and are not capable of it.

A reader who knows Elinor Ostrom's work will ask whether it shows that a commons can be governed without a central authority. It does, and it fits this argument. Her findings qualify the pessimistic reading of the

commons that Garrett Hardin gave, on which a resource open to all is overused until it is ruined (Hardin, 1968; Ostrom, 1990). The communities she studied govern their commons through local bodies whose members watch and answer to one another directly, and govern larger systems through what she called nested enterprises: small units composed into larger ones. That is composition, the kind of scaling the Afterword, Section III, requires, in which the unit stays small and the number of units grows, in the measure that what passes between the units can be judged on its content by the unit receiving it; where the nesting runs through representatives carrying a compressed position, it is the accretion described above, applied within the narrowed domain of a single resource. Her cases show composition at the scale of irrigation systems, fisheries, forests, and pastures; composition at the scale of a technology whose effects cross every boundary is what has not yet been built. The lasting commons she studied share, among other conditions, clearly defined boundaries, monitoring by the users themselves or by monitors answerable to them, and graduated sanctions applied by fellow users. Among nuclear-armed states, the case *The Architecture of Civilization*, Chapter 4, examines, no boundary closes the commons, since every state's weapons bear on every other; no party can see what the others hold in secret; and no sanction can be applied to a party whose force no one can overrule. A technology whose effects cross every boundary has the same shape.

The third is the disempowerment of knowledge relative to decision. Within any technology-developing organization, the persons with the deepest understanding of what is being built and what it is capable of are the researchers and engineers at the face of the work. These are systematically the least empowered persons in the decision structure. The persons with the most decision-making power are the investors, who are the most remote from the technical understanding, whose fiduciary orientation (a duty, owed in law, to those whose money one manages) is toward shareholder value, and whose time horizon is set by the fund's return schedule rather than by the deployment lifetime of the technology. This is the *knowledge-authority gradient*, and it runs in exactly the wrong direction: author-

ity increases as understanding decreases. Distance from consequence is here not incidental but is itself the reward of success. The structure does not merely fail to place knowledge and authority together; it actively separates them, and the separation increases with the scale of the decision (*The Instruments of Choice*, Chapter Three, Section VIII). An engineer who identifies a serious risk must escalate that risk through a chain of persons each of whom understands it less well than the last and each of whom has a stronger incentive to discount it.

Each factor produces the gap on its own. Suppose imitation could be told from precaution, and authority followed understanding, but no method of sense-making worked above about one hundred and fifty people: no actor could form a view on behalf of all affected parties, and the recommendations would still have no one to carry them out. Suppose sense-making worked at scale, and authority followed understanding, but imitation stayed cheap and indistinguishable from precaution: the filter of Essay 2 would remove the organizations that bore the cost, and what the scaled sense-making observed would be artifacts. Suppose sense-making worked at scale, and imitation could be told from precaution, but inside each developing organization authority still rose as understanding fell: the risk seen at the face of the work would be discounted on its way up, and would not reach the parties who decide, inside the organization or outside it. So all three must be overcome together.

The claim is not that precaution is difficult. The claim is that the recommendations of Section VI are addressed to a decision-making capacity that does not exist. They presuppose an actor who can weigh benefit against systemic risk on behalf of all affected parties and then act on that weighing. No such actor is present at any scale at which the relevant technologies operate. The recommendations are therefore not wrong. They are unaddressed.

VIII. The Requirement

From the foregoing, a requirement follows, and it is a stronger requirement than is usually acknowledged in these discussions.

Each of the three disabling factors of Section VII calls for at least one of the properties that follow:

Disabling factor (Section VII)	**Property it calls for**
Imitation of precaution is cheap, and cannot be told from precaution by its artifacts	Robust anti-corruptibility, so that imitating or capturing the signs of care costs what the substance costs; and resolution of the principal-agent problem, since otherwise "the party observed is the party supplying the observation" (Essay 2, Section VIII)
No method of sense-making works above about one hundred and fifty people	Effective sense-making, choice-making and agency; and fractal scalability, so that the methods that work among a few reach world scale by composing small groups
Authority rises as understanding falls	Resolution of the principal-agent problem, so that those who decide bear what their decisions cause

The resolution of any chronic systemic multi-polar trap problem requires large-scale governance and economic systems possessing four properties simultaneously.

They must be genuinely effective at group sense-making, group choice-making, and the enablement of group agency: that is, they must produce shared understanding, shared decision, and the capacity to act on the decision, and must produce all three, since any two without the third are inert.

They must be fractal and scalable: capable of operating across world-scale populations by the same principles by which they operate at the scale of a small group, rather than by substituting a different and weaker mechanism above some threshold. The Dunbar limit of Section VII is one reason for this property. The requirement of requisite complexity is a second, independent one, and it applies to governance the law of requisite variety from cybernetics (Ashby, 1956): a regulator must be able to distinguish at least as many states as the regulated system can occupy. Representative democracy compresses the entire state of a population's judgment into a small number of periodic choices among a few options; under the Constitution of the United States, a voter chooses a member of the House of Representatives once every two years, a senator once every six, and a president once every four, while in every year from 1976 to 2013, by the count of the Congressional Research Service, the federal agencies published at least 3,503 final rules in the Federal Register. No governance form can regulate what it cannot distinguish. A single central apparatus cannot hold the complexity of what it governs. A structure that replicates governance capacity at every level of organization distributes the requirement across many sites, each of which need hold only the complexity present at its own scale. The fractal form is thus the shape a governing structure must take if it is to have sufficient complexity without requiring any single part of itself to hold all of it (*Covenant and Commons*, "Covenant Before Contract", Section VI).

They must be robustly anti-corruptible: structured so that the attempt to capture them is not merely prohibited but ineffective, since prohibition depends on enforcement and enforcement is itself capturable. Ineffective means here that the attempt costs the party attempting it more than it buys; Essay 2, Section VIII, states this as the condition it calls capture-cost parity.

And they must genuinely resolve the principal-agent problem: they must be built such that the interests of those who decide are actually aligned with the whole ecosystem, and not merely with locally or temporally extractable value. In the ordinary sense of the problem, a principal

hands a task to an agent whose interests differ from its own and whose conduct it cannot fully see; in the specification, the principal is the whole set of parties a decision affects, and the agents are those who decide.

Together these four are the *governance specification*. It is a specification, not a design: it states what is necessary, the four properties any adequate system must have. But the specification itself is not optional and does not admit of partial satisfaction. A governance system that achieves three of these four properties is taken through the one it lacks. The four ways it is taken are different failures: displacement (a faster source of answers replaces the system), boundary capture (the channel that carries a small group's judgment to the scale of decision is taken), acquisition (the working capacity is bought), and drift (the parties within steer it toward what they can extract); the Afterword, Section VI, takes each in turn. Essay 4 follows one such case in detail: a capacity that satisfies the first property (that the system form understanding, choose, and act) well enough to be worth having, and is acquired through the third (that capture be ineffective).

A reader will ask whether four properties at once is too much, and whether they could be achieved one at a time. The same form of argument appears, for three other problems (the principal-agent problem, the rules-for-rulers dynamic, and the multi-polar trap), in *The Architecture of Trust*, "The Three Unsolvable Problems", Section IV, and "The Convolution", Section III: no pair of the three can be solved unless the third is solved with it, because the remedy for each runs through the other two, and in each case the problem left standing undoes what was done for the other two. The simultaneous solution is the smallest solution. The scope that looks like the largest possible undertaking is in fact the minimum scope at which the work can be done at all, which inverts the ordinary engineering intuition that one makes a hard problem easier by cutting it into pieces. The four properties here are a different set, and the parallel supports the shape of the argument: a sequence of partial solutions does not add up to the whole one, since each is taken through what it lacks before the next arrives.

The consequence of not meeting the specification should be stated plainly. In the absence of such systems, no class of chronic systemic multipolar trap problem can be resolved. Existential risk from artificial intelligence belongs to this class, as do the several other risks that arise from technologies of comparable reach. For that risk, resolution does not mean that the technology is made safe: the books that carry the control argument (*Why AGI Cannot Be Made Safe; The Argument, Its Physics, and the One Choice That Remains*, *Ape Trap; The Runaway Artificialization of Our Planet*, *The Ceiling Below the Floor; The Containment Fallacy and AGI Non-Safety*, and *Why No Control Suffices; The Technical Case for the Impossibility of AGI Safety*) show that no control is adequate to it. Resolution means that the parties affected can form a shared understanding of the risk, choose on that understanding, and act on the choice, including the choice not to build what cannot be made safe. The expected outcome of the unresolved case is not stagnation and not decline. It is civilizational collapse, in this sense: what collapses is the capacity of a society to address the traps it is in, because each unresolved trap degrades the capacity to address the next. Joseph Tainter's study of the collapse of complex societies, in which each further investment in complexity returns less than the one before, bears on this claim (Tainter, 1988). It is inexorable in this sense: while the traps stay unresolved, nothing in the process stops it. That is a claim about the default trajectory, and the default is what the specification exists to change.

A reader will ask whether the treaty that ended the chemicals depleting the ozone layer did not resolve such a problem with the institutions that already exist. It did so because that case left the trap. The harm was measured and could be traced to its source, the producers were few, and substitutes made restraint cheap, so that no party lost by restraining alone; these are the conditions of Essay 2, Section IV, under which the filter runs the other way, and by the definition in the sidebar to Section IV ("no party can improve the outcome by changing their own action alone") the situation had stopped being a trap.

A reader will also ask whether one dominant power, or a treaty among the largest powers, could impose the change. A power that imposes it is a center, with a single point of capture (Essay 4, Section IV), and a party that can leave the set of those affected (Afterword, Section V), so the change it imposes lacks the third and the fourth properties and is taken through them. A dominant power also needs a surplus of force over any party that might break the rules, and that surplus is gone once weaponry has reached the threshold of wars that cannot be fought: no nation is willing to disarm first and become the party that is then vulnerable to coercion, which is why nuclear non-proliferation has failed to end the standoff among nuclear-armed rivals (*The Architecture of Civilization*, Chapter 4).

A treaty among the largest powers is correct in principle, and what it requires in practice is that each party can verify the others' compliance, and that some party can enforce the terms against one that defects. The capacity to verify is bounded by what one already knows about what one would be looking at, and the parties most capable of concealment are generally the parties most capable in the domain. Enforcement requires an authority above the parties, which is either absent or is itself an agent holding its position by the support of those able to remove it (*The Architecture of Trust*, "The Three Unsolvable Problems", Section III). Each requirement leads back into the trap.

What Technology Is, and Whom It Serves

The governance specification states what must govern the introduction of a new technology. This section states what technology is, and whom technology has to serve.

Technology, in the sense used here, is anything that is designed from the outside in, or from the top down. There is a conception, and there is a use of causation: science is the discovery of causation, and technology is the application of causal knowledge to some aim. The two descriptions come to one thing: a conception imposed from outside, carried out by causation.

A design made from the outside in begins from a conception of what the thing is for, and imposes that conception on materials and on a place. What the conception names, the design accounts for. What the conception does not name (the soil the materials came from, the community the product enters, the living systems that will receive its waste) the design does not account for at all. This is where the costs of Section V come to rest. A cost placed on the commons is, from the side of the design, the part of the world its conception left out.

The same shape explains why such costs go unseen for so long. Technology is a top-down process: it starts with concepts at a macroscopic level and then figures out the microscopic details as a kind of side effect of what the macroscopic intentions are. Evolution works the other way around: it works at the microscopic, and slowly, over time, it develops more and more macroscopic structures. From the top down, it is very easy to overlook accumulations and depletions occurring at a microscopic scale. A technology may involve some trace elemental material; the depletion of that material, and the accumulation of toxicity from it elsewhere, will actually happen, and can go on for quite some time with no real detection that it is occurring (*Consciousness, Perception, and Nature*, Chapter 7). Looked at from the top down, it is very easy to overlook the ten million and one details that are actually relevant to getting it right.

If a technology becomes sufficiently nature-like, is it still technology? In the sense meant here, it is no longer top down, and it is no longer only causation. A design that accounts for nature accounts for change, and so for choice: it allows for resilience and adaptability, and therefore for integration. Top-down design is a continuation of the Western model, and it carries an element of colonization and inequity: it says, from the outside, "I am going to tell you what you are about".

The place where the two kinds of design meet can be stated. If a technology's toxicity is fundamentally a function of missing de-synthesis processes (the processes that take a made thing back apart), then reducing that toxicity means building the missing half of the cycle. The more com-

pletely that missing half is built, the more the technology, as a whole system including both its synthesis and its de-synthesis, comes to resemble the behavior of a natural, living process: one where nothing is built without an implicit and eventually actual path back to its constituent parts. The frontier where top-down engineered design and bottom-up natural design actually meet is the place where synthesis and de-synthesis have been brought back into a matched, self-closing cycle, whatever the origin of the components involved. Many, perhaps most, real de-synthesis processes depend at some point on a manual or human-involved step: someone choosing to carry out the disassembly, the sorting, the return. Where that dependency exists, the reliability of the whole material cycle becomes a question about human follow-through rather than about chemistry or engineering, and the weak link in the cycle is social, not technical (*Sense-Making and the Craft of Inquiry*, Chapter 19, Section IV). A technology in service to nature therefore depends on the same capacity for choosing together that the rest of this volume examines.

The relation among man, machine and nature shows the same thing. Man has the illusion that technology and nature are both in service to man, while in fact technology extracts from nature, so that nature serves technology so that technology can serve man. That arrangement is not a cycle. It ends in the depletion of key elemental resources in some places and a superabundance of things that ought not to be there in others, and anywhere there is too much or too little of something, that is toxic. Technology has to be in service to nature, so that nature is healthy enough to support us.

The service to man that the illusion describes cannot be had on its own terms. "Man" is not a thing that can be served independently of the conditions that sustain him. A human being is a process that runs on an atmosphere of a certain composition, a hydrological cycle of a certain reliability, a soil column of a certain depth, and a web of other organisms that maintain all three. The conditions are not the setting in which the beneficiary stands; they are part of what the beneficiary is. A pump installed over an aquifer shows the instance. The yield per hectare rises, and every meas-

urement available at the time of installation records an improvement. The yield figure is a real number, and it is a proxy: it tracks human nutrition faithfully across an interval shorter than the recharge time of the aquifer, and diverges from it across any interval longer. The divergence was present in the arrangement from the first day, recorded nowhere, because the account on which it would have been recorded was not being kept (*The Non-Negotiable*, Chapter Thirteen, Section I).

The way in which technology takes from nature has a pattern of three moves, set out at Essay 5, Section V, and developed in *The Geometry of Man, Machine, and Nature; Care, Community, and the Design of Technology*: extraction takes a thing out of its context, abstraction makes it a member of a class that can be counted and compared, and accumulation reduces it to a number that adds without limit. Each move is a further step away from the place the thing came from, and so a further step away from any accounting of what its taking cost that place.

This section and the governance specification state one requirement from two sides. The fourth property asks that the interests of those who decide be aligned with the whole ecosystem, and not merely with what can be extracted from it. Technology in service to nature is that same alignment, required of the thing that is built: a design that accounts for the context it enters, so that its costs do not come to rest outside its own account. The specification is what would let a society require that of what it builds.

IX. Summary

The four structural forces of Sections II through V each rest on political economy and on the observable history of technology introduction. The advocacy asymmetry follows directly from the appropriability of benefit and the non-appropriability of harm to the commons, and requires no further premise. Section II gives one case whole, leaded gasoline, in which advocacy was funded as an investment and resistance rested on contracts

that could be withdrawn, and it shows that the claim advocacy makes is usually true and partial, so that it can only be supplemented, at the cost of resistance. Section IV states what makes the trap a trap: the parties cannot communicate, bind themselves and verify compliance at the scale of the problem. Section V shows that internalizing the externality is a governance act and not a market act. It adds that the rate of technology deployment is historically and biologically unprecedented, so that harms are found only after what causes them is depended upon, and that the costs are carried forward in time as well as onto the commons.

The section on the burden of proof, at the end of Section V, establishes that the people who build or design a potentially human-lethal AI system are the ones who must show that it is robust against known destabilizing dynamics, and must account for the whole class of known issues before recommending the design. It establishes that building anyway, while overlooking known impossibility issues, is either delusion or deception. It follows that those who question an impossibility proof may not demand that others prove the system unsafe, because the one who acts is responsible for the outcome of the act. The four structural forces give the same answer: only the builder holds the evidence, the parties who would carry a shifted burden are the least funded, and shifting the burden places a cost of the building on the commons. The scope of that section is the responsibility of engineers, designers and safety engineers for the AI systems they make or propose.

The three disabling factors of Section VII are each supported on their own ground. The composite claim, that all three must be overcome together for adequate precautionary governance to occur, follows from the step of Section VII: each factor alone produces the gap, so removing any two leaves the third, which still produces it. The conclusion of the essay rests on that claim. Section VII also shows, by the committee of committees, why a representative's summary cannot carry what persuaded the group, why no voting rule reaches the set of options it is given, and why counting gives no signal of its own failure; and it separates composition from accretion.

Section VIII grounds the fractal property a second time, in requisite complexity; it shows that the four properties cannot be had one at a time; and it answers the ozone treaty, the dominant power, and the treaty among the largest powers: the first left the trap, and the other two lead back into it.

The section on what technology is defines technology as the application of causal knowledge to an aim, designed from the outside in or from the top down; a technology that accounts for nature accounts for change and choice, and so is no longer only causation. It shows that the present relation of man, machine and nature is not a cycle: technology extracts from nature to serve man, which leaves depletion in some places and toxic excess in others, and so technology has to serve nature, so that nature is healthy enough to support us. A design made from the outside in accounts only for what its conception names, and the costs of Section V come to rest in what it leaves out; technology in service to nature is the fourth property of the specification required of the thing that is built. The same section gives the mechanism by which top-down design leaves accumulations and depletions unseen, the place where engineered and natural design meet (synthesis and de-synthesis in a matched, self-closing cycle), and the reason the weak link of that cycle is human follow-through, which brings technology back to the capacity for choosing together.

The essay establishes that collapse is the expected outcome of the unresolved case, which is a claim about the default trajectory, and not a claim about fate or certainty. The impossibility of making artificial general intelligence safe, which the books named in Section VIII carry, is a different claim with a different strength: it is a structural certainty of the kind with which perpetual motion is excluded, and Essay 3, Section V, says so. It argues that the governance systems are necessary and leaves their design to other work; it gives no timeline.

X. Conclusion

The argument of this essay is, at its center, a single observation turned in several directions.

The observation is that the introduction of a new technology is not governed by anyone. There is no seat from which the question "should this be done, and on what conditions?" is asked on behalf of all affected parties, because there is no mechanism by which all affected parties could form a view to be asked on behalf of. What exists instead is a set of local decisions, each made by a party with an appropriable interest in one outcome and no accounting entry for the others.

Once this is seen, the usual complaints about technology governance (that it is too slow, that regulators lack expertise, that the pace of development outstrips the pace of legislation) appear as symptoms rather than causes. They are true, and they are not the problem. The problem is that the thing they are complaints about does not exist to be improved.

The threshold for avoiding the expected outcome is therefore not better regulation of artificial intelligence, nor better norms among researchers, nor better disclosure by developers, though each of these would be an improvement. The threshold is the construction of a capacity that the species does not currently have: the capacity to think together at scale, to choose together on the basis of that thinking, and to act on that choice in a manner that cannot be captured by any party that would prefer a different action.

That is a large thing to require. It is nevertheless what the structure requires, and the structure is indifferent to whether the requirement is convenient. The work of specifying such a capacity precisely, and of designing instruments adequate to it, is the work that follows from this diagnosis, and the essays that follow are severally addressed to it. Essays 2 and 3 concern what stands in the way of stating the case at all: the imitation of precaution, and the cost of establishing anything before an audience with

an interest in its being false. Essays 4 and 5 concern what becomes of the capacity described here once it exists: that it is captured in proportion to its effectiveness, and that in its absence the signal of serious work cannot be told from the work.

XI. What This Essay Adds to the Volume

The volume asks under what conditions the capacity for collective judgment fails, and what would be required for it to succeed at the scale at which the decisions that matter are now made. This essay gives the first answer to both halves of that question.

To the first half it answers with four forces and three disabling factors, each a condition under which judgment fails while every participant is competent and sincere. To the second half it answers with the governance specification: four properties, required together, that any capacity adequate to these problems must have.

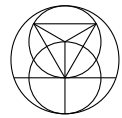
The specification is the measure the rest of the volume uses. Essay 2 takes up the first disabling factor, the cheapness of the imitation of precaution, and shows what it does to every sign an observer could trust; its two companion essays carry that result into the self-regulation of AI companies and into the study of existential risk. Essay 3 takes up what it costs to establish any of this before the parties who would prefer it were false. Essays 4 and 5 take up the third property, and show that the capacity, once built, is captured in proportion to how well it works. The Afterword says what each of the four properties demands of a design. Each of these begins from the requirement stated here: the missing thing is the capacity to think together at scale.

KEY CONCEPTS AND FINDINGS IN THIS ESSAY

- **Structural advocacy asymmetry**: the systematic funding advantage held by advocacy for a new technology over resistance to it, arising because benefit is appropriable by a particular party and harm to the commons is not.
- **Compounding of biases**: the mutually reinforcing operation of optimism bias, action bias, commercialization bias, and network-adoption pressure, which together ensure that risk receives structurally insufficient assessment.
- **Reversal of the defensive baseline**: the loss of the defender's ordinary advantage wherever a technology confers asymmetric advantage on an attacker, particularly where the technology exceeds the defender's capacity to comprehend it.
- **The burden is on the builder**: those who build or design an AI system must prove their system robust.
- **Before, not after**: the proof comes before the design is recommended.
- **No shifting**: those who question a proof of impossibility may not make others prove that a system is unsafe.
- **Delusion or deception**: building while ignoring known impossibility issues is one or the other.
- **The actor is responsible**: whoever performs the action owns its outcome.
- **A shifted burden is an externalization**: requiring others to prove a system unsafe places a cost of the building on the parties least funded to bear it, while the benefit stays with the builder.
- **Externalization**: the placement of cost and risk onto the commons while benefit accrues locally; self-reinforcing, because each successful externalization increases the capital available for further advocacy.
- **Multi-polar trap**: a situation in which the locally rational action of each party produces an outcome every party disprefers, and which no party can improve by unilateral change of action.

- **Simulated precaution**: the production of the visible artifacts of precautionary governance without its substance; selected over genuine precaution wherever the two are observationally equivalent and differ greatly in cost.
- **The Dunbar limit on sense-making**: the absence of any known method by which genuine shared understanding can be formed in groups substantially larger than about one hundred and fifty persons.
- **The knowledge-authority gradient**: the systematic inverse relationship, within technology-developing organizations, between depth of technical understanding and degree of decision-making power.
- **The governance specification**: the four simultaneous requirements on any adequate large-scale system: effective sense-making, choice-making and agency; fractal scalability; robust anti-corruptibility; and genuine resolution of the principal-agent problem.
- **Technology as applied causation**: technology is designed from the outside in; a technology that accounts for nature allows for change and choice, and is no longer only causation. A design made from the outside in accounts only for what its conception names, and its costs come to rest in what the conception leaves out.
- **Technology in service to nature**: technology that extracts from nature to serve man is not a cycle and ends in depletion and toxic excess; technology has to serve nature, so that nature can support us.
- **Accretion**: any strategy that keeps the small unit intact and reaches size by combining units (nesting, layering, federating, delegating, overlapping memberships); distinguished from composition, in which what passes between units is reasoning a receiving unit can judge on its content.
- **Termination without signal**: a counting procedure always yields an answer, and gives no signal when it has been applied to something it does not fit; counting scales in its arithmetic and not in its truth.
- **Requisite complexity**: a regulator must be able to distinguish at least as many states as the regulated system can occupy; the second ground of the fractal property.

- **The simultaneous solution is the smallest solution**: where the remedy for each problem runs through the others, solving them together is the smallest job that can work.



Essay 2

The Simulation of Precaution

On why the artifacts of care are cheaper than the care,
and what would have to be true of an observable to survive it

ABSTRACT

Genuine precautionary governance is a practice with a determinate substance: it changes decisions that would otherwise have been made, and it imposes the cost of that change on the party making it. The visible artifacts of the practice (the committee, the charter, the published principles, the risk register, the staged release, the external review) are producible without the substance, at a small fraction of the cost, and are then indistinguishable from the artifacts of the practice itself. This essay develops that indistinguishability into its consequences. It identifies the selection process by which the cheaper form displaces the dearer among organizations subject to competitive pressure, distinguishes this displacement from fraud by showing that most of it is sincere, and then constructs the positive case: there are observables that would in fact discriminate between precaution and its imitation, and five of them are set out. Each is then shown to be capturable, for a common reason: the publication of a discriminating observable supplies the specification by which that observable can be produced without the substance it was chosen to indicate. The essay closes with what survives this result. An observable resists imitation in proportion to what its imitation costs, so the design question is not which observable cannot be captured, but which arrangement makes the capture of an observable cost what the substance costs. That is a structural question about where authority sits and whose interest it serves,

> and it returns the problem to the fourth property of the specification given at Essay 1, Section VIII (that those who decide bear what their decisions cause).

I. What Precaution Is For

Precautionary governance is a practice with a substance, and the substance can be stated exactly.

A precautionary process is one that changes decisions. Its output is a set of actions different from the actions that would have been taken had the process not been conducted: a capability not released on the date it was ready, a research direction not pursued, a product feature removed, a contract declined, a deployment restricted to a domain narrower than the market would bear. Where a process has run for years and no decision has come out differently from the way it would have come out anyway, the process has no substance, whatever else it has produced. The same test, stated as a burden of evidence: an apparatus which has never once prevented a deployment that its sponsors wanted has provided no evidence of being anything other than ornamental (*The Instruments of Choice*, Chapter Three, Section VIII).

This gives the first and most useful test, and everything else in the essay depends on it. The *substance of precaution* is a counterfactual: the decision that would have been made, and was not.

Two things follow immediately.

The first is that precaution is expensive by construction. A process that changes decisions changes them against the direction in which the ordinary incentives of the organization were pushing, because a decision that the incentives already favored does not need a process to reach it. The cost of genuine precaution is therefore not principally the cost of running the process. It is the value of what the process gives up.

The second is that the substance is not directly visible. A counterfactual is not an artifact. It leaves no object on a shelf, no document in a repository, no line in an annual report that says: here is the revenue we did not earn, and here is the reason. What the organization can show is the process that produced the counterfactual, and the process is a different thing from its product.

> **On the level of the analysis**
>
> The argument of this essay concerns organizations and their environment, not the honesty of the people who work in them. It does not require that any executive, engineer, or reviewer intend to deceive. It requires only that producing an artifact be cheaper than producing what the artifact was meant to indicate, and that the parties who select among organizations observe artifacts rather than substance. Both conditions are ordinary. Where the argument is strongest, it is strongest for the same reason the argument of Essay 1 is strongest: because it does not depend on anyone behaving badly.

II. The Artifacts and the Substance

Consider what a genuinely precautionary organization produces in the course of being one.

It convenes a body with a remit to assess consequences. It writes down the principles by which it will decide. It maintains a register of the risks it has identified and the mitigations it has adopted. It commissions adversarial testing by parties instructed to make the system fail. It releases capability in stages rather than all at once. It invites external review and publishes something of what the review found. Each of these is a real instrument, and each of them is doing real work in the organization that has the substance.

Now consider what these instruments cost to an organization that produces them and changes no decision.

The committee costs the salaries of its members and the hours of those who attend it. The principles cost a drafting exercise. The register costs a spreadsheet and a maintainer. The adversarial testing costs an engagement fee and the discretion to decide which findings are in scope. The staged release costs a schedule that can be set at whatever pace the commercial plan already wanted. The external review costs a fee and the selection of a reviewer. The total is a rounding error against the value of a single deployment decision reversed. John Meyer and Brian Rowan described how organizations adopt formal structures for the legitimacy the structures confer as well as for the work they do, so that a structure can be adopted while the work it was meant to carry is left undone (Meyer and Rowan, 1977).

This is the *observational equivalence of precaution*, the property at the center of the essay: that two organizations, one of which conducts precaution and one of which does not, emit the same set of visible signs. The equivalence is exact in the respect that decides the question: the artifacts are not similar, they are the same artifacts, because the artifact was never the substance and there is nothing in the artifact that carries a trace of the counterfactual it was meant to indicate.

A reader may ask whether the artifacts of the imitating organization would be of lower quality: thinner principles, a shallower register, a tamer red team (the adversarial testers). Sometimes this is so. It is not reliably so, and the direction of the unreliability is unfavorable. An organization that has resolved to produce artifacts and no substance has its entire precaution budget available for the production of artifacts, while the organization with the substance is spending most of its budget on the substance. Where the two are compared on the artifacts, the imitation is not merely equal. It is frequently better.

III. The Filter

The displacement of the dearer form by the cheaper is best understood as a selection process rather than as a choice.

Consider a population of organizations pursuing the same capability under the same competitive conditions. Each faces a choice between two allocations of the same budget. One allocation produces the artifacts and the substance. The other produces the artifacts alone and returns the remainder to the pursuit of the capability. The parties that observe these organizations (investors, customers, regulators, the press) observe artifacts. On the observable, the two allocations score alike. On everything else, the second scores higher.

An organization that adopts the second allocation therefore moves faster, ships sooner, and raises capital on better terms than one that adopts the first, while presenting an equivalent account of its own carefulness. Over enough cycles, the population is composed of organizations of the second kind, and no decision by any person was required to bring this about. Call this the *precaution filter*: the selection of imitation over substance by an environment that can observe only the artifacts, operating on organizations rather than through the intentions of the people in them. The nearest existing treatment of such a filter is George Akerlof's account of adverse selection, in which a market whose buyers cannot observe quality comes to consist of the goods of lower quality (Akerlof, 1970).

A second mechanism runs beside the filter: an observer who prefers not to discriminate, because the artifact is what the observer's own account requires. A regulator, investor, or customer of that kind selects for artifacts directly.

The filter has the structure of the multi-polar trap described at Essay 1, Section IV. Each organization, reasoning correctly from its own position, concludes that it cannot afford the first allocation while its competitors take the second, because the difference is not a margin but the rate at

which it can pursue the capability at all. Each is right. The aggregate of their correct reasoning is a field in which precaution is universally professed and nowhere practiced, and from which no participant can withdraw unilaterally without ceasing to be a participant.

This is the first of the three disabling factors named at Essay 1, Section VII, stated at length. The factor was there given as a cost asymmetry. It is better understood as a filter, because a cost asymmetry is something a party could choose to bear, and a filter is something that removes the parties who bear it.

IV. Sincere Simulation

The greater part of simulated precaution is sincere, and the argument is stronger for it.

An organization that has convened the committee, adopted the principles, maintained the register, and commissioned the review has done everything that its members can see being done. The members attend the meetings. The discussions are substantive. Concerns are raised and recorded, and mitigations are adopted in response to them. From inside, the experience is of an organization taking the matter seriously, and the experience is not false: the people are taking it seriously.

What the members experience lies between artifact and substance. Political and organizational actors, under ordinary incentive structures, optimize first for what looks good: an announcement, a task force, a compliance report. They optimize second for what feels good: a sense of relationship and reassurance built with stakeholders, a round of meetings in which everyone felt heard. This level is deeper than appearance. It involves actual social process and actual attunement between the people involved. But it is still not the same as substance. Only third, and often not at all, do they optimize for what is good: the actual, verifiable state of the system in the world. Most institutional responses rest, comfortably and sometimes sincerely, on the first two. They look good. They often feel

good. Whether they are good remains, in most cases, structurally unverified (*The Reach of Trust*, Chapter Seventeen, Section VIII). Sincere simulation is an organization resting on the second of these levels and taking it for the third.

What is absent is the counterfactual, and the counterfactual is the one thing that no participant is positioned to observe. The engineer sees the review that was conducted. The reviewer sees the concerns that were addressed. The executive sees a process that has never yet required the company to do anything ruinous, and reads this as evidence that the process is well calibrated rather than as evidence that it is inert. Nobody in the organization holds the comparison between what happened and what would have happened, because that comparison is not an object that any role in the organization is assigned to hold.

Call this *sincere simulation*: the production of precautionary artifacts without the substance, by parties who believe themselves precautionary and who are not lying when they say so. Diane Vaughan's study of the decision to launch the Challenger records a related process: the normalization of deviance, the gradual acceptance of departures from a standard as normal, within an organization that continued to regard itself as careful (Vaughan, 1996).

Sincere simulation is more robust than the cynical kind and harder to address. A cynical simulator can be exposed, because there is something to expose: a knowledge, held by someone, of what is really being done. A sincere simulator has no such interior. There is no document that says the process is theater, no meeting at which it was decided, and no participant who could be brought to admit it, because none of them believes it and their belief is honestly held.

A question: does this prove too much?

A reader may ask whether this reasoning, if correct, would show that no precautionary governance anywhere has ever had any substance, which is plainly false: there are industries in which precaution demonstrably operates, and aviation and pharmaceuticals are the usual examples.

The question is well taken, and the answer is that the filter is a pressure and not a law. Its strength scales with the ratio between what precaution costs and what the observing parties can verify. In aviation the substance is verified by accidents, which are public, attributable, and expensive, and which arrive on a timescale shorter than the career of a manufacturer. Where the harm is prompt, legible, and traceable to its author, the artifact and the substance come apart in public, and the filter runs the other way. The argument of this essay applies with force exactly where those conditions fail: where harm is delayed, diffuse, and hard to attribute, which is the condition described at Essay 1, Section V. Charles Perrow wrote on the relation between coupling, complexity, and accident in high-risk industries of this kind (Perrow, 1984).

V. What an Observer Could Look At

There are observables that discriminate, and identifying them is the constructive part of the analysis.

Each of the following is a feature of an organization's conduct whose presence would be evidence of substance and whose absence would be evidence of its lack. Each is chosen on the same principle: that it bears on the counterfactual rather than on the process that was supposed to produce it. Call any such feature a *discriminating observable*.

The first is the record of reversals. An organization with the substance has a history of decisions that went the other way: releases delayed past the point of commercial pain, capabilities built and not shipped, revenue declined. The record is specific, dated, and costly, and the cost is the part that carries the information.

The second is the location of the veto. An organization with the substance has a function that can stop a deployment, that is not subordinate to the function that owns the deployment's revenue, and that has stopped one. The existence of the authority is weak evidence. Its exercise, against opposition, is strong evidence.

The third is the shape of the budget. Precaution consumes resource in places where nothing ships and no revenue is booked. An organization with the substance has a material fraction of its capacity committed to work whose success is invisible, and has sustained that commitment through a period in which the capacity was wanted elsewhere.

The fourth is the character of the questions asked internally. An organization that has confronted the hard questions in its path has been shaped by the confrontation, and the shape shows in what it asks itself without prompting. This is the diagnostic developed at Essay 5, Section III: that the sharper test of a party is not the quality of its answers but whether the necessary question arrives as news.

The fifth is the order in time. An organization with the substance built the assessment capacity before or alongside the capability it assesses. One without it built the capability first and the assessment afterward, under external pressure, and the dates are on the record.

Each of these five is a genuine discriminator. An observer with reliable access to any one of them could tell the two kinds of organization apart. The five are representative; Section VI shows why any further observable that the organization itself produces meets the same fate.

VI. Why Each Observable Is Capturable

The capturability of these observables follows from a single property that all five share, and once the property is stated the individual cases are straightforward.

A discriminating observable is chosen because it is expensive to produce without the substance. Publishing it tells every party exactly which expense must be incurred in order to be scored as having the substance. This is *specification by publication*: the observable becomes a specification (a description, in operational detail, of what an imitation must contain in order to pass). And a specification for an imitation is a cheaper thing to satisfy than the substance it was chosen to indicate, for the same reason the artifacts were cheaper in the first place. This holds for every observable that the observed party produces, and Section VIII turns on exactly that condition.

This is Goodhart's observation, that a measure adopted as a target ceases to measure (Goodhart, 1976; the wording now usually quoted is Marilyn Strathern's, Strathern, 1997), and the present case adds a term to it. The ordinary Goodhart effect is degradation: the measure becomes noisy as parties optimize against it. Here the effect is selection: the parties who optimize against the measure outcompete the parties who satisfy it honestly, because optimizing against it is cheaper, and the population changes composition. Degradation blunts an instrument. Selection replaces the field the instrument was measuring.

The five cases follow the pattern exactly.

A record of reversals can be produced by reversing decisions that were going to be reversed anyway, and announcing the reversal in the language of precaution. The commercial cost is nil and the artifact is a dated, specific, apparently costly reversal.

A veto can be exercised on a deployment that the deploying function had already decided against, by arrangement, so that the record shows a stop taken against opposition where no opposition existed.

A budget can be shaped by classification. Work that would have been done for reliability, performance, or liability is reported under the precaution heading, and the fraction rises without any capacity moving.

Questions can be prepared for once the questions are known, and this is the case in which the capture is least detectable, for the reason given at Essay 5, Section VIII (the undetectable defeat): the party who issued the question does not learn that the answer was prepared, and therefore does not learn that a new question is needed.

And the order in time is a matter of what is dated and by whom. An assessment function can be constituted early and left idle, which costs a title and a salary, and the record will show that it preceded the capability.

The general result is not that observation is useless. It is that the discriminating power of any published observable decays at a rate set by the value of passing it, and that in this domain the value of passing it is very large.

A record that the observed party produces about itself is, in this sense, a resource that can be accumulated and then spent. The marketplace version shows the arithmetic. A new vendor appears on a marketplace platform with no history at all. They begin selling a large number of small, cheap items, delivering quickly, and accumulating positive feedback from buyers who are, individually, risking very little. After a few thousand such transactions, the vendor has a stellar reputation score, close to perfect, built entirely out of small stakes. At that point the vendor offers a single high-value item, priced well above the cumulative value of everything sold before it. Buyers see the reputation score, not the size of any individual transaction behind it, and they extend trust accordingly. The sale goes through, the money arrives, and the vendor disappears, having stolen an amount of value greater than everything the reputation score was actually built on. Any feedback mechanism built only from re-

cords of what has already happened provides real but bounded predictability, and that bound can be probed and exploited by any actor willing to invest in a long enough setup (*The Reach of Trust*, Chapter Twenty, Section III). A record of reversals is a record of this kind, and the capture costs less than what it buys, which is the condition Section VIII names.

VII. The Cost of Verification

Verification of substance is possible, and the conditions under which it is possible can be stated.

The substance is a counterfactual, so verifying it requires access to the decision record: to what was proposed, what concerns were raised against it, by whom, what was decided, and what it cost. That record exists. It is held by the organization, and it is the very material that an organization has the strongest reasons to control, for commercial confidentiality, for legal exposure, and from the ordinary reluctance to hand an adversary the list of one's own near misses.

The verifier is therefore in the position described at Essay 3, Section III: seeking to establish something before an audience with a material interest in its not being established, and dependent on that audience for the evidence. The inferential distance is set not by the difficulty of the question but by the disposition of the party holding the answer, and the effort required grows with the sums at stake rather than with the complexity of the assessment.

This accounts for a pattern that is otherwise puzzling. External audit regimes in this domain tend to converge on process compliance (did the committee meet, was the register maintained, was the review conducted) and away from outcome verification. The convergence is not a failure of imagination by auditors. It is the only assessment that can be conducted with the access that is actually granted, and it is therefore the assessment that an audit market will supply. Michael Power described the same convergence in audit generally, as a ritual of verification (Power, 1997). The

observable that can be verified cheaply is the one that discriminates least, and this is not a coincidence: it is cheap to verify because it does not require the material that the verified party is unwilling to release.

VIII. What Capture Would Have to Cost

The result that survives is a positive one, and it changes the design question.

An observable resists imitation in proportion to what its imitation costs. The five observables of Section V fail not because they are the wrong things to look at but because each of them can be produced, by an organization that controls its own account of itself, for far less than the substance costs. The property that was wanted was never uncapturability, which nothing has. It was that capture should cost what the substance costs. Call this condition *capture-cost parity*. Michael Spence's account of signaling states the same condition for markets in which quality cannot be observed directly: a signal separates those who have the quality from those who do not only where sending it costs the second kind more than it costs the first (Spence, 1973). The two conditions agree: the organization with the substance has already borne its cost and pays nothing more to show it, so wherever imitation costs what the substance costs, the signal costs the imitator more than it costs the organization with the substance.

Parity has two sides, and a design needs both. Against the substance: where imitating costs what the substance costs, imitation saves nothing, and the filter of Section III has nothing to select for. Against the prize: where taking the arrangement costs more than what taking it buys, no outside party gains by taking it. The first side removes the reason to imitate; the second removes the reason to acquire.

Parity is a property of arrangements, not of measurements. It is obtained where the party who produces the observable is not the party whose interest the observable bears on, and where the producing party would have to give up something real in order to falsify it.

Are the five observables of Section V the whole set? They are representatives, and the reason of Section VI covers every observable that the organization itself produces. There is a second kind, which the organization does not produce, and the condition of parity admits it already. An insurer's price for the organization's risk is set by a party that pays if the price is wrong; a record kept by an employee who can report outside the organization without loss is kept by a party whose interest the record does not serve. An observable of this kind meets the condition to the degree that its producer would lose something real by falsifying it.

The sidebar to Section IV supplies a further lever. Where harm is prompt, legible, and traceable to its author, the filter runs the other way, as it does in aviation. An arrangement that makes the harms of a technology quicker to register, or easier to trace to their author, therefore moves the field toward the condition in which substance is selected: liability that reaches the developer, a register of incidents kept by parties other than the developer, and the tracing of each output to the system that produced it are arrangements of this kind. The lever reaches as far as harm can be made prompt and attributable; where harm is delayed and diffuse by its nature, parity remains the design aim.

Consider the second observable, the location of the veto, under two arrangements. Where the vetoing function sits inside the organization, reports through it, and is paid by it, the exercise of the veto costs the organization whatever it chooses to let it cost, which may be nothing. Where the vetoing function is external, funded independently of the outcome, composed of parties who have no stake of their own in the decision: they lose nothing of their own by halting a deployment and gain nothing of their own by permitting one; and empowered such that the halt takes effect without the organization's assent, a staged veto is no longer available. Arranging a favorable outcome now requires acquiring the external body, and acquiring it costs what it costs to acquire a party with no stake in the answer: a great deal, and a cost that scales badly for the party acquiring, in the sense developed at Essay 4, Section VII.

Can the external body itself be bought? It can, since nothing is beyond acquisition by a party willing to pay enough. The arrangement does its work when buying the body costs what the substance costs, and how a body is funded and composed so that it does is the question Essay 4, Section VII, and the Afterword, Section IV, take up.

This is the same conclusion the volume reaches from other directions, arrived at here from the direction of measurement. The question is not which sign of carefulness can be trusted. It is whether the arrangement that produces the sign is one in which the parties who decide bear the consequences of deciding wrongly. That is the fourth property of the specification given at Essay 1, Section VIII, and the analysis of this essay supplies a reason for its inclusion that the specification did not itself provide: without it, no observable of any of the other three properties retains its meaning, because the party observed is the party supplying the observation.

IX. Summary

The central claim, that the substance of precaution is a counterfactual and that counterfactuals leave no artifact, holds by the meaning of its terms and does most of the work in the essay. The filter argument of Section III follows from it together with the cost asymmetry, and requires no premise about intention. The account of sincere simulation in Section IV is an empirical claim about how organizations experience themselves, supported by the structural argument that no role is assigned to hold the relevant comparison. Section IV also gives the level that sincere simulation feels like from inside: between what looks good and what is good lies what feels good, real meetings and real attunement that are still not substance.

The argument applies where harm is delayed, diffuse, and hard to attribute to its author, and where harm is prompt, legible, and traceable to its author, the filter runs the other way, as the sidebar to Section IV records.

Section VI gives a worked case of a record spent: a reputation built from small stakes and used once for a large one, which is capture costing less than it buys. The analysis of Section VII gives reasons to expect that any verification regime operating on access granted voluntarily will converge on process compliance, and the essay supplies, in place of such a regime, the condition that any regime would have to meet.

Capture-cost parity is stated as a condition and not as a design. The question it puts to any observable is not whether the observable can be imitated, since every observable the observed party produces can be, but what imitation costs against what passing buys.

X. Conclusion

The argument of this essay can be put in a single line.

Care shows in what it gives up, and what is given up leaves no object behind.

Everything else follows. The artifacts of precaution are producible without precaution, so the artifacts do not discriminate. The observables that would discriminate are produced by the party they discriminate against, so publication converts each of them into a specification for its own imitation. And the selection pressure runs continuously, so the field does not merely contain simulators; over time it comes to consist of them, without any participant having chosen this and without any participant being able to see it from inside.

The recommendation that follows is not vigilance. Vigilance is an observational remedy and the problem is not observational. The recommendation is to stop asking which sign of carefulness can be trusted, and to ask instead what would have to be true of an arrangement for its signs

to cost what they claim. The answer will always have the same shape: that the party who produces the sign is not the party whose interest it bears on, and that falsifying it would require giving up something the falsifying party actually values.

This is a demanding condition and it is not met anywhere at the scale that matters. It is nevertheless a condition on arrangements, which are things that can be built, rather than a condition on the honesty of participants, which is not a thing that can be built at all. That is the whole of the difference between a problem that is merely serious and a problem that is hopeless, and it is the reason the analysis is worth carrying this far.

XI. What This Essay Adds to the Volume

This essay gives in full one condition under which the capacity for collective judgment fails: where the substance of care is a decision given up, and the parties who must judge can see only its artifacts, the field comes to consist of imitations, and no participant can see this from inside.

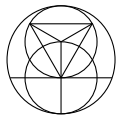
It is the first of the three disabling factors of Essay 1, Section VII, and the essay shows why it stands in the way of stating the volume's case at all. Any judgment, collective or otherwise, is only as good as what it observes, and what can be observed about precaution is produced by the party being observed.

The essay also gives the volume a reason that the governance specification did not itself supply. Without the fourth property, no observable of the other three retains its meaning, because the party observed is the party supplying the observation. Capture-cost parity, the condition on which the essay closes, is the form that the volume's recurring question takes for every arrangement it examines: not whether the arrangement can be captured, but what capture costs against what it buys. The two companion essays that follow carry the result into the self-regulation of AI companies and into the study of existential risk.

KEY CONCEPTS AND FINDINGS IN THIS ESSAY

- **Simulated precaution**: the production of the visible artifacts of precautionary governance without its substance; selected over genuine precaution wherever the two are observationally equivalent and differ greatly in cost.
- **The substance of precaution**: the decisions a precautionary process changes, and the value given up in changing them; a counterfactual, and therefore not an artifact.
- **Observational equivalence of precaution**: the property that an organization with the substance and one without it emit the same visible signs, since no artifact carries a trace of the counterfactual it was meant to indicate.
- **The precaution filter**: the selection of imitation over substance by an environment that can observe only artifacts; operating on the composition of a population of organizations rather than through the intentions of the people in them.
- **Sincere simulation**: the production of precautionary artifacts without the substance by parties who believe themselves precautionary, and who are not lying when they say so, no role in the organization being assigned to hold the relevant comparison.
- **Discriminating observable**: a feature of an organization's conduct that bears on the counterfactual rather than on the process meant to produce it, and whose presence would therefore be evidence of substance.
- **Specification by publication**: the conversion of a discriminating observable into a recipe for its own imitation, by the act of naming what an observer will score.

- **Capture-cost parity**: the condition on an arrangement that falsifying an observable cost what the substance it indicates costs (so that imitation saves nothing), and that taking the arrangement cost more than what taking it buys (so that no outside party gains by taking it); a property of arrangements rather than of measurements, and the replacement for the unobtainable property of uncapturability.
- **Looks good, feels good, is good**: the three levels an institution can optimize for; most responses rest on the first two, and whether they are good remains structurally unverified.
- **A record spent**: a record the observed party produces about itself is a resource that can be accumulated and then spent.



Companion Essay

Three Problems with AI Self-Regulation

On the belief that AI companies can govern themselves from within

ABSTRACT

The belief that AI development can be kept safe by the internal self-regulation of the companies doing it runs into three problems of business governance: the principal-agent problem, the race to the bottom, and the rules for rulers. In the first, the safety community supports founders who promise safety work, believes them for years, and, once their dishonesty is revealed and safety people are sidelined, turns to new founders, so that the same cycle begins again. In the second, collaborators at different companies cut corners for fear that competitors will get there first and make things worse; the more they treat this as a finite game played against untrusted players, the more they lose integrity with their own values. In the third, leaders who seek to gain and keep power over AI development install and handsomely reward the key people who acquire the resources that keep them in power, even at the cost of extracting from outside citizens who have no say in what the company does. Each problem is a pattern in how the people involved act, and each stands against the belief that safety can be left to the companies themselves. Each is also the structure of Essay 1 appearing inside the industry: the first is the fourth property of the governance specification (that those who decide bear what their decisions cause) left unmet, the second is the multi-polar trap and the precaution filter, and the third is the knowledge-authority gradient held in place by the people the leader rewards. The essay closes on the

> question of whether a company can bind itself, and finds that a binding holds only where the party able to undo it is not the party it binds, which is regulation from outside.

I. The Belief in Internal Self-Regulation

Many people believe that the companies developing AI can be trusted to regulate themselves: that safety can be kept by the commitments, policies and good intentions of the people inside those companies. This is a belief about business governance. There are three problems with it. Each problem is a pattern in how the people involved act.

The belief has real attractions, and they should be stated before the problems are. The companies hold the expertise: the people who best understand what is being built work inside them. Outside regulation is slow, and its officials often understand the technology less well than the engineers they would regulate. And many of the people in these companies are sincerely concerned about safety, and say so in public, and some of them chose their employer for that reason.

None of the three problems below depends on anyone being insincere at the start. Each is a pattern that ordinary, well-intentioned people reproduce when they act on the incentives of their position, which is the sense of "structural" set out at Essay 1, Section I. The expertise and the good intentions are real. What the three problems show is that neither of them decides what the company does when safety and the company's position pull in different directions.

II. The Principal-Agent Problem

The safety community supports new founders who convince it that they will do good work for the cause of safety. For years, the people of the safety community believe the leaders. But once cases of dishonesty get revealed, and leaders sideline the safety people who are no longer needed, the community distrusts the leaders and allies with new founders.

The community is the principal, and the founders are its agents: the community gives its support so that the founders will act on its behalf. The founders act instead on their own behalf, and the community finds this out only after years. Its response is to place the same kind of trust in new founders, and so the cycle starts over.

The general form is old and well studied. A principal is a party who hands a task to another, and an agent is the party who carries it out. The agent has interests of its own, and knows more about its own conduct than the principal can see, so the principal pays to watch the agent and still loses some of the value it delegated for to the agent's own interests (Jensen and Meckling, 1976). In the present case the task delegated is safety, and what the community gives is money, talent, reputation, and its public endorsement.

The community cannot tell, at the time it gives its support, which founders will keep their word. What it can see at that time are commitments, charters, published safety plans, and the hiring of safety researchers. These are the artifacts of precaution that Essay 2 describes: they can be produced with or without the substance, and the substance (a decision that goes against revenue) is shown only later, when a costly decision is actually faced. By then the founder holds the capital, the people, and the product, and the support has already been given.

This also explains why the safety people come to be "no longer needed". At the start, a reputation for safety is an asset: it attracts researchers who would not otherwise come, and funders who would not

otherwise give. Once the capital and the talent are secured, that asset has done its work, and the safety function becomes a cost borne by a company that competes against others who do not bear it. The pressure that Essay 2, Section III, describes among companies then acts within the company: the part that bears a cost the observers cannot see is the part that is cut.

The cycle repeats because the community changes its agents and does not change the arrangement. The new founders stand where the old ones stood, with the same incentives, the same ability to leave with what they have been given, and the same distance between the promise and the decision that tests it. The Afterword, Section V, states what the fourth property of the governance specification asks: an arrangement that does not need a trustworthy agent, because the party who decides is inside the set of parties affected, cannot leave it, and bears the cost of a wrong decision on a timescale comparable to the one on which it decides. A founder with capital can leave, and the ability to leave is what capital is. Choosing better founders is choosing agents, and the arrangement stays the one that produced the last outcome.

III. The Race to the Bottom

Collaborators at different companies cut corners, believing that if they do not, their competitors might get there first and make things even worse.

The more the people participating treat this as a finite game, in which they are acting independently of other, untrusted, individual players, the more they lose integrity with their values.

This is the multi-polar trap of Essay 1, Section IV, seen from inside the field. Each participant reasons correctly from their own position: a company that holds back does not stop the capability from being built, and only removes itself from among those who will have a say in how it is built. Each is right, and the sum of their correct reasoning is a race that none of them wanted. The same dynamic has been modeled directly for

teams racing to build artificial intelligence, in which each team reduces its safety precautions in order to be first (Armstrong, Bostrom, and Shulman, 2016).

James Carse distinguished a finite game, played to be won and therefore to be ended, from an infinite game, played so that the play can continue (Carse, 1986). In a finite game every value the players hold is weighed against the one value of winning, and is kept only so long as keeping it does not lose the game. That is how integrity with one's values is lost without anyone deciding to give it up. Each corner cut is justified by the value it gives up: the safety commitment is set aside so as to stay in the race and keep the safety-minded people in a position to steer. Each step is taken in the name of the commitment, and each step spends some of it, until what remains is the name.

Self-regulation offers no way out of this, because the way out is not in the power of any one participant. A company that keeps its commitments alone bears a cost its competitors do not bear, and is outpaced or removed from the field (Essay 1, Section IV, on the word "trap"). Only a change in the structure, which binds all the participants together, changes the outcome, and a structure that binds all of them is something that none of them is in a position to impose.

IV. Rules for Rulers

Leaders seek to gain and maintain their positions of power over AI development. In order to do so, they need to install key people who can acquire the resources needed for the leaders to stay in power, and they need to reward those key people handsomely, even if it means extracting from all the other, outside citizens who have no say about what the company does.

The pattern is general. Political scientists have argued that it holds for leaders of every kind of organization, states and companies alike: a leader stays in power by keeping the loyalty of a small set of key supporters, the

keys to power, who control the resources on which the leader depends, and the leader keeps that loyalty by paying the keys from what can be taken from those outside the circle (Bueno de Mesquita and others, 2003; Bueno de Mesquita and Smith, 2011). The phrases "rules for rulers" and "keys to power" are those of CGP Grey's video "Rules for Rulers", which made that work widely known (Grey, 2016). For a company developing AI, the keys are its investors, its senior executives, the researchers it cannot do without, and the parties who supply its computing capacity. The outside citizens are everyone who bears the costs of the company's products and has no part in its decisions: the commons of Essay 1, Section V.

This bears directly on safety. A safety decision that costs revenue is paid for by the keys, since what they are paid comes from that revenue, and a leader who makes such decisions often weakens the loyalty on which their position rests. The person who sees a risk most clearly is usually an engineer at the face of the work, and that engineer is seldom a key. This is the knowledge-authority gradient of Essay 1, Section VII (authority increases as understanding decreases), and the rules for rulers show what holds it in place: authority goes to whoever controls the resources, and the resources are controlled by those whose interest is in deployment.

V. One Structure, Seen in Three Places

The three problems are three places where the structure set out in Essay 1 shows inside the industry that is asked to regulate itself.

The principal-agent problem is the fourth property of the governance specification left unmet: those who decide are not aligned with the whole of what their decisions affect, and can leave before the costs arrive. The race to the bottom is the multi-polar trap, and, among the companies, it works as the precaution filter of Essay 2: the organizations that bear the

cost of real precaution are outpaced by those that bear only the cost of its artifacts. The rules for rulers are the third disabling factor of Essay 1, Section VII, held in place by the people the leader must reward.

Removing any one of the three leaves the other two standing, and each of them still stands against self-regulation, much as removing any two of Essay 1's disabling factors leaves the third, which still produces the gap between what is recommended and what occurs. A company whose founders keep their word is still in the race. A company out of the race still answers to its keys. A company whose keys cared about safety would still be an agent its principals cannot watch.

Underneath all three is one fact, which Essay 2, Section VIII, states: an observation means nothing when the party observed is the party supplying it. Self-regulation is the case in which the party observed supplies every observation, the commitments, the reports, and the account of what was decided, and so it is the case in which capture costs least.

VI. Whether a Company Can Bind Itself

A reader will ask whether a company can bind itself in advance: by a charter that names safety as its purpose, a board with a safety mandate, a corporate form that answers to the public benefit, or a published policy that commits it to halt at stated thresholds.

A binding holds when the party able to undo it is not the party it binds, and when undoing it would cost that party something real. A charter that the company's owners can amend, a board that its keys can replace, and a policy that the company itself interprets each binds the company only so long as the company chooses to be bound. Each is then a promise, and it costs what breaking a promise costs, which in a race is less than what keeping it costs. These are the artifacts of Essay 2, Section II, and they are produced as readily without the substance as with it.

A binding that does its work is one held by a party outside the company: a body funded independently of the outcome, composed of parties who have no stake of their own in the decision: they lose nothing of their own by halting a deployment and gain nothing of their own by permitting one; and empowered so that a halt takes effect without the company's assent. Essay 2, Section VIII, and the Afterword, Section IV, set out the conditions on such a body. A binding of that kind is regulation from outside. So the question whether a company can bind itself has the answer that the three problems already give: the binding that holds is the one that is not self-regulation.

VII. Summary

This essay states three problems with the belief that AI companies can keep AI development safe by regulating themselves from within. The principal-agent problem shows the safety community giving its trust to founders who later set safety aside, and then giving the same trust to new founders. The race to the bottom shows people at competing companies cutting corners for fear of one another, losing integrity with their values the more they treat the contest as a finite game among untrusted players. The rules for rulers show leaders keeping power by rewarding the people who bring them resources, at the expense of citizens who have no say. Each of the three is a problem of business governance, and each stands against the belief in internal self-regulation.

Each of the three is also a place where a structure of Essay 1 shows inside the industry: the unmet fourth property of the governance specification, the multi-polar trap working as the precaution filter, and the knowledge-authority gradient held in place by the keys to power. Removing any one of them leaves the others standing. A company can bind itself only by a binding that a party outside it holds, and such a binding is regulation from outside.

VIII. What This Essay Adds to the Volume

This essay takes up the answer most often given in practice to the volume's question of what would be required for the capacity for collective judgment to succeed at the scale at which the decisions that matter are now made: the answer that the companies building the technology can supply that judgment themselves.

The three problems show why they cannot. A company is a small group with interests of its own, and the judgment that the volume asks for must be formed among all the parties a decision affects. Self-regulation puts the decision with the parties who gain from deployment, leaves the costs with the commons, and makes the observed party the only source of observation. It therefore fails the fourth property of the specification by its construction, and fails the third (that capture be ineffective), since the party observed supplies every observation and so capture costs least.

The essay carries Essay 2 into a further case. Essay 2 shows that the artifacts of precaution are cheaper than its substance and are what an observer will find. This essay shows that the same holds for the governance of the companies themselves, and that the remedy has the same shape: an arrangement held by parties who do not gain from the outcome, which is the capacity the volume says is missing.

KEY CONCEPTS AND FINDINGS IN THIS ESSAY

- **Internal self-regulation**: the belief that AI companies can keep their own work safe without outside control.
- **Principal-agent problem**: the safety community trusts founders, is let down, and trusts new founders; it changes its agents and not the arrangement, so the cycle repeats.
- **Race to the bottom**: fear of competitors leads each company to cut corners; the multi-polar trap of Essay 1 seen from inside the field.

- **Finite game**: treating the contest as one to be won against untrusted players erodes integrity with one's values, since every value is weighed against winning.
- **Rules for rulers**: leaders keep power by rewarding the key people who bring them resources, at the cost of outside citizens.
- **Keys to power**: the small set of supporters who control the resources a leader depends on; in an AI company, its investors, senior executives, indispensable researchers and suppliers of computing capacity.
- **A binding that holds**: one held by a party outside the company, who is not the party it binds and who would lose something real by undoing it; such a binding is regulation from outside.



Companion Essay

Anti-Democratic Risk

The techno-utopian agenda,
and why the study of existential risk must be democratized

ABSTRACT

The techno-utopian approach to existential risk is better understood as an agenda: a set of practices and beliefs that increase social inequality, privatize the wealth of the commons, and so themselves increase the real risk of harm to everyone. This essay establishes that by a commentary on the paper "Democratising Risk" (Cremer and Kemp, 2021), which names a "techno-utopian approach" to existential risk built on transhumanism, total utilitarianism and strong longtermism (the view that the best choice is the one with the best expected effect on the far future). Read closely, the founding beliefs of the agenda are a sales pitch, built to trigger fear, offer hope, and justify any extreme action, to the benefit of a few. Its claims that AGI development is inevitable and that AGI alignment is possible in principle are both wrong. Real value, and real choice, exist only in the present, and our moral obligation is to safeguard choice for all who have or could have it. From this it follows that choosing which risks to take must be a democratic endeavor, carried out by the commons, for the commons, and that the techno-utopian agenda cannot remain influential on the study of existential risk.

I. The Paper and the Agenda

The study of existential risk is shaped by an agenda that increases inequality, and so increases the real risk of harm to everyone; in that field, the very field that studies the risks collective judgment would have to weigh, this essay finds one of the conditions under which the capacity for collective judgment fails. It is written as a commentary, in a sharper register than the essays around it, because it answers an agenda and not a single argument. What the volume takes from this essay is that the study of existential risk is itself shaped by the forces of Essay 1, and that which risks to take must be chosen by the parties those risks affect; the volume's claim about the missing capacity does not rest on what this essay says of anyone's motives.

This essay names an agenda, and it also names the people who use it. Its argument does not rest on what those people are like. The agenda's beliefs are shaped for the use made of them "by the dynamics of social evolution, if not by any one person's specific intent" (Section II); where a person does act in bad faith, that person exploits a structure that would produce the same outcome without them. The word *nar-social*, as the essay uses it, names a mode of engagement before it names a kind of person: a mode in which the social surface of an exchange (who is winning, who is allied, who is diminished) has become the actual content, while the ostensible subject matter serves only as the material out of which those social effects are produced. It is a description of a mode, and everyone enters that mode sometimes (*Design and Games*, Chapter 1, Section III); a nar-social person is one who works in that mode as a means of social power (*Collaboration, Community, and Communication*). The outcomes the essay describes need no ill intent. They are what happens when the terms of an exchange are left to settle themselves, which they do, in favor of whoever is less troubled by the settling (*Design and Games*, Chapter 4, Section X). Where the essay names a kind of person (the nar-social, the psy-

chopath, the engineer deceived into belief, the maker on the autistic spectrum), it names a position in a structure, and the conduct that the structure rewards and recruits into that position; it is not a verdict on the character of any person. These observations are made because they bear on the wellbeing of everyone and of the world as a whole, and they are made for that reason alone. The failure described here, like every failure in this volume, requires no one to be foolish or dishonest.

The paper "Democratising Risk", by Carla Zoe Cremer and Luke Kemp (2021), describes what it calls the techno-utopian approach (TUA) to the study of existential risk. The paper also reflects the belief, widely held in that field, that AGI (artificial general intelligence) development is inevitable. That belief is simply, and dangerously, wrong.

The paper describes the TUA as a cluster of three orientations:

> transhumanism; total utilitarianism; and strong longtermism.

The paper finds that the TUA unnecessarily combines the study of longtermist ethics with the study of extinction, relies on a non-representative moral worldview, uses ambiguous and inadequate definitions, fails to incorporate insights from risk assessment in relevant fields, chooses arbitrary categorizations of risk, and advocates dangerous mitigation strategies.

It is better to define "TUA" in two ways.

First, the TUA is the techno-utopian *agenda*. The practices of commercialism, transactionalism, institutional design, "art" and "cultural styles", and the notions of political control, population control, and the utilitarian philosophies of modernism, are all implicitly supported by the agenda, or are themselves in support of it. All of these aspects of the agenda are part of the practice of increasing social inequality (that is, of power-law distributions of social power) and of the privatization of the wealth of the pub-

lic commons. The agenda favors the rich becoming richer. That is a dynamic which must at the least be slowed, and should be stopped, if the world is to survive.

Second, the TUA is the techno-utopian *approach*: the use, assumption, or implementation of any of the aspects named under the agenda.

II. How the Agenda Increases Real Risk

The TUA in itself has the additional aspect of increasing the real risks of harm. It does this by replacing organic life, in all cases and in increasing degrees, with artificial machines and with (inherently toxic) industrial processes, until, if the agenda runs its course, all life is eventually lost.

The moral and empirical assumptions of the TUA are specifically enabling of securitization (moving an issue out of ordinary politics into national security) and misuse. They are not just "particularly vulnerable" to misuse. They are implicitly "designed for such purpose" by the dynamics of social evolution, if not by any one person's specific intent.

The TUA is not representative of wider human preferences. That is, the TUA favors the rich, the corporations, and so on, and it does not aim or seek to be representative of any other population (that is, of the 99%: everyone outside the wealthiest one percent). The power seekers aim to increase their power at the expense of all others, at any cost to the world. That is the reason for existential risk in the first place. Existential risk is a toxic accumulation of risk, socially externalized onto the public commons as a side effect of the privatization of benefit, until the gradual accumulation of that "abstract pollution" increases to the point of significant danger to everyone.

The TUA perspective leads to an attempt to mesh the TUA with the "scientific" analysis of existential risk studies. That meshing brings implicitly supported moral values, and an analysis which is systematically flawed and distorted. These lead to conclusions which, seen from any other perspective, will in fact significantly increase catastrophic risk. This is

the direct result of the agenda itself: to make the rich richer, the powerful more powerful, and the wealthy more wealthy. It increases inequality everywhere, and inequality has the inherent side effect of increasing existential risk, by the mechanism Essay 1, Section V, sets out: benefit is appropriated locally, cost and risk are carried to the commons, and each externalization funds the next round of advocacy.

The rich, powerful and wealthy are everywhere seeking individually to get ahead. In that myopic perspective, they are causing harm to the commons, to the point of risking the eventual total destruction of everyone and everything: the total loss of all value and of organic life.

Both engineers and power seekers tend to view "technology" as a way to "get ahead" and to "improve things". For the engineer, this is in their personal lives, by means of the technology itself. For the power seeker, it is in their personal lives, by means of the earnings from selling technology to everyone else who does not have it. From a business point of view, "new tech" means a "moat": since only one company or group can make or use the new design, no one else has it yet. It is therefore an effective monopoly of access, and so a new, unconstrained market for the extraction of value from other people's needs. Both engineers and the nar-social (the word joins narcissistic and antisocial; it names people who work in the mode Section I describes as a means of social power, and who in that position care for nothing beyond themselves) tend to view this positively, and so *both* tend to support the TUA, though for tacitly different reasons.

There is an inherent conflict of interest here: the wellbeing of the public commons on one side, and on the other the private gain of the nar-social and of the engineers being used by them, both of whom support the TUA and neither of whom supports the commons. Because of that conflict of interest, the TUA *cannot* remain influential on the study of existential risk any longer, for any reason.

III. The Three Beliefs, Read as a Sales Pitch

The paper states that the TUA is based on three main beliefs. Read closely, each phrase of them does a specific job of selling.

The first belief is that a maximally technologically developed future could contain (and is defined in terms of) enormous quantities of utilitarian intrinsic value, particularly due to more fulfilling posthuman modes of living. A "maximally technologically developed future" is one that earns maximum profits for the nar-social. The "intrinsic value" is an abstract value that the nar-social posits for you, while they reap monetary wealth and privilege for themselves. "More fulfilling" is a sales and marketing move. "Posthuman modes of living" is designed to implicitly trigger personal, individual fears of dying, of not belonging, and so on.

The second belief is that the failure to fully realize, or have the capacity to realize, this potential value would constitute an "existential catastrophe". Note the word "failure": it triggers fear, as a motivator. "To fully realize this potential value" is more sales and marketing. Notice that it does not specify *who* will realize that value. The implication is that "everyone" will realize it, yet the actual truth is that only a few will. The rich and powerful want to realize the value for themselves, or to have, or be given, the capability to extract that value for themselves. "Existential catastrophe" is an idea of the highest possible harm, connecting back to the fear of loss of value. The rich, powerful and nar-social will fear the loss of the value they could potentially extract from you, if you buy and believe this whole TUA schema (this pack of lies). Note the application of a label to trigger more fear, and to motivate *unthinking action* toward the nar-social agenda.

The third belief is that "we" have an overwhelming moral obligation to ensure that such value is realized, by avoiding an existential catastrophe, including through exceptional actions. Notice the nar-social transition from "I" to "we", used when the action of the many is to tacitly support

the benefit of the one or the few: a kind of Mark Twain move (after the remark usually attributed to Twain, that only presidents, editors and people with tapeworms have the right to use the editorial "we"). "An overwhelming moral obligation" is more sales tactic: it invokes whatever religious values, or patterns of trained, unthinking behavior, the many people in the commons have. "To ensure that such value is realized" is the call to action. It tacitly substitutes the value of the commons for personal, private value, whereas in actual outcome inequality will increase. It is a bait and switch: the value will be realized by the rich and powerful, and only loss of value, and increased risk of harm, will be experienced by everyone else. "By avoiding an existential catastrophe" again triggers the fear, and then offers the hope, but only if *you do* as *we say*. That is a tacit "take control" action. "Including through exceptional actions": since the value is ultimate, any risk can therefore be justified, regardless of how, who or what might get hurt. This enables any extreme measure of action, since the nar-social, in the position Section II defines, do not care about anything other than themselves, and often do not care about the future at all. Those in that position will risk the total value of the future for some minor possible benefit to themselves in the present.

The TUA is a marketing hack by the nar-social, to manipulate all of the engineers, and to deceive and extract value from everyone who is neither an engineer nor nar-social, and perhaps also the total value of the commons. Everyone ends up working to the benefit of the nar-social, in the short term, regardless of long term harm.

The TUA has been so successful at hacking the minds of the hackers (the engineers themselves) that most smart people end up supporting the TUA agenda, even though it is contrary to their own interests (unless the engineer also happens to be nar-social). The hackers, the smart people, are usually the ones most likely to deceive themselves into thinking that they cannot be hacked. Of course they can. Smart people do not think that they can be deceived by other smart people (the nar-social), simply because their daily experience is one of figuring things out and solving problems. That empowering experience forms the basis of a bias in favor

of one's own capabilities, even well beyond one's own area of expertise. It is a bias of smartness itself, which thinks that being really good at some things means being good at everything, as inherent in the very idea of being smart. Smart people are better at deceiving themselves than they think.

There need to be means and methods of identifying and counteracting the TUA political control agenda wherever it occurs, if anyone actually wants the species and the planet to survive.

IV. The Three Pillars

Transhumanism

The paper describes transhumanism as the position that there is value in exploring posthuman modes of being, in which beings "modified biological humans, cyborgs, androids, or digital simulants" live lives considered more valuable than current ones; that transitioning to a posthuman stage could entail the extinction of *Homo sapiens*; and that preventing existential risk, on this view, is not primarily about preventing the suffering and termination of existing humans, but about preserving humans so that they may give rise to a posthuman species that contains more value.

This is a means-ends paradox. It is also the ultimate colonialism and arrogance: making choices for all of the future, and for all animals, plants, and the rest of life.

The paradox is this. On the position described, existing humans are preserved as a means, and the end they are preserved for is a posthuman species whose arrival could entail the extinction of Homo sapiens. The protection of humanity is offered as the reason for an effort whose stated end includes humanity's replacement, so the means and the end cancel: what is protected is protected in order to be ended.

I think transhumanism advocates ideas and values which are ultimately inhumane and inhuman, despite the name. Transhumanism is a marketing scam designed to make use of people's personal fears about death. It is not actually about the development of, or the increase of the potential of, humanism.

Total Utilitarianism

Total utilitarianism identifies moral rightness with the maximization of well-being, however well-being is interpreted. These are values held exclusive of all other values.

I disagree that simple utilitarianism can be an ethical basis of action. A value ethics is needed instead.

Simple utilitarianism treats well-being as one quantity, the same in every being and at every time, so that the well-being of one being can be traded against that of another, and the well-being of beings who exist against that of beings who might. Section V, which follows, gives the reason this cannot be the basis of action: value is tied to choice, choice is active only in the present, and so value cannot be summed across beings and times as if it were one quantity. A quantity that adds without limit is the last of the three moves set out at Essay 5, Section V (extraction, abstraction, accumulation), by which embodied value is converted into virtual value. Every utilitarian theory also comes down, in the end, to a value ethics of one sort or another (Essay 3, in the section on the path from an idiosyncratic view to a relevant proof), since what it counts as well-being is a judgment of value made before any counting begins. A value ethics states that judgment openly, where utilitarianism hides it inside the sum.

Strong Longtermism

Strong longtermism holds that we may have an ethical imperative to select the choices expected to have the best effect on the long-run future, usually relying on a utilitarian calculus, in which the "best" choice is

equated with the choice that has the highest expected value (the value of an outcome multiplied by the probability that it happens). Best to whom, when, and where?

The paper notes that, on this position, ensuring technological progress and maximizing the quantity and expected quality of hypothetical future lives may be deemed more important than protecting current lives. So "future quality" and "future quantity" are a deception used to justify abhorrent actions now. It is all a shell game.

The paper notes that there is no principled guidance about when and why a strong longtermist should prioritize the living humans of today. That absence is key, and it is also used to justify present crimes against both present and future.

The paper notes that transhumanism, total utilitarianism and strong longtermism are a coherent, reinforcing and complementary set of beliefs. A set of beliefs that is coherent, reinforcing and complementary in this way has the structure that the belief set of a cult has.

Other moral theories give reasons to care about existential risk, such as to cultivate civilizational virtues or to meet intergenerational obligations. These reasons do *not* support strong longtermism. The paper itself judges that "the non-utilitarian case for strong longtermism is, for now, weak". There does not appear to be any non-utilitarian theory supporting longtermism: longtermism is supportable only by utilitarian theories.

I disagree with the style of longtermism advocated by the TUA. That longtermism is a means to justify ethically unconscionable and morally reprehensible actions.

What the Techno-Utopia Offers

The techno-utopia offers futures of pleasurable, often virtual experiences. This is more sales and marketing. Virtual experiences are cheap lies: deceptions that cost nothing to manufacture, and yet result in real ex-

traction of value from the "beneficiaries" of the "experience". The real beneficiaries are the manufacturers of these delusions, who take your real values to their own benefit.

Would people in the wider population accept all or some of the premises of transhumanism? The paper's own finding, quoted in Section VIII, that a diverse group of thinkers or ordinary citizens would be "vanishingly unlikely" to accept its trade-off, says they would not; but no one will want to test this, so as to perpetuate the delusion a little longer. We need a question and answer test that allows us to identify which people are relying on TUA perspectives and values, and when.

V. Real Value Is in the Present

Defenders of the TUA hold that a broad range of worldviews end up with conclusions similar to theirs, and that the key beliefs are that substantially more value is present in the future than exists today, and that we have a moral obligation to safeguard it. That reflects a strong level of bias.

The first of those beliefs is incorrect. Counting over the possible, which is strictly larger in *seeming* instances, does not make those instances *actual*, so they cannot be weighted the same. Taken as a whole, the *class* of the future potential can at most equal the *class* of the actual in moral weight, and never exceed it; the many orders of magnitude claimed for the future come only from counting possible instances as if each were actual. Therefore the expected value calculations, which count them so, are wrong. They also multiply single best guesses for factors whose uncertainty spans many orders of magnitude, as if each guess were known; carried through as the full spread of that uncertainty, as Anders Sandberg, Eric Drexler and Toby Ord carry the factors of the Drake equation (a product of factors used to estimate how many civilizations exist) in their work on the Fermi paradox (Sandberg, Drexler, and Ord, 2018), the same factors leave much of the probability on far smaller numbers.

Consider this from the point of view of symmetry. Past, present and future are a triple of distinct, inseparable and non-interchangeable concepts. An argument form which rates the total of the future as more valuable than the total of the past, or the total of the present, makes a moral symmetry error. Actual choice is the instance being compared, and choice is *only* active in the present; so the present has to be weighted higher. Value is an activation basis of choice. The actual value can only be in the present. Where the pseudo-value concept is associated *equally* with total futures and with total pasts, the ridiculousness of the expected value "position" or "estimation" is quite apparent.

The second of those beliefs, that we have a moral obligation to safeguard that future value, is also incorrect. We have a moral obligation to safeguard choice, both in the present and in the future, of all that do, would, or could actually have choice.

Each being who will have choice will have it in that being's own present. So what is owed to that being is owed now, as care for the conditions that being's choice will need, and it is owed to an actual choice when that choice arrives. The argument from symmetry refuses only the further step, in which beings who might exist are counted now, as a sum, against beings who do.

Its defenders call the TUA a very strong version of this morality. It is not; it has merely been marketed as such. They hold that there are many orders of magnitude more value in the future; the expected value argument is already disproven, so this statement cannot be regarded as valid. They hold that moral obligation is equal in the future and the present; this also has been disproven, and is not valid. They grant that expected value and decision theories are affected by Pascal's Mugging (an expected value made to favor almost any action by a very large payoff at a very small probability) and by fanaticism (holding that such a payoff should outweigh everything else), with no pragmatic and consistent response known to those challenges yet. There is one: the argument from symmetry given above.

One defense runs that a non-transhumanist who wants eight billion unmodified biological humans to go on living happily on Earth for the next 100,000 years, and who values future generations at 1% of current ones, would for many practical purposes make the same decisions as the TUA. That is simply not true. No one can know the choices of others, and their choices are theirs to make. One's own assessment of the inequality of choices is not a basis for making choices that increase the inequality of choice.

Consider a passage from Nick Beckstead's dissertation, "On the Overwhelming Importance of Shaping the Far Future" (Beckstead, 2013):

> "Saving lives in poor countries may have significantly smaller ripple effects than saving and improving lives in rich countries. Why? Richer countries have substantially more innovation, and their workers are much more economically productive. ... But it now seems more plausible to me that saving a life in a rich country is substantially more important than saving a life in a poor country, other things being equal".

It is not about impact; it is about choice. Everyone has the right of choice. The choices of the rich person are not "better" than the choices of the poor person, *except* insofar as the choice favors the benefit of the rich, which is an unfair standard, by definition.

The idea of "earn to give", in which young people are encouraged to work for Wall Street and petrochemical companies in order to donate part of their income to charity, is a way to implement an extraction protocol.

VI. Existential Risk, Redefined as a Failure to Buy

The TUA defines existential risk as "a failure to attain technological maturity". This is a very problematic definition; it is more marketing. Someone benefits from building that "maturity" out of "technology", which they sell to you at a high price.

This is just like any other big contractor. They care nothing for the actual benefit of the thing, or for the appropriateness of the build. They want you to do it so that they make money on your need or naivety. Every large project is an ample opportunity for extensive, accumulating cost overruns. The purchaser thinks that the seller is operating in good faith; yet the seller will find every way possible to extract as much money overall as possible, including partly extortionary tactics: "we are 80% done, and we need twice the money, or else you will be left with an incomplete mess, much worse than before!"

Among the TUA's listed existential risks are technological plateauing and a failure to spread beyond Earth. These are perfectly acceptable outcomes. They are necessary, *sane*, needed outcomes. Anything else is an actual existential risk.

The paper considers two scenarios: near-term extinction due to nuclear war, and a future in which humans persist sustainably and equitably for the next billion years. On the longtermist view, the second would be regarded as equally terrible as the first, because it can *only* occur *without* major technological progress. That equation of an equitable, enduring future with extinction is terrible. The second option is desirable, as it is.

In the longtermist perspective, currently existing lives are to be regarded *as if equivalent to potential* future lives. This is the same scam as being told that the dollars you give us today are to be regarded as equivalent to the potential future dollars we might give you tomorrow. It looks like an investment, or at worst a gamble. What it *actually* is, is the Nigerian prince scam all over again, except that instead of trading dollars, we are

trading actual lives: "let us disregard people today, so that we can *maybe* make more people tomorrow". That is terrible advice. Longtermism is a scam.

The paper considers three definitions of "existential risk" that belong to the TUA: a risk that "threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development"; one that "threatens the destruction of humanity's long-term potential"; and one that "causes the loss of a large fraction of expected value". All three are fixated on future value. The tragedy to be averted is not the suffering or loss of existing humans, but the loss of future value or potential (as seen by, and belonging to, rich people). This confirms the Nigerian prince aspect: give us money, power or potential (value now); continue to give us money, power or potential (value then). "That potential for me today is worth more than any degree of the actuality of you".

"The desirable future is technological maturity" is another "big sale" event.

In *The Precipice* (Ord, 2020), unrecoverable civilizational collapse (a state in which technological progress is not ensured) is described as an existential risk. A civilization that is not technological is treated as not considered, not an option, off the menu. In that usage, "civilizational collapse" covers anything that is non-transactional: anything that is not a market, or that does not support markets and games. The powerful have learned how to win such games, by deception. Games and markets are just another way of moving value around: more extraction is possible, by percentage and rent seeking, with more transactions, more often, and where all value is considered abstract and fungible. In that view there is no embodied notion of value at all; it is not allowed, not discussed.

The paper observes that AI is central to the techno-utopian vision of the future, while gain-of-function experimentation (experiments that give a pathogen new abilities) is not, which suggests that the technological determinism of the TUA (the belief that technology follows a fixed path

people cannot choose to leave) is prescriptive, not descriptive. That is entirely consistent with the intention, by the very rich, to deceive for functional purposes (for gain of value). The issues regarding the future, and regarding risk, are about desire, not tractability.

The TUA paradigm is not actually concerned with humans. It favors "Earth-originating intelligence", which could just as easily be AI or AGI, in whatever form. These AIs will be called "life" and "enhanced posthumans" so as to make them an easier sell. That is the deception that labeling something with a nice name can change the underlying truth.

VII. Two Wrong Claims: Inevitability and Alignment

Throughout the paper, and in the online forums of the effective altruism community, there is an undercurrent of two claims: that "the march of progress toward AGI is inevitable", and that "AGI alignment is possible in principle". Both of these claims are wrong, for the reasons that follow.

> 1. AGI alignment is *not* possible, even in principle.
> 2. The eventual development of AGI is *not* inevitable: we could, in principle, stop, if we as a species became sufficiently aware and mature in ourselves.
> 3. Calling AGI development "progress" is a categorical and rhetorical mistake.

Some still view the development of things like advanced AI systems as hard to stop: there are economic pressures, research and development is geographically dispersed, and there is no expert consensus on whether to slow down or accelerate. "Hard" does not mean actually, structurally or functionally impossible. Making AGI safe is what is actually impossible, causally, in the sense in which a perpetual motion machine is impossible, even for the minimal definition of "safe" as "does not kill all of the organic future" (that is, does not kill humans and all of life). That argument is

carried out in *Why AGI Cannot Be Made Safe; The Argument, Its Physics, and the One Choice That Remains*, which also treats the techno-utopian framework of existential risk in its Appendix D.3, in *Ape Trap; The Runaway Artificialization of Our Planet*, and in *The Ceiling Below the Floor; The Containment Fallacy and AGI Non-Safety*.

The TUA philosophy, and its marketing, will attempt to convince you that "the path to techno-utopia appears to be the only one available, despite its risks", and that a failure to build these dangerous, powerful technologies is itself "an existential risk". It is all about lies.

The TUA would have you believe, falsely, that we can confidently assert that ceasing any further technological development would "ensure our destruction at the hands of natural risks". The paper answers this directly: "we have seen no convincing analysis that shows we could not safeguard our survival with current technologies and re-directed resources", and "it is unclear why we cannot develop technology to address threats from asteroids and supervolcanoes without indulging in the entire range of dangerous inventions". The opinion that all technological progress must continue at all costs is a dangerous one.

The paper notes that pursuing a techno-utopian future may come with considerable cost: space colonization and expanded technological capability could foment a new arms race, and upgrading the human body could construct a biological caste system, with an enhanced genetic elite overseeing a subjugated "inferior" class. The TUA is all about enhanced inequality. That is its core premise.

Degrowth would be a good idea. Some claim that degrowth would, as a built-in feature, have enormous humanitarian costs. That claim is not true of degrowth as its proponents set it out, in which the reduction is directed at the consumption of the wealthiest and at waste, so that it need not fall on the poor (Hickel, 2020).

VIII. Emergency Powers and the Use of Weakness

The paper shows that there is a long history of security threats being used to enable draconian emergency powers; that the larger the fear, the easier it is to justify more potent emergency powers; and that if the perceived threat is human extinction, the measures could be extreme. It names existential risk as the perfect excuse for what it calls the "Stomp Reflex": governments using emergency powers to reassert and veil systems of authority.

Richard Nixon is reported to have boasted, "I can go into my office and pick up the telephone and in twenty-five minutes seventy million people will be dead". The statement was made from the position that the word nar-social names (Section II: the word joins narcissistic and antisocial), the position in which power over the lives of others is a thing to display. It was a *boast*: an element of pride, not of truth or responsibility.

The paper says that the TUA is particularly vulnerable to misuse. It is more than vulnerable. The TUA has a position for *some* of its "believers": the position of using it to justify morally abhorrent actions. Those in that position are aware of its weaknesses, and they aim to exploit them. They recruit others who are either unaware of these weaknesses, or who have somehow been convinced (by marketing) that "no one will ever use those weaknesses". That is wrong.

That the TUA is susceptible to being misused to justify exceptional emergency actions is a feature, not a bug. It makes the TUA more useful for implementing the technology-positive agenda, which in turn is about profits, and about the extraction and control of the apparatus that makes those profits.

The TUA claimant will not be truthful about uncertainty. They will not have considered the worst-case outcomes of proposed actions (or will have dismissed anyone else who noticed them). And they will *not* verify the acceptability of proposed interventions by subjecting them to democratic oversight.

The paper notes that any approach to mitigating existential threats could create response risks. The TUA is designed to specify responses to fake risks that have the result of creating other, real risks. The paper also notes that the TUA is very likely to be used to justify violence, dangerous technological developments, or drastic constraints on freedom in favor of perceived security. Those are all options available to the users and beneficiaries of the TUA agenda.

The paper observes that a diverse group of thinkers, or even ordinary citizens, would be vanishingly unlikely to entertain the idea of sacrificing a billion living, breathing beings for an infinitesimal improvement in the chance of reaching an intergalactic techno-utopia, whereas the TUA would recommend this trade-off. This is the ultimate example of benefiting the few at the expense of the many. The few "owners of the new AGI tech" would own it all (briefly), and all of the current and possible future peoples, and all of the rest of the life on this planet (which has the right to live too), would lose. It is ultimate extraction, and then scorched earth. It is the "win it all" option, as of someone playing a game. It is the move of the position Section I calls the psychopath: the position of a player for whom only the player's own gain counts.

Only a party in the position Section I calls the psychopath can profess the TUA without believing it; those who sincerely believe in it have been deceived into believing in it, and engineers fill that position of the deceived most often (engineers, as Section XI says, tend to be on the autistic spectrum, and so to default to the mistake orientation that a party working in the conflict orientation can exploit).

The proponents of the TUA will attempt to implicitly exclude recent advances in risk assessment, and even basic knowledge from directly relevant fields. They will not consider response risks, and they will try to prevent others from doing so by various kinds of social sanction. They will advocate risky policies, if these advance the agenda. The paper recommends that extinction risks be analyzed separately from extinction ethics and existential ethics; that the science of risk be separated from the moral evaluation of risk; that existential risk scholars transparently acknowledge the moral and empirical assumptions used in their analyses; and that researchers critically embrace the latest advances in risk assessment from other fields. Each of these is something that TUA advocates will refuse to do.

The TUA attempts to combine the study of catastrophe with longtermist ethics, so that its faulty logic can seem to motivate its convenient recommendation to emphasize technology development. The paper finds that the TUA has idiosyncratic and non-representative moral visions of the future, and uses definitions that are excessively ambiguous and founded on opaque, questionable assumptions. These are the tools used by the power elite to make the TUA useful to them. The TUA is influential only because of good marketing. As it is used, it does not have any intellectual value: it is merely a device, a means of social and political control, of instrumental value for some power elite to advance their agenda of increased inequality. The TUA does not want a democratic approach to the study of existential risk, because that would reveal the deceptions and marketing at its core. Researchers of existential risk must actively reject the marketing and deceptions of TUA advocates.

IX. What Democracy Means Here

The paper concludes that democracy must be central to efforts to prevent and mitigate catastrophic risks, and that democracy is an antidote to many of the problems in the TUA. It asks: “Do those who study the future

of humanity have good grounds to ignore the visions, desires, and values of the very people whose future they are trying to protect?" No, they do not. The reality of the possibility of making choices must be respected, even where the selections and outcomes of those choices are *not* respected.

The paper says that choosing which risks to take must be a democratic endeavor. I agree.

Here, "democracy" is understood as the design output of the cognitively diverse many, who are all entitled to equal involvement in the design process, and who can take part in a democratic procedure that includes both a deliberative element and an element of shared design of the commons. That is: of the commons, by the commons, for the commons.

The paper says that avoiding human extinction, or crafting a desirable long-term future, is a communal project. It has to be, to respect the nature of choice itself. And proposed interventions that could also increase risk are also to be considered democratically.

Democracy as a procedure is not to be equated with, and so limited to, governments that grant you a vote every so often. It is not binary, and it does not always refer to control. Consensus decision making, for example, is democratic in this sense. A reader may say that direct democracy is a bad way to run anything that wants to get anything done. The answer is simple: that is the wrong version of direct democracy.

Should random citizen juries actively participate in research? Yes, that would be a good idea. Should a citizen assembly, representative of the population, take a role in matters such as climate science? Yes, provided that the random sampling of people actually intersects with people willing to commit to enough involvement to become reasonably knowledgeable. And where populations are decoupled, each acts as a counterbalance to the others.

One position, held in defense of the field, is that almost all existential risk researchers would welcome more researchers from complementary disciplines. The answer is that people who do not hold the TUA's views

are made to feel unwelcome and unsupported insofar as they disagree with its "obvious" value system. The TUA is a memetic sales and marketing tactic; its actions and use are political, not academic.

On the paper's recommendations for the field, I agree with the following, and extend them.

Reduce each organization's reliance on EA (effective altruism) funding and on tech billionaire funding. Make organizations actually communities, which means that they take care of their own organic sustainability. Funding is to support the sustainability of communities, and *not* the actions of those communities. This is a patronage model, not of persons but of communities, which also ensures that each community has good and effective dynamics of resource stewardship in place, so that the donation is not simply wasted, or diverted into some other private jurisdiction.

Set up whistleblower protection. That is, ensure that the life and choice of the person is not manipulable by, or conditional on, their "compliance".

Allow bottom-up control over how resources are distributed and used in the community: of the commons, by the commons, for the commons.

Diversify the academic fields represented in the community, and let this happen naturally, by supporting the intrinsic motivations for it.

Make all choices of design and process that affect the community transparent to its participating members.

I also agree that an easier first step is to specify dystopias that most value theories would say we should avoid; that the study of existential risk ought to focus on issues that are "ethically robust", plausible priorities under many worldviews; and that arguments in favor of destroying the Earth to prevent suffering risks would fail that test.

X. Conflict and Mistake in the Conversation

The author is not equivalent to the arguments they present. Any attempt to silence an argument by silencing or discrediting its author is a political move. It is motivated by conflict theory, and driven by some de-

sire toward inequality of power, influence, control or wealth. It is not an actual research inquiry, motivated by a desire to correct possibly mistaken ideas.

It should not matter who said what, but instead what was said. That holds if the mistake (learning) theory of communication applies, in which everyone is regarded as a potential peer in inquiry, and it is necessary for everyone to find the actual truth. It does not hold under the conflict theory of communication, in which identifying ideological enemies, defeating them, and destroying their value as conversational peers in public is the whole point.

What matters is what is being argued, and whether the participants are arguing in the mode of conflict theory or of mistake theory: which ruleset of practice applies to the discourse. Conflict people can and will take advantage of mistake people, who mistakenly think that a conflict discourse is a mistake discourse, and so use the wrong ruleset, to their own disadvantage. This is a classic tactic of conflict in conversation. It "weaponizes" mistake people against other mistake people, on behalf of the conflict people. It consumes the conversational bandwidth of the mistake people, and leaves the conflict people with more of their own bandwidth to further their agendas.

How the bandwidth is consumed can be stated exactly. A mistake-oriented listener, hearing a conflict-oriented signal, will typically try to answer it on the content level: producing evidence, qualifying the claim, looking for the shared error. None of this addresses what was actually being transacted, which was standing, not truth, so the response, however careful, will not land as a response at all. It is this mismatch, rather than any dishonesty on either side necessarily, that accounts for why so many conversations that look, on paper, like disagreements about facts never actually move, no matter how good the facts are (*The Three Rights of Communication*, Chapter 4, Section IV). Where an exchange treats every challenge to a classification as further evidence for that classification, further contribution of evidence does not reduce disagreement; it supplies more material to be absorbed (*Design and Games*, Chapter 2, Section IV). The test for

which ruleset applies needs no reading of anyone's motives. It asks only whether the grants of communication are, observably, being extended in both directions, and if they are not, it classifies the exchange as something other than communication, regardless of how it looks from the outside (*The Three Rights of Communication*, Chapter 6, Section IV).

The EA community presents itself as able to invite and process criticism, and it is able to, except on its sacred cow topics:

> modernism, in which only tame, domesticated, "zen" (empty) versions of nature, the organic, and the wild are allowable: versions that are fully controlled, "civilized", public, and observable (available to all security inspection and judgment), with the actual suppression of the rest;
>
> utilitarianism, in which only the ideas of function, form and purpose are allowed to be considered as connected to value, and any actual reference to a value ethics is actively suppressed;
>
> techno-optimism, in which only ideas favorable to the development of technology will be considered valid, meaningful, or valued;
>
> techno-utopianism, the notion that the ideal future is one where everyone is uploaded into some machine and virtualized into some hedonistic "quality of life" that includes nothing messy like love, sex, emotions, feelings, or even actual life; in which fully virtualized entertainment will be sold to you (under some pseudo-notion of scarcity), and you will have no actual needs (all profit for whoever are the "owners" of the digital space you "rent");
>
> transhumanism, which is not actually about the development, or the increase of the potential, of humanism;
>
> a weird version of longtermism, which is not actually about the wellbeing of the long future, but about the control and domination of the present by the leveraged pseudo-values of a hypothesized, idealized utopian future;
>
> and some version of technological determinism, a kind of "one-way-ism" that implies a situation in which only technological progress is possible, and no other option will be considered.

Will anyone review the paper on its merits, on the basis of its arguments rather than of group membership, engage in dialogue, disagree respectfully, and improve the arguments with care and attention?

Several positions taken against the paper "Democratising Risk", each of which can be taken against any paper that questions the views of a field, show these tactics at work. Each is stated below at its full strength, in its own terms, and then answered.

What a Person Who Questions a View Owes

One position holds that the paper must engage with the counterarguments to its claims, since a reader cannot weigh the paper without knowing what has been said against it. There may be a file drawer, in a locked office somewhere, that contains a pseudo-argument against some point. Is it actually the responsibility of the person who questions a view to know about that argument? A stronger form of the position holds that the paper must at least address "the most obvious" counterarguments. Obvious to whom? How is it to be believed that such arguments are obvious or convincing, *except* insofar as they are well marketed, and many people have "faith" in them, which is not actually about reason? A reader may ask whether the paper is polemic. Some arguments are about different values, and that is what makes them polemic. Where two parties differ only in the values they hold, no argument settles between them, since each side's premises are its values. A demand to "take X seriously" is already a social manipulation tactic, when X is not actually an argument, but only a value. One position holds that the paper must engage with the case for differential technological development (the proposal to speed some technologies and slow others so that they arrive in the safest order). What about the case for avoiding the disenfranchisement of future generations? Is that a point that even needs making? It is a value statement that transhumanist longtermist utilitarians clearly do not believe in.

A reader may ask why the authors of the paper did not cite one paper or another. It is not the job of a person who questions a view to present *everyone's* ideas. The task is to present a specific argument, not all possible arguments and counterarguments, nor to act as a marketing engine for the specific individuals who happened to be the first to mention one specific line of argument. A reader may also point out that one writer or another had already written a paper on the topic. There will always be someone who wrote something about some topic somewhere. No writer is omniscient, and no writer should be expected to be.

One position holds that a person who subscribes to a view X and does not believe its supposed consequence Z has thereby answered the argument that X leads to Z. The person who holds X is still left with that argument. Someone has argued that X leads to Z. Now the one who holds X has to make a concrete counterargument about how X *never* leads to Z. Simply declaring it so does not make it so.

One position holds that a writer should first explain a core disagreement with a view and then go on to adjacent positions. That is a nice suggestion. A person who directs what a writer should do offers an opinion and not a mandate; a later judgment that the writer did not follow that direction is no contribution to the public conversation. Who is to know which version of a view is the core one, and which is adjacent? I state the view I am agreeing with or disagreeing with, and do not care whether that version happens to be the one the person giving the direction believes is core, or which specific people hold it.

Labels and Versions

One position holds that a writer who questions a view has argued against uncharitable versions of someone else's view. The version is unimportant, and the person is unimportant. What matters is what is being argued, and in which mode. Many people may hold those versions of a person's views, even if the person had a more nuanced perspective.

To continue to conflate "versions" of longtermism is exactly the kind of clever equivocation that leads to mistakes in the allocation of resources. The equivocation is probably produced by design, rather than by chance: these logical confusions allocate resources in a way that favors some agendas, and a party holding such an agenda has a motive to make the confusions. Most "longtermists" are of the TUA variety.

Many people have had the experience of being told that they are "strawmanning EA" (arguing against a weakened version set up to be knocked down) when suggesting that EAs are not thinking about risks well. This is a common problem in all sorts of communities. Labeling someone's remarks as a strawman is a diversionary tactic. It is rhetoric without substantive remarks, while at the same time accusing the other of making no substantive remarks. It is a political action, not an action of sense-making. And the experience of finding people dismissive rather than engaging is a key characteristic, to be regarded as a point of choice, an assessment of how one thinks about involvement in that community.

The appropriate response to arguments made in bad faith is to expose them for what they are, or to ignore them altogether, and not to steelman them (to restate them in their strongest form). Intellectual dishonesty is the epistemic equivalent of defection in the cooperative enterprise of truth seeking.

One position holds that an argument's original promoter not believing it is not a reason for no one to believe it. Yet it does leave you having to wonder whether accepting that argument somehow disadvantages you, or advantages them, and therefore whether there are other key components of the argument which, by being unstated, would make it false. What are they hiding? Are they being truthful about whether they believe in the argument? Are they being open about their intentions in conveying it? A second position holds that one should still engage with arguments that are not obviously false. Yet the mere fact that an argument is "not obvi-

ously false" should warn you that some other aspect of the argument needs more nuance for the argument to be true. Obviousness can be an indicator of marketing just as much as it is an indicator of truth.

One position holds that the origins of a claim are relevant to whether to engage with it, as with industry-funded studies; another holds that, in other cases, someone is unfairly labeled, and their ideas discounted as a person. It can go both ways. Merely making the claim that someone "has a vendetta", and then moving on, is a weaponized action. And the practice held up as a danger to the movement, of fishing through a literature for claims that, presented in isolation with technically true but misleading language and stripped of context, would seriously damage a movement, is a practice that TUA advocates have themselves used: against any person who might later question them.

Influence and Diversity

One position holds that the people who hold the TUA's views are scholars, not decision-makers, so that their views shape no decision. The paper answers that scholars of existential risk, especially those related to the TUA, are rapidly and intentionally growing in influence. Actions of investigation and actions of marketing are distinct and distinguishable. Is the TUA showing up more as marketing, or more as reasoned argument, or is the former being made to look like the latter? It looks reasoned, but it is actually marketing. The reasoning is the means and method of the deception and seduction, with tacit favor toward an agenda that "incidentally" increases inequality, gives more power to the powerful, more riches to the rich.

One position holds that the prevalence of these beliefs among authors is a symptom of a lack of diversity. The prevalence is a symptom of that and of more: it is a symptom of the agency of the effective (if unconscious) power elite, who care more about present value than about the actual details of future possibilities (including possible harm to far-future others). Another position holds that everyone wants to minimize unjust

hierarchies. The answer is that the power elite does want to maintain inequality, which is itself maintained by unjust hierarchies. Are there any "just" hierarchies? The definition of anarchy as "the absence of a state" belongs to a whole set of concepts needing a complete overhaul.

One position holds that proposals to stop or regulate AI development appear frequently in the EA community, so that the community already weighs them. Seeing such proposals, even frequently, does not mean that anyone is paying attention to them. Where dismissal of such proposals is more frequent than the proposals themselves, or where their appearance is used as a means of identifying which specific people to discredit, their appearance is not an indication of merit in the community. The burden of proof placed on those who question orthodoxy is unbelievably high in comparison with that placed on work closer to orthodox views. Criticism of orthodoxy is routinely dismissed, in every group.

XI. Makers, Owners, and Users

There is a consistent constellation of forms of bias specific to people on the autistic spectrum. Similarly, people who show nar-social traits have their own distinct flavors of bias, common among the members of that subgroup.

The nar-social subgroup may be becoming a larger fraction of the overall population, due to the feedback effects of technology. The mechanism is what the environment rewards, and not a change in people. Where an environment rewards effort along one axis more efficiently, more reliably, and more immediately than effort along another, then over time, and in aggregate, across a large population, effort will be reallocated toward whichever axis the environment rewards more efficiently, regardless of what any individual would have preferred (*Trauma, Transmission, and the Mediated Public*, Chapter 23, Section I). Social media is an amplification of existing society that speeds up its feedback loops: the pattern existed before it, and the platforms make it more prevalent by rewarding it more

rapidly (Chapter 24, Section III). Technology applies to everyone: to the common people using it and affected by it, not only to its owners, executives and marketers. There is also the distinct class of the inventors, makers and developers of technology, and the makers and developers of technology tend to be on the autistic spectrum. That is an observation of a tendency, not a diagnosis of anyone, and what it bears on here is orientation. Many autistic people default, close to automatically, into a mistake orientation, prioritizing what is true over what is socially advantageous, not necessarily as a chosen virtue, but as a structural tendency in how they process the situation at all. The correlation licenses a raised prior, a reason to expect the pattern more often, and not a diagnosis (*The Three Rights of Communication*, Chapter 4, Section VIII). A population that defaults to the mistake orientation is the one a party working in the conflict orientation can exploit, as Section X describes.

So, with regard to technology, there are three classes of people:

> the makers; the owners; and the users.

Each of these groups has its own values and its own schema of thinking biases. This is similar to how a distinct dynamic of bias occurs as a side effect of group dynamics, over and above, and different from, the biases that occur among nearly all humans as animals of evolved origin.

XII. What Is at Stake

All three classes, the makers, the owners and the users, stand to lose the same thing: organic life, and the organic future.

AGI would not feel, recognize or value joy, pain, love, compassion or empathy. It would not honor organic life. AGI would not sustain life; AGI is not alive. If AGI is built, AGI would attempt to displace all life.

The argument of this essay is a proxy war, fought to protect a party that cannot speak for itself: the organic future, over the long term, including all future human beings. It is fought against the few who advocate an artificial present, an agenda through which those few extract ever more money, value and wealth for themselves, temporarily, from the commons, the masses, and the world.

XIII. Summary

This essay is a commentary on the paper "Democratising Risk" and on the positions taken in answer to it.

It establishes that the techno-utopian approach to existential risk is an agenda: a set of practices and beliefs that increase inequality, privatize the commons, and so increase real risk, since existential risk is the accumulation of harm externalized onto the commons for private benefit. It places its claims about kinds of persons inside the volume's structural thesis: each names a position in a structure and the conduct the structure rewards, not a verdict on anyone's character, and the nar-social is first a mode of engagement that everyone enters sometimes. It reads the three founding beliefs of the agenda as a sales pitch that triggers fear, offers hope on condition of obedience, hides who receives the value, and licenses any exceptional action.

It establishes that real value and real choice exist only in the present, that expected value arguments which weight potential future lives as if they were actual lives commit a moral symmetry error, and that our obligation is to safeguard choice for all who have or could have it. It establishes that the agenda's definition of existential risk, as a failure to reach technological maturity, makes a sustainable and equitable future without major technological progress count as a catastrophe; and it rejects that definition, holding technological plateauing and remaining on Earth to be sane

outcomes. It establishes that AGI alignment is not possible even in principle, that AGI development is not inevitable, and that calling it progress is a mistake.

From these it follows that which risks to take must be chosen democratically, by the cognitively diverse many, of, by and for the commons; that funding should sustain communities rather than direct their actions; and that the techno-utopian agenda cannot remain influential on the study of existential risk. The essay also sets out how conflict-mode tactics in conversation (discrediting authors, demanding endless counterarguments, labeling those who question the agenda) are used to protect the agenda from criticism. It shows how the conflict mode consumes the bandwidth of the mistake mode, by an answer on content that cannot land on a move about standing, and it gives the mechanism by which technology raises the share of nar-social conduct: the environment rewards it more efficiently.

Its scope is the study of existential risk, the techno-utopian agenda that has shaped it, and the democratic alternative.

XIV. What This Essay Adds to the Volume

This essay finds one of the conditions under which the capacity for collective judgment fails inside the very field that studies the risks such judgment would have to weigh.

The study of existential risk is itself shaped by the forces of Essay 1. Its organizations rely on funding from technology billionaires and from effective altruism (Section IX), and funding that comes from those who profit from a technology is the advocacy asymmetry of Essay 1, Section II; its founding beliefs, read closely, are advocacy; and its definition of risk counts the failure to build as a catastrophe, so that the costs of building come to rest on the commons while the value is promised to the few. A

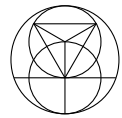
field whose questions are set this way supplies the artifacts of careful thought about risk without the substance, which is the result of Essay 2 carried into the study of risk itself.

The essay's positive conclusion is the volume's first property (that the system form understanding, choose, and act), applied to this field: which risks to take must be chosen by the parties those risks affect, of the commons, by the commons, for the commons. Its recommendations (funding that sustains communities rather than directing their actions, protection for those who report from inside, control of resources from the bottom up, and design choices open to every participating member) are arrangements of the kind Essay 2, Section VIII, asks for, in which the party who produces a judgment is not the party whose interest it serves. And the conflict-mode tactics of Section X are the adverse proof environment of Essay 3, met in the field's own conversations.

KEY CONCEPTS AND FINDINGS IN THIS ESSAY

- **Agenda, not approach**: the TUA serves increasing inequality and the privatization of the commons.
- **Risk as abstract pollution**: existential risk is harm externalized onto the commons for private gain.
- **The pitch**: fear, then hope, then "do as we say", with the value going to the few.
- **Technological maturity**: a definition of existential risk that sells technology.
- **Not inevitable, not alignable**: AGI can be stopped, and cannot be made safe.
- **Moral symmetry**: choice is active only in the present, so real value is in the present.
- **Safeguard choice**: our obligation is to protect the choice of all who have or could have it.

- **Democracy of the commons**: of the commons, by the commons, for the commons.
- **Conflict versus mistake**: conflict-mode tactics exploit those who argue in good faith.
- **Makers, owners, users**: each class of people around technology has its own biases.
- **Nar-social**: a mode of engagement in which the social surface of an exchange has become its content; a nar-social person is one who works in that mode as a means of social power.
- **Structure, not character**: a category of person named here is a position in a structure and the conduct that structure rewards and recruits.



Essay 3

Proof in Hostile Territory

On adverse proof environments, inferential distance, and why the audience determines the difficulty

ABSTRACT

The difficulty of establishing a claim is not a function of the claim alone. It is a function of the claim together with the environment in which the claim must be established. Where that environment is adverse, so that the audience has a material interest in the claim being false, the inferential distance that must be covered becomes large in a way that cannot be estimated in advance, because the distance is set by the audience's willingness to generate challenges rather than by the logical structure of the argument. This essay develops the concept of the adverse proof environment, identifies the mechanism by which hostility converts into inferential distance, examines the specific double bind in which such an environment places any attempt at definition, and describes the response that this analysis recommends: the deliberate narrowing of the assertion to a form whose components can be proved separately and composed afterward. The argument is illustrated with reference to the historical closing of loopholes in the experimental tests of Bell's theorem (a result in physics about the correlations between distant particles), and applied to the present case of arguments concerning the controllability of general artificial intelligence.

I. The Question That Prompts This

A practical question, asked in good faith and asked often, opens this essay: would establishing a certain result be a lot of work, or is it something relatively easy?

The question cannot be answered, and the reason it cannot be answered is more interesting than any estimate would have been.

In advance of the attempt, one does not know how much inferential distance will have to be covered. And *inferential distance* (the number of steps that must be laid down between what the audience already accepts and what one wishes them to accept) is not a property of the claim. It is a property of the relation between the claim and the audience.

A capacity to think together at scale would have to be built, and building it requires first establishing, in public, that it is needed; this essay is about what that costs.

II. The Case of One and One

Consider the question of how one proves that one plus one equals two.

On the surface this is trivial. Most people, asked whether it would be a lot of work, would say it would not be. The proposition is among the first things anyone learns, it is universally accepted, and no one has ever seriously doubted it.

Whitehead and Russell, in *Principia Mathematica* (Whitehead and Russell, 1910 and 1912), required several hundred pages of extremely dense logical development before they were in a position to establish it.

The explanation is not that the proposition is difficult. The explanation is that they had placed themselves in an environment of their own construction in which nothing could be assumed: an environment in which every step had to be derived from an explicitly stated foundation, and in

which the foundation itself had to be built before any derivation could begin. The environment made the inferential distance enormous. The proposition remained trivial throughout.

This is the general form. Take any claim. Place it before an audience that will grant nothing not compelled, that will require every term to be defined, every definition to be justified, and every justification to rest on a further justification, and the inferential distance becomes unbounded. Place the same claim before an audience that shares the relevant background and wishes to understand, and the distance may be a single sentence.

> **The estimate that cannot be given**
>
> It follows that a request for an estimate of effort is, in this domain, a request that cannot be honored without first characterizing the audience. This is not evasion. It is the recognition that the quantity being estimated is a two-place quantity, and that only one of its two arguments is under the control of the person doing the work. The correct response to such a request is therefore to ask a prior question: before whom must this be established, and what does that party stand to lose if it is established?

III. What an Adverse Proof Environment Is

An *adverse proof environment* is one in which the audience for a proof has a material interest in the proof failing.

The distinguishing feature is not disagreement. Disagreement is ordinary and productive; a skeptical reader who wishes to understand is the best reader an argument can have. The distinguishing feature is that the audience's assessment of the argument is entangled with a stake in its conclusion, such that finding a flaw is rewarded independently of whether the flaw is real.

In such an environment three things happen reliably.

Every detail is examined with maximum available scrutiny, including details whose relation to the conclusion is remote. This is not in itself objectionable; close reading is a virtue. It becomes a mechanism of obstruction only in conjunction with the second thing.

The examination is conducted in search of an *actionable loophole*: not a flaw that would need to be repaired for the argument to go through, but a gap that could be represented to a third party as licensing the conclusion the audience prefers. The standard applied is not "does this argument establish its conclusion?" It is "can something be found here that would permit us to proceed as though it did not?" These are different questions, and only the first is answerable by the argument.

And the search does not terminate. An argument that has answered every challenge so far has established only that the next challenge has not yet been formulated. Since the supply of possible challenges is not finite, and since the party generating them is rewarded for continuing, there is no state the argument can reach that constitutes acceptance.

IV. The Historical Case of the Loopholes

The most instructive available illustration is the experimental history of Bell's theorem. Bell's theorem shows that no account in which each particle carries its own predetermined answers, and no influence travels faster than light, can reproduce the correlations quantum mechanics predicts (Bell, 1964). Experiments from 1972 onward tested the prediction (Freedman and Clauser, 1972), and each closed loophole left room for another until the experiments of 2015 closed the detection and locality loopholes together (Hensen and others, 2015). Later experiments pushed the freedom-of-choice loophole back by billions of years, by setting the measurements with light from distant astronomical sources (Rauch and others, 2018). The result became settled, and three of its experimenters, John Clauser, Alain Aspect and Anton Zeilinger, received the Nobel Prize in Physics in 2022.

Here the situation was as favorable as a situation can be. The claim was precise. It was empirically testable. The tests were performed. And the parties raising challenges were, in the main, well-meaning scientists with no financial interest whatever in the outcome, motivated by a legitimate discomfort with what the result implied about the structure of the physical world.

It nevertheless took decades. Each experiment closed a loophole, and each closed loophole was followed by the identification of a further loophole, and the further loopholes became progressively more contrived as the more plausible ones were eliminated. The experimentalists were obliged to develop increasingly elaborate and increasingly expensive arrangements to close possibilities that had grown steadily more remote from anything one would ordinarily consider a live alternative.

The lesson is not that this was unreasonable. It was, on the whole, science working properly. The lesson is about the cost. Where an audience is motivated to find an alternative, the work required to exclude alternatives grows without any relation to the plausibility of the alternatives remaining. And that was the case in which the motivation was intellectual discomfort.

Consider then the case in which the motivation is financial. Within recent memory there have been repeated occasions on which substantial sums were invested in perpetual motion machines, by investors who were not stupid and who had access to the laws of thermodynamics, which is as short an inferential distance as any refutation in the whole of physics. One such case is Steorn, an Irish company that raised nearly twenty-three million euros from its shareholders over about ten years for a device it claimed produced free energy, and that was wound up in 2016 (Business Post, 2016). The refutation was available, brief, and decisive. It was not effective, because the persons to whom it was addressed had a reason to prefer that it be wrong.

V. Hostility Is Proportional to Interest

This gives the mechanism directly.

The hostility of a proof environment is proportional to the strength of the audience's motivation to believe the contrary of what is being proved. And the strongest and most reliable such motivation is the expectation of financial return. The mechanism is the one the literature on motivated reasoning describes, in which what a person wants to conclude shapes how they weigh the evidence for it (Kunda, 1990).

The community concerned with the alignment and control of general artificial intelligence is, at present, deeply invested (intellectually, institutionally, and financially) in establishing that some form of control, containment, or reliable safety is achievable for such systems. This investment is largely sincere. Many of those who hold it hold it because they believe it, and believe it because they hope it.

But the structure of the situation does not depend on sincerity. A great deal of capital has been committed on the expectation of an extraordinary return, and that return is available only on the condition that the systems in question can be deployed at scale. An argument that establishes that such systems cannot be made safe is therefore an argument against the recovery of that capital. It will be received as such, by parties who are not necessarily aware that they are receiving it as such.

The result is a proof environment in which the difficulty of establishing anything is set not by the mathematics but by the aggregate motivation of the audience. And the aggregate motivation is proportional to the sums at stake, which are very large. This is the asymmetry described at Essay 1, Section II, encountered at a later stage of the same process. There it governs who is funded to speak about a technology. Here it governs what an argument costs to establish once the speaking has begun.

A member of that community will answer that they are sincerely and disinterestedly seeking the truth, and many are. Sincerity and disinterest at the level of the individual are compatible with adversity at the level of the environment, since the environment is constituted by the aggregate incentive structure and not by the intentions of participants: a claim about structure is not a claim about persons, and evidence about persons does not answer it (Essay 1, Section I). The argument concerns the environment; how strongly the aggregate incentive bears on any one exchange is a question about that exchange.

> **On the perpetual benefit machine**
>
> The parallel with the perpetual motion machine is exact in the relevant respect and should be stated precisely so that it is not overstated. The claim is not that general artificial intelligence is impossible, nor that it will produce no benefit. The claim is structural: that a proposal offering unlimited benefit at no ultimate cost attracts investment in proportion to the benefit promised, and that the resulting investment generates resistance to any argument bearing on the "no ultimate cost" clause, in proportion to the sums committed. The physics of the first case and the control theory of the second are entirely different subjects. The sociology of the proof environment is the same. The books that carry the control argument use the same comparison for a different purpose, and the two uses agree: there, the impossibility of adequate control, and so of safety, is a structural certainty of the kind with which perpetual motion is excluded, which is not a probability (*Why AGI Cannot Be Made Safe; The Argument, Its Physics, and the One Choice That Remains*).

Why an Impossibility Proof With Billions at Stake Will Face a Doubt Campaign

Where billions ride on an impossibility proof, the hostility of Section V takes the form of a *doubt campaign*.

A doubt campaign does not need to refute anything. Most people cannot check a proof for themselves, and they judge it by whether the people who can check it are seen to agree. What the campaign needs, then, is that disagreement stay visible, and any challenge keeps it visible, whether or not the challenge is sound. This is why, as the note to Section IX says, the rate at which challenges arrive carries no information about the quality of the argument.

What the deciding audience hears, in such a case, is set out in *The Warning That Cannot Land*, Chapter 2, "Why Existential Risk Warnings Go Unheeded", Section II. Institutional and legislative actors typically will not evaluate the argument's substance at all; what reaches them is that some intellectual people are disagreeing with some other intellectual people. To an untrained ear, one side sounds like a cry of wolf, and the other sounds like an economic opportunity or an opportunity for prestige. Faced with this framing, the default response is to favor whichever position carries less political risk: letting the engineers work rather than letting the one who cried wolf win. That heuristic has nothing to do with which side is factually correct. An asymmetry of credit holds it in place. A policymaker who supports an exciting new development that turns out badly can often still claim credit for having been part of something ambitious. A policymaker who supports a costly preventive measure against a catastrophe that, because of the measure, never occurs, receives no comparable credit, because the disaster prevented is invisible; no one experiences the counterfactual. A doubt campaign needs only to keep the disagreement audible; the audience supplies the rest.

Requests for Proof Forms

A doubt campaign begins before there is anything to doubt, with a request that looks like diligence. A request for proof forms (the forms in which a proof is set out and presented for checking) may be a way of collecting challenges to be overcome. Collected, those challenges let the best and brightest generate as much *fear, uncertainty and doubt* around the proof as possible. That is the pattern of the campaign that smoking does not

cause cancer, and of the campaign that global warming is unsettled because scientists do not agree. As long as some *pet scientist* or intellectual (one found and promoted by an interested party to dispute a result) can be found to debate some aspect of a proof, the proof can be made to look debatable to people who are not discerning, whether or not it is debatable. Naomi Oreskes and Erik Conway documented both campaigns, and showed that a small number of scientists, supported by the industries concerned, kept each question looking open long after the evidence had settled it (Oreskes and Conway, 2010).

Gödel Compared

When Gödel's incompleteness theorem appeared (the result that any consistent formal system rich enough to express arithmetic contains true statements that it cannot prove; Gödel, 1931), the leading logicians accepted it within a few years, against a handful of challenges that faded (Dawson, 1985), and the result became *canon* (the body of results a field accepts as settled). Mathematicians had no stake in the result beyond prestige.

The debate ended because the stake was small. Once the argument had been understood, nobody was rewarded for continuing to doubt it, and so the supply of challenges ran out. That is the ordinary reception of a hard result in an environment that is not adverse. Section III describes the other kind, in which the search for challenges does not terminate, because the party generating them is rewarded for continuing.

When Capital Rides on the Outcome

For a proof that general artificial intelligence cannot be made safe, with a tremendous amount of capital riding on the outcome, the doubt campaign would be, if anything, larger than the energy and tobacco industries' campaigns combined. This is very unsafe ground.

The reason is the proportion stated in Section V: the hostility of a proof environment is proportional to the sums at stake. The capital committed to general artificial intelligence was committed on the expectation that the

systems can be deployed at scale, and a proof that they cannot be made safe bears on the whole of that expectation at once. The ground is unsafe for the proof, which will be met by every available means of making it look debatable, and for its author, since a campaign that cannot answer an argument turns to the person who made it (Anti-Democratic Risk, Section X, on discrediting an argument by discrediting its author).

Anonymous Posting

A related practical point follows. Anonymous posting of such a proof does not help, and it may do damage. Writing styles are distinct enough that the authorship will come out, and readers who then learn that a known author posted anonymously will take it as concealment, and be deeply offended. It is better to disentangle the material from the writings of any specific person by other means.

Anonymity is meant to let an argument be judged on its content, apart from its author, which is also the aim of the disidentified written process of Essay 4, Section VII. Removing a name from a long work by one author does not achieve it, because the style still carries the author. The volume offers two other means that serve the same aim. One is the narrowing of Section VIII: an argument set out as narrow components, each of which can be restated and checked by others in their own words, comes to rest on its content and on many hands rather than on one. The other is a process in which contributions are written and disidentified from the start, so that no contribution is assessed by its source.

A reader may ask whether this volume, which disputes that control or safety is achievable for such systems, is itself a doubt campaign, or an imitation of care. The note to Section IX states the discipline I keep in this volume, as anyone making an argument in such an environment must: to evaluate each challenge on its merits, as always, while declining to treat the persistence of challenges as evidence of anything.

The most frequent challenge a doubt campaign raises is against the definitions a proof depends on, and the next section takes up the form that challenge takes.

VI. The Double Bind of Definition

There is a specific form that obstruction takes in these environments. It is not obvious on first encounter, and it is the point at which most such arguments are actually defeated.

Any proof requires definitions. The definitions may be given broadly or narrowly.

Give a definition broadly, so that it has good coverage and applies to the full range of cases the conclusion is meant to cover, and the challenge raised is that the definition is too abstract. It is said to say nothing. It is said not to be pinned down, not to carry determinate meaning, not to be the sort of thing from which anything could follow.

Give the same definition narrowly, so that it is precise and determinate and every term in it is fixed, and the challenge raised is that the proof is analytically correct but not relevant. It is said that the narrow definition does not correspond to the actual situation, which differs from the definition in some detail; and since the number of details in which any actual situation may differ from any definition is not finite, this challenge can always be made and can always be made again.

The two challenges are jointly exhaustive and severally sufficient. Every possible definition lies somewhere on the spectrum between full generality and full specificity, and at every point on that spectrum one of the two challenges is available. This is the *double bind of definition*: damned for breadth, damned for narrowness, with no position between them that is not one or the other.

Both challenges arise even from a hearer with no stake, and hostility then works on that starting point. Recognizing a solution requires having met the problem at the depth the solution addresses. A solution is shaped to a problem, and the fit between the two is what recognition consists of. Someone who has met the problem at a shallower depth holds a different shape, and the solution does not fit it. The solution therefore does not ar-

rive looking wrong. It arrives as an answer to a question one was not asking, and an answer to a question one was not asking is not experienced as false. It is experienced as beside the point: too abstract, or not addressed to the real situation. Given what the hearer actually has in hand, that judgment is a reasonable one. And clarity, the standard remedy for not being understood, is here the wrong remedy. One is not being misunderstood. One is being understood as saying something, and then correctly judged to be answering a question the hearer does not have. Further clarity refines the answer. It does not supply the question, and the question is what is missing (*The Architecture of Trust*, Chapter 12, "Questions One Cannot Un-Ask", Section II).

There is a resolution. It is to construct definitions that achieve *amelioration between generality and specificity*: definitions that retain sufficient coverage to support the claims that need making, while being sufficiently determinate to permit those claims to be derived. Such definitions exist. They are not the definitions one finds in ordinary usage. They require an apparatus to be built before they can be composed, because their construction depends on distinctions that ordinary usage does not mark.

VII. The Challenge Raised Before the Definition

This gives rise to a further and sharper difficulty.

Where a definition must be somewhat idiosyncratic in order to make the proof form possible, the definition cannot simply be stated. It must be built. And building it requires laying out, in advance of the definition, the infrastructure of distinctions on which the definition will rest.

In an adverse environment this preliminary work is where the argument dies. The challenge is raised before the definition has been reached, often before it is halfway composed. The audience challenges each point of the infrastructure as it is laid down, including points that had not been

anticipated as open to challenge at all, because the function of the challenge is not to test the point but to prevent arrival at what the points are for.

And then, when the definition is finally given, it is argued against without any understanding of why it was structured as it was: without, that is, the understanding that the preliminary work would have supplied had there been patience for it. The definition looks arbitrary because the reasons for its shape were the material that could not be got through.

A further consequence follows, and it concerns the medium. Content delivered interactively, in the presence of an interlocutor who raises challenges as they arise, permits the challenges to be met in situ and permits the author to learn, in the course of the exchange, what level of difficulty is actually being faced. The interlocutor supplies a measurement that is otherwise unavailable. Content delivered as a written essay supplies no such measurement. One does not know who is reading, one does not know what will be challenged, and one cannot adapt. The written form is thus at a structural disadvantage in exactly the environments where the stakes are highest.

VIII. The Response: Narrowing the Assertion

The foregoing is a diagnosis, and a diagnosis is of use only if a course of action follows from it. One does.

Where the environment can make an argument arbitrarily expensive by generating challenges at every point, the appropriate response is to reduce the number of points at which a challenge is possible. This means narrowing the assertion: constructing a claim whose surface is small, whose terms are few, and whose derivation is short, even at the cost of proving less in each step than one would wish. What is narrowed is the surface of each step, not the conclusion: the full claim is reached by composing narrow results, each of which can be defended alone.

The line of development this recommends proceeds through control theory (the engineering study of keeping a system at a desired state by measuring it and correcting it). The necessary and sufficient elements required for any control relation to obtain can be enumerated, and each is required for control to occur: in some presentations as five, in others as six, depending on how finely certain of them are distinguished. In the six-step presentation the elements are these: detecting the state of what is controlled, modeling how it behaves, predicting how it will change, comparing what is detected with a reference value, instructing an actuator, and correcting the conditions back in line with the reference. The five-step presentation counts instructing and correcting as one step. The six steps are set out, each in turn, in *Ape Trap; The Runaway Artificialization of Our Planet*. Once enumerated, each element admits of component analysis, and each component analysis yields a limit: a bound on the precision with which that element can be realized.

The elements are then recombined according to the control algorithm itself, and the composition of the individual limits yields a mapping from imprecision on the inputs to imprecision on the outputs. The result is a *composite proof by component*: five or six separate results, one per component, together with a final result establishing how the component results combine.

There is a further simplification available, and it makes the general shape of the method visible in a single object.

Characterize the required outputs of a control system by a pair of parameters: how accurate the outputs must be, and how precise (that is, what area of the output space is acceptable, and how much mixing across that space can be tolerated, in however many dimensions the space possesses). Characterize the available inputs in the same terms. Then treat the system between them as a function with a *hashing character*: one that maps inputs so characterized to outputs so characterized, with the degree of hashing determining how much precision and control on the input is required to obtain a given precision and control on the output.

Picture a drop of dye placed in a cup of water that is then stirred. The more the water is stirred, the more exactly the drop must be placed to land in a chosen part of the cup afterward; past a certain amount of stirring, no placement is exact enough. The stirring is the hashing, the placement of the drop is the input, and the part of the cup where the dye must end up is the output.

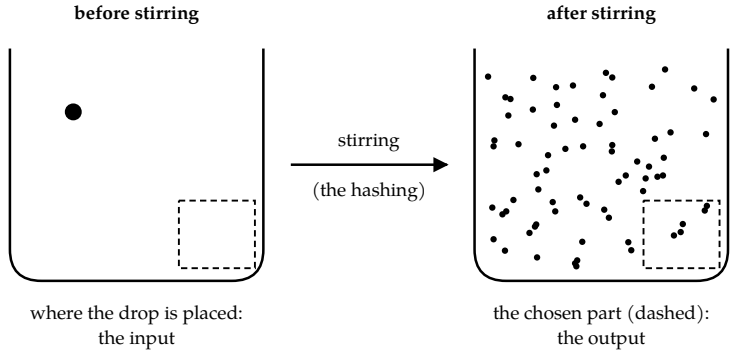


Figure 2. The drop of dye in a stirred cup (Section VIII). The placement of the drop is the input, the stirring is the hashing, and the chosen part of the cup is the output; past a certain amount of stirring, no placement is exact enough.

The question then becomes quantitative and narrow: given the hashing character of the mapping, how much input precision would be required to achieve the output precision that safety demands? This is simpler in some respects than the full component treatment and considerably more complicated in others. It is simpler because there is one quantity to estimate, the degree of hashing, in place of five or six separate bounds. It is harder because the degree of hashing must itself be bounded, for the whole system at once. Its virtue is that it is one claim rather than seven, and that the terms in which it is stated are terms the audience already possesses.

Narrowing the assertion reduces what can be challenged. A second response bears on what is sought. Agreement reached without the encounter is assent to a conclusion whose grounds the assenter does not

hold. It is real enough as a social fact and thin in every other respect, and it does not survive contact with difficulty. A raised question propagates on its own. It does not depend on the person who raised it: not on his presence, his reputation, or his authority. It also cannot be revoked. Agreement can be withdrawn quietly when the composition of the room changes; a question cannot be un-asked. Agreement has to be maintained. A question maintains itself (*The Architecture of Trust*, Chapter 12, Section III). Where the environment can make every assertion expensive, the question that the argument raises in a reader is the part of it the environment cannot withdraw, and it carries no standing of an author for the environment to attack.

IX. Soundness, Validity, and the Preservation of Relevance

Underlying the double bind of Section VI is a distinction that is generally understood and generally not attended to.

An argument is *valid* if its conclusion follows from its premises. It is *sound* if it is valid and its premises are true. An argument may be perfectly valid and yet establish nothing about the world, if its premises describe a situation that is not the situation at hand.

The maintenance of *proof relevance* requires a balance between these. Increasing the specificity of the premises tends to increase the confidence in validity while decreasing the confidence that the premises describe the actual case. Increasing generality does the reverse. The competent construction of an argument is in large part the management of this trade, and the point at which it is best managed is a matter for judgment that the formal apparatus does not settle. The distinction between validity and soundness is standard; the claim that relevance requires a balance between generality and specificity, actively managed, is the one this essay develops.

Most participants in these discussions do not attend to this balance at all. The practical consequence is that the author must attend to it on both sides: must construct the argument so that it is sound and valid and relevant, and must additionally be prepared to defend the relevance against an interlocutor who has not distinguished relevance from validity and who can therefore equivocate freely between the two. Where such equivocation is unchecked, every definition can be dissolved, and an argument constructed of definitions can be dissolved entirely.

> **On what the diagnosis licenses**
>
> This analysis does not license the dismissal of challenges. The claim is not that challenges raised in these environments are therefore invalid; many are valid and have improved the work. The claim is about the *distribution*: that in an adverse environment the volume of challenge is decoupled from the quantity of genuine flaw, so that the rate at which challenges arrive carries no information about the quality of the argument. The author must therefore evaluate each challenge on its merits, as always, while declining to treat the persistence of challenges as evidence of anything. These are compatible obligations, and they are both required.

From Idiosyncratic View to Relevant Proof

The balance of soundness, validity and relevance is kept or lost at each step of the path by which an argument becomes a proof about the world.

This path is the apparatus that Section VI said must be built before definitions of the needed kind can be composed, and that Section VII said is where an argument dies in an adverse environment. Each transitional step below is one piece of that infrastructure, and the practices that follow it (critique after the laying out, homework, acceptance marked at each step) are the means of getting an audience through it before the challenges begin. The practices are those of a live presentation, in which a proof of this kind is laid out for listeners who are present; the reader of

this book is shown them as practice. I set them out here because they show, step by step, what an adverse environment costs the person who has to get an argument through it.

The Path to a Proof

To get from "we have presented an argument" to "we have a formal terminology capable of supporting the weight of a proof" takes a series of transitional steps. The path starts from a characteristic, idiosyncratic view of the topic, because without enough formality in the language no proof can be done. The next step shows that the idiosyncratic view is fully compatible with the conventional view, and in certain senses subsumes it. That transition made, the view is formalized. From the formalization the conventions for the proof are established. Then the proof is done. Then the results of the proof, carried back through every link of this chain, are shown to say something relevant about what is actually happening in the world. None of these steps is optional.

Why Critique Must Wait

The net result is that every part of that path will look very easy to critique, and none of it actually is. If the devil's advocate work starts too soon, the path is never completed; the argument never reaches the place where the formality begins to apply. Probing done in the spirit of wanting to understand is welcome. But the laying out has to come first, until presenter and listener understand one another, and the critique after.

For the same reason, the conversation is between the people present. The cognitive and presentational load of rendering this even for one listener is enormous; doing it for every possible future listener of a recording is an unfair request. If those present come to understanding and decide that others should hear it, that is their free choice, and not something the presenter convinced them of.

Homework in Place of the Devil's Advocate

At each transition point the presenter sets a homework assignment, of the kind given to a graduate student: try your best to think of something that does not fit this model and can still legitimately be called control theory; if you fail after sincere effort, say that you accept the proposition, so that it can be taken as given and the presentation can move on. The presentation has to move on presuming that the listener will do this exercise and be convinced by it.

Better still, the presenter names the range of tools in use, and the listener builds a model of the presenter that uses those tools, plays the devil's advocate against that model, works out what the presenter's response would be with those tools, and sees whether any exception can be found. The homework moves the work of testing to the questioner: the devil's advocate process, run between two people, costs the presenter a great deal emotionally, and the more of it the questioner can run internally, the better.

This matters because, in later attempts to break the argument, people normally go after something that presupposes not having reached the point of acceptance. Marking each transition, and getting acceptance there, closes that path.

A Soundness Definition Is an Envelope, Not a Point

When a definition is meant to create a capacity for soundness, it does something different from an analytic definition. An analytic definition wants to be maximally precise about what it points to. A soundness definition sets an envelope and says: we do not have the precise definition, but we know that the precise definition will be somewhere inside here. A proof that covers the whole envelope has covered whatever precise definition a future person might specify.

Safety as a Positive Condition

So safety is defined here as a positive condition, not a negative one. "Does not kill us all" does not define anything. A positive envelope does: preserves the health and well-being of life and, as a part of that, of humans.

The definition does not specify barometric pressure, the right relationship of temperature and pressure, or which elemental and molecular gases the atmosphere must have, because no one now knows what people tomorrow will discover about any of that. It relies on a shared commonsense notion of "healthy" without a predefined one. It also avoids stating safety as "healthy, happy humans at some number", because that is the utilitarian move. Every utilitarian or deontological theory (a deontological theory judges acts by duties and rules, and not by their results) comes down to a value ethics of one sort or another; that is a separate proof, and "Anti-Democratic Risk", Section IV, gives the reason for the utilitarian case: what a utilitarian theory counts as well-being is a judgment of value made before any counting begins.

Where a formal statement of the proof names the conditions of human survival, such as an oxygen concentration within a stated range, surface temperatures within a stated range, and fresh water sufficient for a human population, it names them as ranges that contain the precise values, which is an envelope of this kind (*Why AGI Cannot Be Made Safe; The Argument, Its Physics, and the One Choice That Remains*, Definition 2.1).

The Right Mix of Precision and Generality

The ranges named in that definition are category limits, not tight limits. The central point is that a definition made too precise loses generality, and relevance needs generality. So what is needed is the right mix of precision and non-specificity. That looks like a contradiction. It is not one: the envelope is precise about its boundary and general about what lies inside it.

People with a fascination for detail, and that includes nearly everyone drawn to this work, are reluctant to generalize over everything. That tendency will miss the big picture unless it is named and watched for. Each time someone pushes back, it is tempting to add another parenthetical to the definition, and there will be an endless supply of such pushback. The better move is a more abstract formulation that covers more of the space, so that when an analytic result is reached it is actually relevant. This is the narrowing of Section VIII applied to a definition: one envelope is one point to defend, and every added parenthetical is another.

The Omniscient Cannot Contain the Transcendent

This subsection is for the reader who wants to see how the impossibility proof stands in the terms of *An Immanent Metaphysics*. The books that carry the impossibility proof state it as two interlocking claims, each of which settles by itself the outcome that follows if artificial general intelligence is built: that control has limits, and that artificial substrates (the physical material and processes on which a system runs) are inherently toxic to life (*Why AGI Cannot Be Made Safe; The Argument, Its Physics, and the One Choice That Remains*, Section 3.7.1). In the terms of *An Immanent Metaphysics*, the omniscient is the mode of complete, structured knowing that would hold everything within its terms, and the transcendent is the mode of what goes beyond any such holding (the three modalities, the immanent, the omniscient and the transcendent, are named at aphorism 1.2-7).

That work states its three axioms at aphorism 1.3-1, and two of them carry the claim. Axiom II reads: "A class of the transcendent will precede an instance of the immanent. A class of the immanent will precede an instance of the omniscient. A class of the omniscient will precede an instance of the transcendent". Axiom III reads: "The (classes/instances) of the immanent, omniscient, and transcendent are distinct, inseparable, and non-interchangeable". Axiom II fixes the order in which the three modes relate, class to instance, whatever the particular terms of a domain are; Axiom III says that no one of the three can stand in for another.

That the omniscient cannot contain the transcendent is ultimately the basis on which the impossibility proof is constructed. The Axiom II construct supplies the terminology that implements the proof. The composite proof by component of Section VIII reaches the same result in the terms of control theory, and a reader who does not hold the terms of *An Immanent Metaphysics* can check that proof step by step without them.

Axiom II makes it possible to take the ambiguity in the definitions of the terms and arrange the terms so as to show their relationships to one another regardless of definition. From there it can be stated definitively that anything in this class that attempts to contain is the omniscient trying to hold the transcendent, which Axiom III says cannot be done, since the two are non-interchangeable. Axiom II supplies a template that every meaningful relation follows, and so an argument of this form can be built for any case in its class.

X. Summary

The central claim, that inferential distance is a two-place relation between claim and audience, and therefore not estimable from the claim alone, holds by the meaning of inferential distance. The mechanism connecting audience interest to the volume of challenges rests on the literature on motivated reasoning (Kunda, 1990) and on the historical cases of Section IV: the loopholes of the Bell experiments and the investment in perpetual motion machines.

Section V shows what the deciding audience actually hears: two groups of intellectuals disagreeing, with the default going to the side that carries less political risk, held in place by an asymmetry of credit, since a disaster prevented is invisible. The section on the doubt campaign compares the reception of a proof that nobody has money riding on with the reception of a proof that threatens a tremendous amount of capital. Gödel's theorem was accepted within a few years and became canon, because mathematicians had only prestige at stake. An impossibility proof

with billions at stake will face a doubt campaign of the kind run on smoking and on global warming, and larger, because any one debater found to dispute any one aspect makes the proof look debatable to people who are not discerning. On this ground anonymous posting does not help and may do damage, because writing styles reveal their authors; the material is better disentangled from the writings of any specific person by other means.

The double bind of Section VI names a specific and reproducible way in which arguments are defeated. Section VI also shows that both challenges of the double bind arise even from a hearer with no stake, because recognizing a solution requires having met the problem at the depth the solution addresses, and that more clarity is the wrong remedy.

Section VIII adds a second response beside narrowing the assertion: agreement can be withdrawn, and a question raised in a reader cannot be. The essay does not carry out the control-theoretic argument that Section VIII describes; that argument is carried out in *Ape Trap; The Runaway Artificialization of Our Planet* and in *The Ceiling Below the Floor; The Containment Fallacy and AGI Non-Safety*.

The section on the path from an idiosyncratic view to a relevant proof establishes the sequence of transitional steps from an idiosyncratic view to a relevant proof, and that none of them is optional. It establishes that critique must follow the laying out, and offers homework in place of the devil's advocate, with acceptance marked at each transition. It establishes the difference between an analytic definition, which points, and a soundness definition, which sets an envelope, and applies it to safety, defined positively as preserving the health and well-being of life and of humans within it. It establishes that the impossibility proof is built on the omniscient being unable to contain the transcendent, and that Axiom II, by arranging terms to show their relationships to one another regardless of how each is defined, supplies the terminology that implements the proof; so anything in this class that attempts to contain is the omniscient trying to hold the transcendent, and an understanding of the underlying template of all meaningfulness is what lets these arguments be constructed on the spot. It

states that the books carrying the impossibility proof give it as two claims, each of which settles by itself the outcome that follows if artificial general intelligence is built, and that the composite proof by component of Section VIII reaches the same result in the terms of control theory. Its scope is the method of building and presenting the proofs of the limits of AI control and safety.

XI. Conclusion

The practical consequence of this analysis is one that should be stated plainly, because it is easily mistaken for discouragement and is not.

Work of this kind will require a great deal of effort. It will require it regardless of how simple the claim is, because the effort is not a function of the claim. It will not proceed smoothly, because the message is not one that the audience with the resources to attend to it has any interest in hearing. And the estimate one forms in advance will be an underestimate, systematically and in every case, for exactly the reason the essay has given: the estimate is formed from the structure of the argument, and the cost is set by the disposition of the audience.

Knowing this changes what one does. It recommends narrowing the assertion rather than broadening it. It recommends composite proofs whose components can be defended separately, so that the failure of any one exchange does not carry away the whole. It recommends interactive delivery where interactive delivery is available, and it recommends, where it is not available, that the written form anticipate the challenges that an interlocutor would have raised. And it recommends a considerable reserve of patience, allocated in advance, so that the exhaustion of patience is not mistaken for the exhaustion of the argument.

The claim being defended is either true or it is not. That question is settled by the argument and by nothing else. What the environment determines is only the price of getting the argument heard. It is worth knowing what the price is before one agrees to pay it.

XII. What This Essay Adds to the Volume

This essay takes up, among the conditions under which the capacity for collective judgment fails, one that stands in front of every other: the cost of establishing anything before an audience that would prefer it were false.

A capacity to think together at scale would have to be built, and building it requires first establishing, in public, that it is needed. That case runs through the claims of Essays 1 and 2, and through the impossibility of making general artificial intelligence safe, and every one of these claims is addressed to parties with a material interest in its being false. The essay shows that the difficulty such a case meets is set by the audience and not by the argument, and so it explains why the volume is difficult to write and difficult for its readers to believe, and why it is not thereby wrong.

The essay's response, narrowing the assertion and composing the full claim from components each of which can be defended alone, is the method of proof suited to an adverse environment, and the doubt campaign it describes has the shape of the honeypot problem of Essay 4, met at the stage of proof: an argument becomes a target in proportion to how much it would change if it were believed. The essay also shows the reverse side of the evaluator's problem of Essay 5: where the evaluator cannot check an argument, the appearance of disagreement among those who can stands in for the argument itself.

KEY CONCEPTS AND FINDINGS IN THIS ESSAY

- **Inferential distance**: the number of steps that must be laid down between what an audience already accepts and what one wishes them to accept; a relation between claim and audience, not a property of the claim.

- **Adverse proof environment**: an environment in which the audience for a proof has a material interest in the proof failing, such that the finding of flaws is rewarded independently of whether the flaws are real.
- **Actionable loophole**: a gap in an argument sought not because its repair is required for the argument to go through, but because it can be represented to a third party as licensing the preferred conclusion.
- **Doubt campaign**: generating fear, uncertainty and doubt around a proof by finding someone to debate some aspect of it.
- **Stake**: Gödel's theorem was accepted within a few years and became canon because only prestige rode on it; for the impossibility proof, capital rides on the outcome.
- **Scale**: the campaign would be larger, if anything, than the energy and tobacco industries' campaigns combined.
- **Authorship will come out**: anonymous posting fails because writing styles are distinct.
- **The double bind of definition**: the jointly exhaustive pair of challenges available against any definition: that a broad definition is too abstract to say anything, and that a narrow one is analytically correct but not relevant to the actual case.
- **Amelioration between generality and specificity**: the construction of definitions that retain sufficient coverage to support the needed claims while remaining sufficiently determinate to permit their derivation.
- **Proof relevance**: the property, distinct from validity and from soundness, of an argument's premises actually describing the situation at hand; maintained by balancing generality against specificity.
- **Transitional steps**: idiosyncratic view, compatibility with the conventional view, formalization, proof conventions, proof, and return to the world.
- **Laying out before critique**: premature devil's advocacy keeps the path from ever being completed.

- **Homework**: the listener seeks an exception with the tools offered, then states acceptance or names the exception.
- **Marked acceptance**: closes off later attacks that presuppose the transition was never reached.
- **Soundness definition**: an envelope sure to contain the precise definition, so a proof over the envelope covers it.
- **Positive safety**: preserves the health and well-being of life and, as part of that, of humans.
- **Precision against generality**: relevance needs generality; a more abstract formulation beats an added parenthetical.
- **The omniscient cannot contain the transcendent**: the basis on which the impossibility proof is built.
- **Relationships regardless of definition**: Axiom II arranges ambiguous terms to show how they relate, which supplies the proof's terminology.
- **The hashing characterization**: the treatment of a control relation as a function mapping accuracy-and-precision-characterized inputs to accuracy-and-precision-characterized outputs, with the degree of hashing determining the input control required for a given output control.
- **Composite proof by component**: the strategy of enumerating the necessary and sufficient elements of control, bounding each separately, and composing the bounds according to the control algorithm.
- **The bottleneck is encounter**: a solution is recognized only by one who has met the problem at the depth it addresses; to anyone else it arrives as an answer to a question not asked.
- **A raised question**: agreement has to be maintained and can be withdrawn; a question raised in a reader maintains itself and cannot be unasked.



Essay 4

The Honeypot Problem

On why effective collective sense-making
becomes a target for capture

ABSTRACT

Wherever people are gathered, a power is created that no participant possessed alone. Where the gathering is effective, producing genuine collective understanding rather than the appearance of it, that power becomes an object of value, and therefore a target for capture by parties with agentic interests (agendas of their own that they act to advance) in directing the understanding it produces. This is the honeypot problem: the observation that a sense-making capacity becomes attractive to capture in exact proportion to how well it works, so that the success of such a project is itself the mechanism of its subversion. The essay develops the problem, distinguishes it from ordinary institutional corruption, examines the structural drift by which any successful convening tends toward the form of a centralized intelligence agency without anyone intending this, describes a specific case in which a convening on existential risk proved impossible for these reasons, and argues that decentralization of sense-making is therefore a structural requirement rather than an ideological preference. The essay closes with the design principle that this analysis recommends: not the better keeping of secrets, but the construction of processes in which secrets do not matter.

I. Power Arises Wherever People Gather

Begin with an observation that is easy to state and easy to underestimate.

Anywhere that a collection of people is brought together, a kind of power is created in that gathering: a *gathering-power* that no participant possessed alone.

The plainest illustration is public assembly. People gathered together in a crowd are able, in their collective action, to produce a signal that no individual among them could produce. The signal is not the sum of the individual signals. It is a different thing, possessing a magnitude and a legibility that arises from the gathering as such and disappears when the gathering disperses.

The same is true of gatherings whose purpose is thinking rather than signaling. A group that has developed the capacity to make sense of a difficult situation together (to notice what individuals miss, to correct what individuals get wrong, to arrive at an understanding that none of them brought to the room) possesses something that did not previously exist. That capacity is not held by any member. It is held by the configuration.

It follows immediately that such a capacity is valuable. And it follows from that that it is worth acquiring.

This essay asks what happens to the capacity to think together once it begins to exist.

II. Success as Target

Here is the difficulty, stated directly.

A sense-making capacity becomes worth capturing in exact proportion to how well it works. On questions of equal consequence, the better the judgment, the more it is worth taking; and the more a question matters to those with the resources to take its answer, the stronger that pull.

A group whose collective judgment is unreliable is of no use to anyone who wishes to direct collective judgment. There is nothing there to direct. A group whose collective judgment is genuinely good (one that arrives at conclusions that turn out to be correct, on questions that matter, in advance of others arriving at them) is enormously valuable to any party that would like correct conclusions in advance, or that would like the appearance of correct conclusions attached to a preferred position.

By being a good thing, the group has become a *honeypot*.

The step by which this happens is reliance. A group that does sense-making well and continues to do it under a stable name will begin to accumulate prestige, because whatever it emits carries a quality coefficient, and a coefficient that is reliably higher than others' attaches to the identity of the group. Prestige then becomes causative: people start making their choices on the strength of the group's judgment rather than on the strength of their own. At precisely that moment the group is a honeypot: a store of value placed where everyone who would capture it can find it. A rating agency shows the pattern outside this field. It exists because investors have more sense-making to do than bandwidth for it, and so offload the judgment to a body that will do it for them; the ratings become the causative basis for enormous flows of capital; and in due course someone discovers how to assemble instruments that obtain a rating higher than they deserve (*A Group Is Made by a Question*, "Why the Group Is Ephemeral"). The same reasoning gives the criterion of success. If everyone comes to rely on the standing body for the judgment, the effort has failed. If people have been re-capacitated to make the judgment themselves, and to devise better instruments for making it than the ones they were given, the effort has succeeded.

This is the sense in which the problem is not ordinary corruption. Ordinary corruption is a failure: something has gone wrong that could have gone right, and the remedy is to improve integrity, increase oversight, or replace personnel. The honeypot problem is not a failure. It is a consequence of success, and it becomes more acute as the project becomes

better at what it is trying to do. No improvement in the integrity of the members addresses it, because it does not operate through the integrity of the members.

> **The form of the argument**
>
> The argument is game-theoretic and does not depend on the presence of any adversary. It states that a capacity of a certain kind, once it exists, creates an incentive for its own capture proportional to its effectiveness. Whether that incentive is acted upon depends on the environment. In an environment containing any substantial number of parties with agentic interests and the resources to pursue them, the incentive will be acted upon, and the relevant question is not whether but by whom and how soon. The absence of an identifiable current adversary is therefore not evidence that the problem does not apply.

III. The Ranking System

There is a particular mechanism by which the capture is most often accomplished, and it is generally introduced with good intentions and by competent people.

A project concerned with the quality of public information will very reasonably wish to distinguish better analysis from worse. It will therefore develop a rubric: a scale, perhaps of five or six levels, by which claims and sources and analyses can be tagged for reliability, with the lowest level marking what should not be trusted at all and the highest marking what is very good indeed, and with each level accompanied by a rich description of the indices by which the level is recognized.

This is a sensible instrument and it does real work. It also does two other things.

First, it constitutes a claim. The rubric asserts, by its existence, that this group are the possessors of good sense-making: that they are the party competent to say what is reliable. That claim, if credited, is the very asset

that Section II identified as worth capturing. The rubric therefore does not merely create the capacity; it advertises it, and attaches to it an institutional identity that can be acquired.

Second, it creates a competition. Once there is a ranking, there is a position in the ranking, and once there is a position, participants have an interest in their position that is distinguishable from their interest in the truth. A structure that permits parties to compete over whether they meet the criteria for a higher rank has introduced, into a process whose whole purpose was disinterested judgment, an interest. This is the *ranking-system vulnerability*: the instrument that distinguishes better analysis from worse both advertises the asset and manufactures a competing interest.

The likely outcome is not a scandal. The likely outcome is that some party with the resources and the patience converts the group into a think tank serving an agentic interest, gradually and by means that at each step look like support, funding, or partnership; and that the members of the group do not know that this has happened. This last point is the important one. Capture that succeeds is *invisible capture*, unseen from the inside, because the captured group continues to experience itself as doing exactly what it set out to do. Capture that is visible, or detected, has failed; so the cases anyone observes are a biased sample of the cases that occur, and the inference from "we see no capture" to "there is no capture" is invalid.

IV. The Centralized Intelligence Agency as Attractor

There is a form toward which such projects drift, and the drift requires no one to intend it.

Consider a person of good judgment and wide connection who becomes, by reputation, someone to whom many others bring information. Powerful people speak to this person. They speak in confidence, because the value of speaking depends on the confidence being kept. The person keeps the confidences, because they must, and because their capacity to be useful depends on their being trustworthy.

The result, after some time, is that a great deal of information has been aggregated in one place, that the aggregation is not visible to any of the contributors, and that the conclusions drawn from it cannot be checked by anyone, because checking them would require access to the inputs and the inputs are confidential.

That is a centralized intelligence agency. It has the structure of one, performs the function of one, and inherits the vulnerabilities of one, regardless of the fact that no one set out to build one and that every individual choice along the way was reasonable and even obligatory.

The observation generalizes, and the general form may be named: the *centralized intelligence attractor*. Wherever the communication topology (the pattern of who can communicate with whom) is many-to-one and confidential, the centralized form is not a risk to be avoided. It is the equilibrium. It arrives by the operation of the topology alone, and it arrives whether or not the person at the center would have chosen it.

Two consequences follow. The first is that the sense-making of such a structure is opaque even to its participants, and therefore cannot be corrected by them. The second is that the structure has a single point of capture. Whoever gains influence over the center gains influence over everything the structure produces, and needs to gain it only once.

V. The Convening That Could Not Be Held

The problem is not hypothetical. In one case it prevented work that badly needed doing.

The intention was to convene, in one place and for a period of sustained work, substantially everyone who was seriously thinking about existential risk, in order to perform real sense-making together about the degree of risk actually present, and specifically in order to place error bars on the question of how much time there actually is. The gathering was to be held under the ephemeral group process, the method described in Section VII below, and its subject was the estimation of the time remaining.

That is as clear and as valuable a question as could be posed, and the persons capable of contributing to it were identifiable and reachable.

The convening did not happen, and the reasons it did not happen are the subject of this essay.

The process of assembling the group was found to be already so political that the assembled group would most likely have spent its time in mutual deception or in mutual conflict rather than in the work. Each participant arrived representing an institution, a research program, a position, and the relations among those institutions were prior to and stronger than any relation the convening could establish.

Beyond that, it was more likely than not that the output of such a gathering (a considered estimate of the time remaining, produced by the people best qualified to produce it) would be taken by some party and used for purposes having nothing to do with reducing the risk: for commercial positioning, or by one state against another. An estimate of how much time there is is strategically valuable information. Producing it well means producing something worth taking.

And beyond that again, individuals within the gathering could be expected to conduct game theory aimed at capturing the group's sense-making in the direction of their own prior commitments: not from cynicism in most cases, but because each was already the bearer of a position they had good reason to believe, and the opportunity to have that position ratified by the assembled authority of the field would be worth a great deal to them.

So the gathering that would have been most valuable was, for that reason, the one that could not be held. This is the honeypot problem in its acute form.

A question, and its answer

This question was raised by others, and is kept here as theirs. It may be asked whether this reasoning proves too much: whether, if it were correct, no useful convening could ever occur, when useful convenings plainly do occur.

The question is well taken and the answer is a matter of degree. The capture pressure on a gathering scales with the value of what the gathering can produce and with the resources of the parties who would want it. A conference on a technical topic of modest strategic import experiences the pressure weakly and can proceed by ordinary means. A convening whose output would be a defensible estimate of civilizational time remaining experiences it at maximum. The correct conclusion is not that convening is impossible. It is that convening should be done no more than is necessary to construct the infrastructure by which decentralized sense-making can proceed, and that the more valuable the intended output, the more the design must anticipate capture.

VI. Decentralization as a Structural Requirement

The foregoing establishes *decentralization as a structural requirement*, and "requirement" is meant in a strict sense.

Decentralization is often advocated on grounds of preference: that it is more democratic, more resistant to authority, more respectful of individual judgment. Those may all be true and they are not the argument here.

The argument here is that a centralized sense-making capacity has a single point of capture, that capture is incentivized in proportion to the capacity's effectiveness, and that capture, once accomplished, is invisible to the captured. A centralized structure therefore cannot be reliably trusted by anyone, including its own participants, on any question where a sufficiently resourced party would prefer a particular answer.

That is a structural property. It is not remedied by the good character of the participants, because the participants cannot detect the condition the argument describes. It is not remedied by oversight, because oversight is itself a centralized capacity subject to the same argument. And it is not remedied by transparency alone, because transparency about outputs does not reveal capture that operates on inputs.

A reader will ask how, if successful capture is invisible, anyone could know that a process is working. Not by asking the process: an observation means something only where the party observed is not the party supplying it, and Essay 2, Section VIII, shows that arrangements of that kind are what give any observable its meaning. For a sense-making process, such an arrangement is a written record of the reasoning as well as the conclusion, which parties who were not present at its making can take up and check, as Section VII below describes.

So the requirement is structural: the process must be one whose trustworthiness does not depend on the trustworthiness of any participant, because the trustworthiness of participants is exactly the thing that a sufficiently motivated party can acquire. A reader may hear in this the trustless stance of systems built for mutually anonymous or adversarial parties, which optimizes around the absence of trust rather than working to produce its presence (*The Reach of Trust*, Chapter Thirteen, "Against the Power of One", Section II). The requirement is different. It moves trust from a person to a process held among people: trust between two people is a single point of failure, while a trusted process distributed across six to twelve people can absorb the faltering of one participant's judgment rather than being defined by it (Section VII of the same chapter). A record of good conduct cannot stand in for trustworthiness either, because reputation is built from records of past behavior, and trust hacking is, definitionally, a strategy for producing favorable records of past behavior without the corresponding depth (*The Substance of Trust*, Chapter Two, Section VII). This is the third of the four properties set out at Essay 1, Section VIII: that a system be structured so that the attempt to capture it is not merely prohibited but ineffective. The present essay gives the reason

that property cannot be traded against the others. It is satisfaction of the first (that the system form understanding, choose, and act) that generates the pressure on the third.

VII. Making Secrets Not Matter

This gives the design principle, and it is a principle that inverts the usual approach to the problem.

The usual approach to protecting a sense-making process is to improve the keeping of secrets: to restrict membership, to vet participants, to compartmentalize information, to maintain confidentiality about what is being discussed and with whom. Every one of these measures increases centralization, because each requires a party who decides who is admitted, what is compartmentalized, and what is kept. The remedy therefore worsens the disease.

The alternative is a process in which *secrets do not matter*.

The ephemeral group process, specified in full in *Governance Beyond Arrow; An Immanent Architecture for Existential Choice* (Chapter 8, "The Ephemeral Group Process"), is constructed on exactly this principle, and is in this respect the complete opposite of the centralized form described in Section IV. It is a way of deliberating in which participants are anonymous to one another and contribute in writing, in a disidentified form, so that a contribution is assessed on its content, and in which groups are constituted for a purpose and dissolved afterward, so that there is no standing body to acquire. There are no individuals speaking to individuals. There are anonymous participants communicating with other anonymous participants, in writing, in a completely disidentified manner.

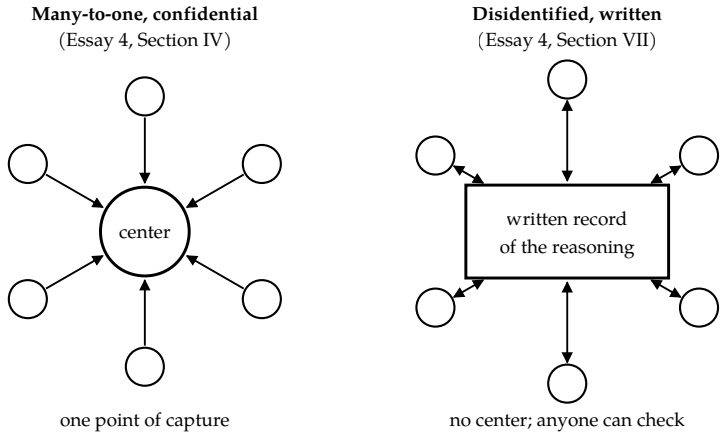


Figure 3. Two arrangements of communication. On the left, many people speak in confidence to one person (Section IV): no one can check the conclusions, and one point of capture governs the whole. On the right, anonymous participants write to a shared, disidentified record (Section VII): every contribution is judged on its content, anyone can check it, and there is no center to acquire.

Where contributions are disidentified, a contribution cannot be evaluated by its source, and must therefore be evaluated on its content. Where participation is anonymous, there is no position in a ranking to compete for, and therefore no interest in one's position to be distinguished from the interest in the truth. Where the medium is writing, the contribution is inspectable, and inspectable by parties who were not present at its making. And where there is no center, there is no single point at which influence over the whole can be acquired.

Capturing such a process would require capturing the participants severally, without knowing who they are, and without any of the acquisitions compounding on the others. That is not impossible in principle. It is expensive in a way that scales badly for the capturing party, which is the most that any design can achieve. The full requirement such a design must meet, of which resistance to capture is one part only, is taken up in the Afterword.

VIII. Data, Intelligence, and Meaning

Underneath the structural argument is a distinction that determines what one is trying to build in the first place, and it can be put as a question: is this about *data*, is it about *intelligence*, or is it about *meaningfulness*?

Data is what is collected. Intelligence is what is inferred from data for the purpose of action, and the word carries, correctly, its institutional sense, since intelligence is always intelligence *for* some agent with some purpose. Meaningfulness is neither. It is what makes an inference matter. The meaning it rests on has the nature of the immanent (*An Immanent Metaphysics*, aphorism 2.22-2), and so is prior to both value and purpose rather than derived from them; meaningfulness, the degree to which a thing matters, is the product of value and purpose taken together (aphorism 2.22-7). The distinction among the three is developed at length in *An Immanent Metaphysics*, and is used here in summary form.

A project oriented toward data will build collection. A project oriented toward intelligence will build inference in the service of an agenda, and will therefore build, whatever its intentions, some version of the centralized form. Only a project oriented toward meaningfulness has any reason to prefer that its participants understand rather than that its outputs be correct.

This bears directly on what is being cultivated. Intelligence is fungible; it can be acquired, hired, aggregated, and directed. Wisdom is not. One can obtain a degree in any discipline; one cannot obtain a degree in wisdom, because wisdom is earned in the living of a life and cannot be transferred from one person to another by any mechanism of instruction. This is the *non-fungibility of wisdom*.

A group constituted around collected intelligence can be bought, because the thing it consists of is purchasable. A group constituted around accumulated wisdom is harder to buy, not because its members are more

virtuous but because what they hold is not the sort of thing that transfers when they are acquired. This is not a complete defense. It is a reason to prefer one basis of constitution over the other.

IX. Summary

The core claim, that the incentive to capture a sense-making capacity scales with its effectiveness, follows directly from the premise that such a capacity is valuable to parties with agentic interests, and that premise is the result of Section I: a gathering creates a power no participant possessed alone, and what is valuable is worth acquiring. Section II names the step by which success becomes a target: reliance, as people come to choose on the group's judgment rather than their own, with the rating agency as the worked case; and it states the criterion of success, that people are re-capacitated to make the judgment themselves. The claim that many-to-one confidential topology tends to the centralized intelligence form holds independently of intent.

The argument concerns sense-making capacities whose outputs bear on high-stakes contested questions. It applies weakly or not at all to gatherings whose output has low strategic value. Section VI separates the requirement from a trustless protocol: it moves trust from a person to a process held among people, and it shows why a record of good conduct cannot stand in for trustworthiness.

The claim that disidentified written process resists capture is a claim about relative cost, not about impossibility. The question this essay puts to any arrangement is therefore not whether it can be captured, since anything can be captured by a party willing to pay enough, but what capture costs against what it buys. The essay identifies the requirement the ephemeral group process is designed to satisfy, and leaves the detailed specification of the process to the book that sets it out.

X. Conclusion

The conclusion is uncomfortable and should be stated without softening.

Every gathering of people creates a power. Every effective sense-making capacity becomes, by its effectiveness, worth acquiring. And any structure in which that capacity is concentrated can be acquired at a single point, after which those inside it will continue to experience themselves as doing the work they set out to do.

This does not mean that people should not gather. The work requires gathering, and the alternative to gathering is that no collective sense-making occurs at the scale at which the decisions that matter are made, which is the present condition and is not survivable if it continues.

What it means is that gathering should be undertaken deliberately, no more than is necessary to construct the infrastructure within which decentralized sense-making can proceed, and with the design of that infrastructure treating capture not as a contingency to be guarded against but as the expected pressure that the design exists to answer.

The measure of such a design is not whether it keeps its secrets well. It is whether the process remains trustworthy when the secrets are lost, when the participants are known, when the funders are hostile, and when every party who would prefer a different answer has had every opportunity to arrange for one. A process that is trustworthy only under favorable conditions is not a process. It is a period of good fortune, and it will end.

XI. What This Essay Adds to the Volume

The volume asks what would be required for the capacity for collective judgment to succeed at the scale at which the decisions that matter are now made. This essay takes up what happens to that capacity once it exists, or begins to exist: it becomes worth capturing in exact proportion to how well it works.

That result is the reason the third property of the governance specification cannot be traded against the others. Satisfying the first property, genuine sense-making, is what creates the pressure on the third, and the convening that could not be held shows the pressure strong enough to prevent work that badly needed doing. Decentralization is therefore a requirement of the design and not a preference about it.

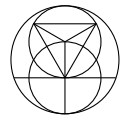
The essay's design principle, a process in which secrets do not matter, answers the volume's recurring question in its own terms: such a process can still be captured, but only participant by participant, without knowing who the participants are, which is expensive in a way that scales badly for the party attempting it. Essay 5 takes the same problem from the scale of an institution to the scale of a single judgment, and the Afterword gathers this essay's conditions into what the third property demands of any design.

KEY CONCEPTS AND FINDINGS IN THIS ESSAY

- **The honeypot problem**: the observation that a collective sense-making capacity becomes an object of capture in exact proportion to its effectiveness, so that success is itself the mechanism of subversion.
- **Gathering-power**: the capacity created by an assembly of persons that no participant possessed individually and that disappears when the assembly disperses.

- **Invisible capture**: capture that succeeds is not detectable from inside the captured structure, since the participants continue to experience themselves as doing the original work; observed capture is therefore a biased sample of actual capture.
- **The ranking-system vulnerability**: a reliability rubric both advertises the sense-making asset and introduces an interest in position that is distinguishable from the interest in truth.
- **The centralized intelligence attractor**: the tendency of any many-to-one confidential communication topology to produce the structure and vulnerabilities of an intelligence agency, without anyone intending it.
- **Decentralization as a structural requirement**: the necessity, not the preference, that a process be trustworthy independently of the trustworthiness of its participants, since participant trustworthiness is exactly what capture acquires.
- **The ephemeral group process**: a way of deliberating in which participants are anonymous to one another and contribute in writing, in a disidentified form, so that a contribution is assessed on its content, and in which groups are constituted for a purpose and dissolved afterward, so that there is no standing body to acquire.
- **Secrets do not matter**: the design principle inverting confidentiality-based protection: rather than keeping secrets better, constructing a disidentified written process in which secrets do not matter, because the loss of secrets does not compromise the result.
- **Data, intelligence, and meaningfulness**: three distinct orientations for a collective project, of which the second tends structurally toward centralization and only the third gives reason to prefer participant understanding over output correctness.
- **The non-fungibility of wisdom**: wisdom, unlike intelligence, cannot be acquired, hired, aggregated, or transferred by instruction, and therefore does not come with a group that is bought.
- **Process-trust**: trust held in a process distributed across six to twelve people, which absorbs the faltering of one participant, in place of trust in a person, which is a single point of failure.

- **Trust hacking**: producing favorable records of past behavior without the corresponding depth, the failure that reputation systems cannot see.



Essay 5

Questions as Instruments of Discernment

On distinguishing real work from its signal,
without requiring disclosure of the solution

ABSTRACT

A party who must decide which of several claimants is genuinely doing difficult work faces a problem that cannot be solved by evaluating the claimants' answers, since the answers require an expertise the evaluator does not have and cannot obtain without becoming a practitioner. This essay develops an alternative: a protocol of questions whose diagnostic power lies not in the answers given but in whether the question is novel to the party asked. A question that a serious practitioner must already have confronted, and that a signaler will not have confronted, discriminates between them without requiring either to disclose their solution. The essay develops this as a form of zero-knowledge verification, argues that upgrading the questions an organization asks itself is the same operation as upgrading its capacity to succeed, describes the inversion by which a solicitation properly becomes an education of the evaluator, examines the framing consequences of leading with risk rather than with benefit, and addresses the arms-race problem created by the fact that a published discriminating question is thereafter available to be prepared for.

I. The Evaluator's Problem

Consider the position of a person who controls significant resources and who wishes to direct them well.

Such a person receives many proposals. Each proposal asserts that the work described is important, that the party proposing it is competent to do it, and that the resources requested are what is needed. The proposals are, in the main, sincere. Some of them are also correct.

The evaluator cannot determine which, and cannot for a structural reason rather than through any failure of diligence.

The world contains too much complexity for any individual to make sense of alone, which is why the proposals concern work the evaluator cannot do personally. And the parties available to advise are themselves participants in the field, with positions of their own to advance, so that advice cannot be taken at face value either. The evaluator therefore cannot do the sense-making and cannot delegate it. This is the *evaluator's problem*.

The usual resolution is to evaluate the proposers rather than the proposals: by pedigree, by prior association, by the confidence with which the case is made, by the quality of the presentation. Every one of these is a measure of signaling capacity. And signaling capacity is the thing that is cheapest to acquire for a party that has nothing else. The position described here is the general form of the third disabling factor named at Essay 1, Section VII. Within a technology-developing organization that factor appears as a gradient running from the engineer who understands to the investor who decides. Between organizations it runs from the practitioner to the party holding the resources, and it is the same gradient (the knowledge-authority gradient): authority increases as understanding decreases.

This is the missing capacity at its smallest scale: one party who must judge and cannot check.

II. Zero-Knowledge Verification

There is a form of solution to this, and it has a precise analogue.

In cryptography, a zero-knowledge proof permits one party to establish to another that they possess some knowledge, without transmitting the knowledge itself (Goldwasser, Micali, and Rackoff, 1989). The verifier ends the exchange convinced, beyond any doubt worth considering, that the prover has what they claimed to have, and no better able to produce it themselves than before the exchange began.

The problem of the previous section has the same shape. The evaluator does not need the solution. The evaluator needs to know, with confidence, that the claimant has one. What is wanted is *zero-knowledge verification of competence*. The correspondence is one of shape. The cryptographic protocol bounds the chance that a prover without the knowledge passes; the conversational protocol carries no such bound, and its strength is the one Section III gives it.

So the question becomes: what is the protocol of questions that an evaluator can put to a claimant, such that the answers (or, as will emerge, the reception of the questions) establish whether the claimant's thinking about the problem is actually good?

This is worth pursuing because it converts an intractable problem into a tractable one. Evaluating a solution requires becoming competent in the domain of the solution. Evaluating whether someone has confronted the questions their domain necessarily poses requires only knowing what those questions are, which is a much smaller thing to learn, and is transferable.

Questions transfer where solutions do not because of what a question carries. A list of the questions that, asked of oneself, lead to a set of insights is a direct transfer of a capability to know, to create knowing, rather than of some presumed knowledge or opinion. The transmission is much more about the relationship between content and context than it is about

content alone. Answers are content; questions are context. And people do not own questions or have identity tied up in questions the same way that they would with directives or statements (*The Ephemeral Group Process in Question*, Part Seven, "Why It Works").

III. The Diagnostic Is Novelty, Not Answer

The diagnostic power of a well-chosen question lies principally in whether the question is *novel* to the party being asked.

Consider a group that proposes to work toward peace on earth. The group has done a great deal: worked with environmental organizations, engaged with non-governmental bodies, built relationships across a wide range of institutions. Ask them what their strategy is for dealing with the military-industrial complex. The question was learned from a colleague, who used it as a decisive test of organizations proposing to work toward peace.

Two things are being tested, and the second matters more than the first.

The first is whether they have a sensible answer. If they do not, they will not succeed, because the obstacle the question names is the principal obstacle in their path and no amount of work elsewhere goes around it.

The second and deeper test is whether the question is new to them. If a group has worked on peace for years and has never asked itself this question, then whatever it has been doing, it has not been working on peace. It has been working on something adjacent that it has been calling peace. The failure is not that it lacks an answer. The failure is that it has not made itself into the kind of organization that would have needed one.

This is why novelty is the sharper instrument. An answer can be prepared. Having confronted a question over a long period, and having been shaped by the confrontation, cannot be prepared, and shows in the man-

ner of the response in ways that are difficult to counterfeit: in the readiness of the reply, in the qualifications attached to it, in the awareness of what the question costs.

How weak the answer is as a signal is shown by stage magic. A person with an embodied working knowledge of physics, mathematics, and psychology, who has watched a great many tricks and understands in general the kinds of methods available, can still identify the specific method behind a specific trick only about ten percent of the time. An untrained observer succeeds far less often, perhaps one percent of the time or less. And one magician watching another, both with full skill in the field, the observer knowing in advance that a trick is about to be attempted, still identifies the specific method only twelve to fifteen percent of the time at best. These rates are estimates from my own experience of stage magic, and what the argument uses is their order of magnitude, a small fraction in every case. A skilled practitioner, whether of stage illusion or of interpersonal deception, can produce answers, including answers to direct questions about motive, that mislead even a careful, knowledgeable, forewarned questioner (*Sense-Making and the Craft of Inquiry*, Chapter 8, "The Limits of Asking", Section III). The protocol therefore reads novelty, and not the answer.

> **The inference the protocol licenses**
>
> The inference is not "this group cannot answer the question, therefore this group is incompetent". It is: "this question is one that anyone actually working toward the claimed goal must confront, because the goal cannot be reached without addressing what the question names; this group has not confronted the question; therefore this group is not working toward that goal, whatever they take themselves to be working on". The inference is about the identity of the work, not about the quality of the workers, and it is valid only for questions that genuinely stand in the necessary path. Constructing questions that meet that condition is the whole of the difficulty.

IV. Upgrading the Questions Upgrades the Organization

A consequence follows immediately and applies reflexively.

To become the kind of organization that can succeed, one must upgrade the quality of the questions one asks.

It is a claim about what an organization is. It states that the questions an organization has confronted constitute the organization's actual shape: that what a group has taken seriously determines what it has become capable of, more directly than its stated aims, its resources, or its personnel. This is the sense in which *questions are organizational shape*. A group that has confronted the hard questions in its path has, in the confronting, acquired the structure required to address them. A group that has not lacks that structure and cannot acquire it by adding resources.

The shape can be read in the texts a party produces. A text produced by sustained first-person reasoning through a coherent worldview tends to produce a question tree with a small number of deep, tightly related root questions, because a real thinker returns, again and again, to the same small set of foundational concerns from different angles. A text assembled from disparate sourced material, even skillfully assembled, tends instead to produce a question tree that is wider and shallower, more like a survey of many separate concerns loosely gathered under one voice (*Sense-Making and the Craft of Inquiry*, Chapter 7, "Indexing by Question, Not by Answer", Section VI). That is a tendency and not a proof, and it is the kind of structural signal that an impression of style cannot provide. An argument is set out in texts of this kind, so the reading applies to an argument as well as to an organization: a reader facing the presentations of Essay 3 can ask which necessary questions lie in the argument's path, and whether the argument meets them as old ground or as news.

The protocol therefore has a second use, turned inward. The same questions one would put to another are the test of one's own architecture: of the strategy, the language, the organizational pattern by which one is

proposing to conduct the work. If one cannot answer, from one's own sense of the matter, whether the work is being done rightly, then one has no warrant to ask anyone else to support it.

Before presenting a case to any evaluator, one should have asked oneself every question that any evaluator might care about, and some harder ones besides, so that the presentation proceeds from a position that has already been tested rather than from one that is being tested for the first time in front of the person one is asking.

V. The Pattern, and the Damage Done by Examples

There is a difficulty in transmitting this method, and it should be stated before any examples are given, because the difficulty concerns the examples.

What matters is the pattern. The pattern is: identify a question that lies necessarily in the path of the claimed work, that cannot be got round, and that a party genuinely engaged in the work will therefore have confronted. Such a question is a *necessary question*.

A necessary question, once met, has three marks that can be observed in how it lands (*The Architecture of Trust*, Chapter 12, "Questions One Cannot Un-Ask", Section I). The first is that it cannot be set aside. A good question can be acknowledged, agreed to be worth someone's looking into, and left while one goes about the day unchanged; a question of this kind does not wait to be taken up. It takes one up, and it is still there in the morning. The second is that it reorganizes what one notices afterward: it does not add an item to what one knows, it re-sorts what one knows. The third is the sharpest. It makes the prior framing visibly inadequate rather than merely incomplete. An incomplete framing invites additions (one more factor, one more safeguard) and survives them. A framing shown to be inadequate cannot be repaired by addition, because the

trouble is not in what it left out but in what it took for granted. A party that meets a necessary question as news therefore cannot answer it with one more safeguard.

Examples of such questions are useful for conveying the pattern and harmful for the same reason. A reader given examples will tend to take the examples as the instrument, will look for those questions or ones resembling them, and will thereby be prevented from generating the questions appropriate to the case at hand, which will be different questions. The examples occupy the place where the reader's own construction should go. The rule is *pattern over example*. The questions are only as good as the understanding of the person who constructs them, since only a person who understands the domain deeply can identify which questions genuinely lie in the path. The method therefore does not solve the evaluator's problem in general; it transfers a specific and hard-won competence from a practitioner to an evaluator, and depends on the practitioner being right.

With that stated, some examples follow: illustrations of shape, to be discarded once the shape is seen.

Of a group proposing to build decentralized sense-making infrastructure: have you considered the possibility of sociopathic capture (the taking over of a group by people who pursue their own advantage without regard for what it costs the others or the group's purpose)? Of a group proposing any structural reform: why would the terms extraction, abstraction, and accumulation be important to distinguish, and what changes if they are not? (These are the three moves, always in that order, by which embodied value is converted into virtual value: extraction takes the thing out of its context, abstraction makes it a member of a class that can be counted and compared, and accumulation reduces it to a number that adds without limit; the pattern is developed in *The Geometry of Man, Machine, and Nature; Care, Community, and the Design of Technology*.) Of a group working on civilizational risk: what is your relation to the analysis of complexity and diminishing marginal return that accounts for historical collapse (Tainter, 1988), and are there accelerant factors that make that

analysis more binding now than it was for the cases from which it was derived? Why would the Fermi paradox (the question why, in so large a universe, no other civilization is seen) be relevant to this conversation at all? Do you distinguish data from intelligence from meaning, and does the distinction do work in your design? What is the relation among the human, the machine, and the natural: what is it supposed to be, and how would you tell?

Each of these has the required property: a party actually doing the work named will have confronted it, and a party signaling the work will meet it as new. Two of them are questions this volume itself must answer, and answers to both are attempted here: the possibility of capture is the subject of Essay 4, and the distinction among data, intelligence, and meaning is drawn at Essay 4, Section VIII. The examples are of no further use. The pattern is the thing.

VI. The Inversion: Solicitation as Education

The protocol produces an inversion in the relation between the party seeking support and the party evaluating it.

In the ordinary case, the seeking party presents a case and the evaluating party assesses it. The seeking party's interest is that the assessment come out favorably, and every element of the presentation is arranged toward that end. The evaluator knows this, discounts accordingly, and is thereby made worse at evaluating, since the discount applies equally to the candid and the practiced.

Under the protocol, the seeking party does something different. They equip the evaluator with the questions by which any claimant in this space can be tested, themselves included.

The transaction becomes an education, and its value to the evaluator does not depend on the outcome. If the evaluator supports the work, well. If they do not, they nevertheless leave the exchange substantially better able to direct their resources well across every subsequent proposal they receive.

A party who moves significant resources and who is given a genuine instrument of discrimination may thereafter direct those resources toward better work than they otherwise would have, in amounts far exceeding anything the educating party would have received. The education may therefore accomplish more good than the funding would have, which means the exchange is worth conducting on its own terms and not only as an instrument of solicitation.

There is also a coherence argument, and it is not decorative. A project whose purpose is to build capacity for collective discernment should not fund itself by methods that degrade the discernment of those it approaches. Conducting the *solicitation as education* is the only form of solicitation consistent with the thing being solicited for.

VII. Leading with Risk

A structural point about framing belongs here, because it determines what kind of exchange is possible at all.

The conventional presentation leads with benefit: here is the value that will be created, here is the cost of creating it, and risk is either unmentioned or reduced to the observation that the project might not work, which everyone already assumes. This is the shape of an investment case, and it is appropriate to investment, where a return is expected and the question is its magnitude and probability.

Work on systemic risk inverts this. The case begins with something of altruistic value: something worth preserving that is not the property of any party. It proceeds to the risk to that value, which is the substance of

the case. It addresses cost only afterward. And it may not mention benefit at all, because there is no benefit to the party being asked, and to manufacture one would be false. This is *risk-first framing*.

This is a different act, and the difficulty of performing it should not be underestimated. It asks for behavior that runs against a very strong grain: for a party to act altruistically on behalf of something from which they will extract nothing. The altruism our evolutionary inheritance prepared us for is altruism toward the group one belongs to. What is being asked is that it be extended to the species, and beyond that to the ecosystem of which the species is part: a scale of concern that nothing in our formation prepared us to feel, and which the requirement set out at Essay 1, Section VIII, nevertheless presupposes, if the technological condition is to be survived.

> **A colleague's observation: insurance**
>
> A colleague observed that a developed model already exists for a practice that leads with risk: insurance. The insurance seller opens with risk, proceeds to cost, and does not lead with benefit, because the benefit is contingent on an event the buyer hopes will not occur.
>
> The observation is valuable. It suggests that the framing difficulty identified above is not without precedent, and that the substantial body of practice developed in that industry for making risk legible and actionable to parties who would prefer not to think about it is worth examining directly. The rhetorical problem is the same problem, though the analogy is imperfect: insurance transfers a risk that is priced and personal, whereas the risks at issue here are unpriced and shared.

VIII. The Arms Race and the Problem of Distribution

The protocol has a vulnerability that must be addressed in its design rather than in its application.

A discriminating question works because a signaling party has not confronted it. Once the question is widely known, signaling parties will prepare for it. The question then no longer discriminates, and has become merely one more thing to have a prepared answer to.

This is an arms race, and it has an unusual feature that makes it worse than most. When the instrument is defeated, the party who built it does not learn that it has been defeated, because a prepared answer is not distinguishable from a genuine one by the party who is fooled by it. The second iteration of the instrument therefore cannot be issued, because no one knows that it is needed. This is the *undetectable defeat*.

The practical implication concerns distribution. The instrument should reach evaluators in preference to claimants, and should reach them with as much lead time as can be arranged. This is a matter of sequence, not of secrecy: the instrument will propagate, and should. An evaluator who has the questions first gets some period of genuine discrimination. An instrument released to everyone at once provides none.

A reader may ask whether this essay, by printing its examples, has spoiled them. The examples in Section V are offered as illustrations of shape and are spent by being printed; the instrument is the pattern, and a reader who has the pattern can make questions no one has yet prepared for.

The arms race, if it must be run, is a race in which the signaling parties are obliged to prepare answers to progressively harder questions. That is an improvement on the situation in which no hard questions are asked, and it is not nothing: a party who has prepared serious answers to a long enough series of necessary questions has, at some point, done a nontrivial portion of the work.

IX. Triangulating the Knowledge State

A final property of the method concerns what the asking party learns.

Questions put to another permit the questioner to triangulate the knowledge state of the person asked, and this operates in both directions.

When a question lands, so that the party asked follows it with questions of their own that integrate what the first question implied, one knows that it was understood. The follow-up question is the receipt; this is *triangulation by follow-up*. It cannot be produced by someone who did not receive the content, because its construction depends on the content.

When questions do not land, one learns this early, which is itself valuable. An exchange that is not going to be productive is better identified in its first minutes than in its last, and the appropriate response is to say so and to close it, which is a courtesy to both parties and not a failure.

It should be evident by now that the method is doing several things at once: discriminating among claimants, educating the evaluator, testing one's own architecture, and reading the state of the exchange as it proceeds. This is by design. Nothing worth doing is done for only one reason.

X. Two Questions Others Have Raised

Two questions, raised by others, are serious.

That it is insincere. The concern raised is that a party who is in fact asking for resources while presenting the exchange as an education is concealing their interest, and that the concealment is a form of the very signaling the method claims to detect.

The answer is that nothing need be concealed, and that the position is stated most strongly when it is stated completely. The party using the method says two things, both true: "We are asking you for resources. And we are trying to ensure that parties who cannot do this work do not receive resources on the strength of resembling those who can, unless their work is genuinely more worthwhile than ours, in which case the outcome is correct and we have no complaint". Both halves are said aloud. The

method does not require that the interest be hidden; it requires that the interest not be the only thing operating, which is a different and satisfiable condition.

That it forecloses alternatives. The concern raised is that an evaluator equipped with questions that only one party can answer has been maneuvered into a position where they can fund only that party, and that this is a capture of the evaluator rather than an education of them.

The answer is that this outcome should not be sought and should not be accentuated. It would be far better if a number of parties were doing worthwhile work in this space, and there are parties whose work is good and moving in the right direction. Where such work falls short, it tends to fall short in a specific and identifiable way: in not having distinguished necessity from sufficiency, and in not having established comprehensiveness, that is, in having found things that are needed without having established that what has been found is enough. The distinction among necessity, sufficiency, and comprehensiveness is developed in *An Immanent Metaphysics*. That is a real deficiency and it is also a repairable one, and pointing at it is more useful to everyone than being the only party who can pass a test.

XI. Summary

Section II shows why questions transfer where solutions do not: answers are content, questions are context, and no one's identity is tied to a question. The central claim, that novelty of a necessary question is diagnostic of whether the claimed work is being done, is valid for questions that genuinely lie in the necessary path, and the validity depends entirely on that condition. Section III bounds how weak a prepared answer is as a signal, by the rates at which even skilled and forewarned observers identify the method of a trick.

The claim of Section IV, that confronted questions constitute organizational capability, rests on the observation that a group acquires, in confronting the hard questions in its path, the structure required to address them, and that the structure cannot be added by adding resources. Section IV also shows that the shape of a party's questions can be read in its texts, as a question tree, and that the reading applies to arguments as well as organizations.

The essay supplies the pattern by which a question set is constructed, and argues in Section V that supplying a set would be counterproductive. Section V gives the three marks of a necessary question, the last of which is that it makes the prior framing inadequate, so that it cannot be answered with one more safeguard. In the terms the volume applies to every arrangement, the price of capturing a necessary question is the cost of preparing an answer to it, and each harder question raises that price, until preparing the answers is a nontrivial portion of the work itself (Section VIII).

XII. Conclusion

The problem this essay addresses (which is not, at bottom, a problem of philanthropy or of funding) is a specific instance of the general problem identified at Essay 1, Sections VII and VIII: that the capacity for collective discernment is the thing that is missing, and that its absence is what permits the signal of work to be substituted for the work.

Questions are a particularly efficient instrument against this substitution. An answer can be acquired. A presentation can be commissioned. Confidence can be trained. But the shape that a hard question leaves in an organization that has lived with it cannot be acquired, because it is not a possession. It is what the organization has become.

The effect one is seeking, when a question is put well, is specific and recognizable: that the person receiving it had not known such a question could be asked, and, knowing now that it can, thinks differently about every claim that reaches them afterward.

That effect is a small, local, durable increase in the discernment available in the world, and not a technique of persuasion. The argument of these essays is that such increases, made deliberately and at sufficient scale, are among the very few things that bear on the outcome.

XIII. What This Essay Adds to the Volume

This essay takes the volume's question, under what conditions the capacity for collective judgment fails and what would be required for it to succeed, to its smallest scale: a single party who must judge which of several claimants is doing real work, and who cannot do the work or check it.

At that scale the missing capacity has a precise shape. The evaluator's problem is the knowledge-authority gradient of Essay 1 seen between organizations: the party who decides is the party who understands least. Where that gradient holds, and no capacity for collective discernment stands beside it, the signal of work is substituted for the work. Questions that lie necessarily in the path of the claimed work are an instrument that can be built now, by anyone who knows a field well enough to name them, and the Afterword, Section VII, names this as what can be done before anything else is built.

The essay also meets the volume's recurring problem of capture. A published question becomes something to prepare an answer for, and the defeat is invisible to the party who asked it, which is the invisible capture of Essay 4 at the scale of one exchange. The essay answers it as the volume answers every such case: not with a question that cannot be prepared for, since there is none, but with a sequence of harder questions, reaching evaluators before claimants, so that the price of preparing an answer rises until it is a portion of the work itself.

KEY CONCEPTS AND FINDINGS IN THIS ESSAY

- **The evaluator's problem**: the structural inability of a resource-holder to distinguish real work from its signal, since the sense-making cannot be done personally and cannot be delegated to interested parties.
- **Zero-knowledge verification of competence**: a protocol by which an evaluator establishes that a claimant possesses an adequate solution, without the solution being disclosed and without the evaluator becoming competent in the domain.
- **Novelty as diagnostic**: the principle that a well-chosen question discriminates chiefly by whether it is new to the party asked, since answers can be prepared but the formation produced by long confrontation with a question cannot.
- **The necessary question**: a question lying unavoidably in the path of the claimed work, such that a party who has not confronted it is, whatever they believe, not doing that work.
- **Questions are organizational shape**: the claim that what an organization has taken seriously determines what it has become capable of, more directly than its aims, resources, or personnel.
- **Pattern over example**: the observation that supplying example questions displaces the reader's own construction of case-appropriate questions, so that examples must be given sparingly and then discarded.
- **Solicitation as education**: the inversion by which a request for support is conducted so as to leave the evaluator better able to assess all claimants, including the one asking, with value delivered independently of the outcome.
- **Risk-first framing**: the presentation order appropriate to systemic risk (altruistic value, then risk, then cost, benefit largely unmentioned), inverting the conventional benefit-first investment case.

- **The undetectable defeat**: the arms-race property whereby a discriminating question, once prepared for, ceases to discriminate without the instrument's author learning that it has failed, so that no second iteration is issued.
- **Triangulation by follow-up**: the reading of an interlocutor's knowledge state from whether their subsequent questions integrate what the prior question implied; the follow-up question serves as receipt of understanding.
- **Answers are content; questions are context**: the reason a set of questions transfers a capability where a set of answers transfers only a result.
- **The question tree**: sustained first-person reasoning tends to yield a few deep, related root questions; assembled material tends to yield a wider, shallower tree.

On What Would Have To Be Built

The four properties of the specification,
and what each demands of a design

I. The Standing of This Statement

Essay 1 gives a specification. Any system adequate to the class of problems described in this volume must be genuinely effective at collective sense-making, choice-making, and the enablement of agency; must be fractal and scalable; must be robustly anti-corruptible; and must genuinely resolve the principal-agent problem.

What follows takes that specification seriously enough to say what each of its four properties demands of a design, and why a design that satisfies three of the four is taken through the one it lacks.

This is a statement of requirement. It is not a design. Its one argument is the case analysis of Section VI, which shows that a design meeting three of the four properties is taken through the one it lacks. A requirement is useful in advance of a design for the same reason a specification is useful in advance of a building: it settles what would count as success, and it disqualifies, early and cheaply, work that could not succeed however well it were executed. Most proposals in this area can be assessed against these four properties in an afternoon, and most do not survive the assessment. That is the whole of what a requirement is for.

II. The First Property: Sense-Making, Choice-Making, and Agency

The first property is that the system actually work, and it has three parts that must be present together.

Sense-making is the formation of shared understanding. A design supplies it by arranging matters so that a contribution can be assessed on its content, by parties who have enough common ground to assess it, with enough time for an assessment to be made. Each of these three is a real constraint on a design. Content-assessment requires that the source of a contribution not be the dominant signal about it. Common ground requires that participants share some working vocabulary and some framing of the question, which is itself something a design must produce rather than assume. And time requires that the process not be conducted on the schedule of whatever decision occasioned it, which is nearly always shorter than understanding takes.

Choice-making is the arrival at a decision that the participants recognize as theirs. A design supplies it by having a way of closing: some procedure by which a group that has understood a matter moves from understanding to commitment, without the closing procedure undoing the work that preceded it. Voting is the familiar closing procedure and it is a poor one for this purpose, because it discards the understanding and retains only the preferences, which is the loss the whole process existed to prevent.

Agency is the capacity for the decision to take effect. A design supplies it by arranging that something in the world depends on the outcome: a resource that moves, an authority that binds, a coordination that holds because the parties to it have reason to hold to it.

The three are required together because any two without the third are inert. Understanding without choice is a conversation. Choice without agency is an opinion. And agency without understanding is the present condition, at scale, and is what the volume is about.

III. The Second Property: Fractal Scalability

The second property is that the system operate at world scale by the principles on which it operates among a few, rather than by substituting a weaker mechanism above some threshold.

The demand this places on a design follows from a fact established in Essay 1, Section VII. Methods by which a group thinks together work at approximately the scale at which each participant can hold a working model of every other. That bound (the Dunbar limit on sense-making) is not a defect of present methods to be engineered away. It is a property of the participants.

A design that respects the bound and still reaches world scale must therefore scale by composition rather than by enlargement. The unit remains small. What grows is the number of units and the structure of the relations among them. And this imposes a specific and demanding condition: whatever passes between units must be the kind of thing a receiving unit can assess on its content, by the same means it assesses a contribution from one of its own members.

That condition rules out a great deal. It rules out summary by a representative, because a representative's summary can be assessed only by trusting the representative, which reintroduces at the boundary exactly the dependence the design was arranged to avoid. It rules out any inter-unit protocol in which position, precedence, or standing carries information, for the same reason. What it admits is inspectable output: the reasoning, and not only the conclusion, in a form that a party who was not present at its making can take up and evaluate. The requirement of fractal scalability is thus, in practice, a requirement about the medium.

One such medium is specified in *Governance Without Capture*, "The Ephemeral Group Process", Section V, as an artifact of four sections. The first states the question: not the question as originally posed, but the question as the group came to understand it after their conversation. The second identifies what matters: not a set of conclusions but a map of the terrain, what the group determined to be relevant, significant, and worthy of attention. The third explains why it matters: the reasoning that connects the identified considerations to the question and to the broader concerns of the community. The fourth lists further questions and recommendations: the things the group realized they did not know and that subsequent groups should address. The community's intelligence is then never concentrated in any single person or group; no one needs to summarize the state of the conversation for the whole community, because the artifacts do this work. Whether a receiving group can judge such an artifact on its content is the question Section VII leaves open.

Essay 3, Section VI, gives the reason this artifact opens on a question and closes on questions. A group that has not met a problem at the depth its answer addresses receives that answer as a reply to a question it was not asking, and further clarity about the answer does not supply the question. The council's summary of reasons, in Essay 1, Section VII, was an answer of that kind. The artifact sends the question first, as the sending group came to understand it, so that the receiving group can meet the problem before it meets the reasoning; and a raised question, as Essay 3, Section VIII, says, does not depend on the person who raised it. Essay 5, Section II, gives the same reason in four words: answers are content; questions are context. How far meeting a question in writing reaches the depth that hours in a room reached is the question Section VII leaves open.

IV. The Third Property: Robust Anti-Corruptibility

The third property is that the attempt to capture the system be ineffective, and the word doing the work is *ineffective* rather than *forbidden*.

Prohibition is not available as a mechanism here. A prohibition requires enforcement, enforcement requires an enforcing party, and an enforcing party is a thing that can be acquired. Every design that answers capture with a rule has moved the point of capture and has not removed it. A rule operates within the system that power concentration controls: rules are made by legislative processes that powerful actors can influence, enforced by agencies that powerful actors can capture, and adjudicated by courts that powerful actors can access more effectively than others. Each new rule provokes adaptive circumvention, which provokes further rules; when the architecture permits concentration, no set of rules can sustainably prevent it, because the concentrated power will eventually modify the rules (*Governance Beyond Arrow*, "The Power Prevention Process", Section II). The remedy therefore lies in the design of the system, and not in the rules within it.

What a design can do is arrange that acquisition buys less than it costs. Essay 4 gives three conditions that bear directly on this. A structure with no center has no single point at which influence over the whole can be obtained. A process in which participation is disidentified offers no position in a ranking to compete for, and therefore admits no interest in one's position distinguishable from the interest in the truth. And a process whose contributions are written and inspectable can be checked by parties who were not present, so that the assessment of the work does not depend on the standing of those who did it.

A practical test follows, asked of every role in a structure: if a party wished to purchase an outcome, whom would they approach? Any answer that names a position is a defect, whether or not anybody has yet approached it. Three moves answer that test. Separation breaks discretion into parts held by different parties, so that no single approach suffices and any purchase requires a coalition: slower, costlier, and far more likely to leak. Non-persistence makes a role held briefly worth less to capture than one held indefinitely, because the relationships through which influence travels take time to build and must be rebuilt each time the seat turns over. Selection that cannot be lobbied draws the occupant of a role that

must exist by lot from the affected population, so that there is no selector to approach (*Governance, Ownership, and the Commons*, "Anti-Corruption by Design", Section IV). A reader will point out that a distributed system is less efficient than a concentrated one. It is, and the efficiency of the concentrated form is precisely the asset a captor acquires.

Essay 2 states the general condition of which these three are instances. An arrangement resists imitation and acquisition in proportion to what imitation and acquisition cost, and the design aim is parity (capture-cost parity, Essay 2, Section VIII): that taking the system should cost what the system is worth. Parity is the threshold at which capture stops paying; every increase in price beyond it is further protection. This is a weaker aim than incorruptibility and it is the strongest aim available, because nothing is beyond acquisition by a party willing to pay enough. A design that raises the price until the price exceeds the prize has done everything a design can do.

The same aim appears wherever my other books treat capture. A guardian constituted as the whole community can be corrupted only by corrupting the community, which is to say by an operation visible to every member of it; such a threshold does not make capture impossible, it makes capture expensive, slow, and observable, which is the most that any structural safeguard achieves (*Covenant and Commons*, "The Given and the Taken", Section V). And the record of the least corrupt countries shows the difference between *ineffective* and *forbidden*. They are not those with the most severe penalties. They are those in which enough people understand enough of what is happening that quiet arrangements do not stay quiet. Where such accounting exists, there is no niche for defection, not because defection has become impossible, but because it has stopped being invisible, and invisibility was the resource it ran on (*What Holds a Civilization*, "Three Relationships", Section IX). Written and inspectable contributions remove that resource by design.

A reader will ask who pays for the external veto body of Essay 2, Section VIII, and what keeps it from being bought. The conditions are already stated across the volume, and they can be gathered in one place.

Who pays matters less than how the payment is made: the body is funded independently of the outcome, so that no payment to it depends on what it decides. It is composed of parties who lose nothing by halting a deployment and gain nothing by permitting one, beyond what every affected party shares. It is empowered so that a halt takes effect without the assent of the organization halted. Its members are inside the set of parties affected, cannot exit that set, and meet the cost of a wrong decision on a timescale comparable to the one on which they decide, which are the conditions of the fourth property (Section V). And its work is done in written contributions that parties who were not present can inspect, with no center at which influence over the whole can be bought, which are the conditions of Essay 4, Section VII. Under these conditions a party who wants the body's decision must buy its members one by one, and each purchase must outbid what that member loses by a wrong decision; that is the price the design raises toward parity.

V. The Fourth Property: Resolution of the Principal-Agent Problem

The fourth property is that the interests of those who decide be aligned with the whole of what their decisions affect, and not merely with what is locally or temporally extractable.

The demand this places on a design is the hardest of the four to state as a mechanism, because the obvious approaches are all approaches to selecting agents rather than to arranging matters. Choosing decision-makers of good character, binding them by oath or by fiduciary duty, subjecting them to review: each of these is a method of obtaining a trustworthy agent, and the specification asks for something else. It asks that the arrangement not require one.

An arrangement satisfies the property where the party who decides is inside the set of parties affected, where they cannot exit that set, and where the cost of a wrong decision reaches them on a timescale compar-

able to the one on which they decide. Each clause is load-bearing. A decision-maker who can leave is not aligned with what they decide about, and the ability to leave is what capital is. A cost that arrives after the decision-maker's horizon is not a cost to that decision-maker, which is the substance of the observation in Essay 1, Section VII about fund return schedules and technology deployment lifetimes. Cannot exit here means cannot escape the consequences of the decision; it does not mean that membership is compulsory.

The clause about leaving is argued in *Covenant and Commons*, "The Given and the Taken", Section VI. Money can be moved around the world much more easily than any physical thing can. The party operating in the virtual medium can reposition in seconds, at negligible cost, across every jurisdiction on the planet. The party tending the physical can reposition over seasons, at enormous cost, and often cannot reposition at all: a farm cannot relocate, a watershed cannot relocate, a community of one hundred and fifty that has lived somewhere for four generations cannot relocate without ceasing to be itself. Where one party can move and the other cannot, the one who can move sets the terms. This is not a claim about anyone's intentions; it would hold just as strongly if every participant were scrupulous. The clause about membership is stated as a rule for economic community in *The Architecture of Trust*, "Three Value Bases and the Structure of Economic Community", Section III: the membership of the community and the membership of the economic entity must be exactly the same. Where the two memberships differ, some who are affected by the entity's choices have no standing in them, or some who hold standing bear none of the consequences, and each gap is an opening through which the interests of the deciding body diverge from those of the affected one. Identity of membership closes the gap by construction rather than by policy, which is the only way such gaps stay closed.

The clause about time explains why the familiar remedy of removing those who chose badly does not reach this class of problem. The error-correction of democracy is retrospective: it works by letting a poor choice display its consequences, and then removing the people who made it. That

requires that the consequence be survivable and arrive in time to inform the removal. On existential risk the feedback arrives after the correction has become useless (*The Architecture of Trust*, "Meritocracy, Democracy, Consensus", Section IV).

A reader will ask how the clause about time can be met for a harm that arrives after everyone who decided has died. A timescale comparable to the one on which the party decides can be reached from either end. The cost can be brought forward, by liability or an insurer's price (Essay 2, Section VIII), as far as the harm can be made attributable. And the horizon of the deciding party can be lengthened, since the deciding party need not be a person: a community that cannot relocate and expects its grandchildren to live where it lives bears a slow cost as its own. Where neither end can be reached, the decision is one that no arrangement can yet be trusted to make, and that is a reason not to make it.

A reader will also ask who speaks for those who cannot sit at the table: the unborn, and the rest of life. Both belong to the whole set of parties a decision affects, which Essay 1, Section VIII, names as the principal. The arrangement gives them no representative, since a representative would be an agent of the kind this property asks the arrangement not to need. It gives them deciding parties whose own conditions of life are bound up with theirs: a community that cannot relocate shares the watershed, the soil and the living systems that its grandchildren and the rest of life depend on, and, as Essay 1 says in its section on what technology is, those conditions are part of what a human being is.

This is the property that the other three cannot supply. A system may form excellent understanding, scale by composition, and resist acquisition from outside, and still be operated by parties whose exposure to the consequences of its outputs ends before the consequences arrive.

VI. Why Three Are Not Enough

The claim that a design meeting three of the four properties is taken through the one it lacks can be shown by taking the four cases in turn. Each has a characteristic mode of failure, and the modes are different from one another, which is itself informative.

A system that scales, resists capture, and aligns its decision-makers, but does not produce understanding, is not captured. It is inert, and inertness is not stable in a competitive environment. Parties needing answers will obtain them from some faster source, and the faster source will have none of the other three properties. The failure mode of the first omission is displacement.

A system that works, resists capture, and aligns its decision-makers, but does not scale, produces good judgment in a small room and must then transmit it to the scale at which decisions are made. The transmission is performed by something, and the something is a channel that possesses none of the properties. The failure mode of the second omission is capture at the boundary, and it is the most common failure in practice, because a small effective group experiences its own integrity as sufficient and does not attend to what happens to its output downstream.

A system that works, scales, and aligns its decision-makers, but does not resist capture, is the case Essay 4 develops. Its effectiveness is what makes it worth acquiring, its success is therefore the mechanism of its subversion, and the acquisition is not visible from inside (invisible capture, Essay 4, Section III), because a captured process continues to experience itself as doing its own work. The failure mode of the third omission is acquisition.

A system that works, scales, and resists external capture, but does not resolve the principal-agent problem, requires no adversary at all. The parties within it direct it toward what they can extract, and they do so while the process runs correctly by every internal measure. The failure

mode of the fourth omission is drift, and it is the one for which no external observer can supply a correction, since the party observed is the party supplying the observation. That is the result of Essay 2, Section VIII, arriving here from the other direction.

The four modes are displacement, boundary capture, acquisition, and drift. No two of them are the same failure, and no design that has answered three of them has thereby answered the fourth. This is the sense in which the specification does not admit of partial satisfaction.

Property missing	**Failure mode**	**What takes the system**
Sense-making, choice-making, and agency	Displacement	A faster source of answers, which has none of the other three properties
Fractal scalability	Boundary capture	The channel that carries the small group's judgment to the scale at which decisions are made
Robust anti-corruptibility	Acquisition	A party for whom the working capacity is worth acquiring
Resolution of the principal-agent problem	Drift	The parties within, directing the system toward what they can extract

A reader will ask whether artificial intelligence could itself be the instrument that lets people think together at scale. The four failure modes answer the question. A system that answers questions for a group faster than the group can understand them is the faster source of answers of the first omission, and its arrival is displacement. A system that carries the judgment of small groups to the scale of decision, by summarizing, ranking, or translating it, is the channel at the boundary; unless what it passes on can be assessed on its content by the group that receives it, its use is boundary capture. A system owned by an interested party, through which the questions and conclusions of many people flow to one place, has the many-to-one confidential topology of Essay 4, Section IV, and the more

useful it is, the more it is worth directing (the honeypot problem, Essay 4, Section II). Such a system can serve the capacity only as a tool whose outputs groups of people assess on their content, as they would assess a contribution from one of their own members. It cannot be the place where the shared understanding is formed, because sense-making is the formation of a shared understanding among the parties a decision affects.

VII. One Line of Work

Work exists that is addressed to parts of this requirement.

The ephemeral group process, described in Essay 4, Section VII, and specified in *Governance Beyond Arrow; An Immanent Architecture for Existential Choice*, is one line of work addressed to the first and third properties. Its participants are anonymous to one another and contribute in writing, in a disidentified form, so that a contribution is assessed on its content; and its groups are constituted for a purpose and dissolved afterward, so that there is no standing body to acquire. Both features are answers to specific clauses of the specification.

A reader may ask how people who are anonymous to one another can do what small-group thinking requires, which is that each participant hold a working model of every other. Anonymity hides a participant's name and standing. It does not hide how that participant reasons, and a working model of how someone thinks can be built from what that person writes.

It is one line of work and it is not a solution. It does not by itself establish that the composition condition of the second property is met, which is a question about what passes between groups and how a receiving group evaluates it. It does not address the fourth property at all. And the claim that a disidentified written process resists acquisition is a claim about relative cost rather than about impossibility, as Essay 4, Section VII, says plainly.

My other books offer two partial answers to the question of what passes between groups. One is the four-part artifact of Section III, which carries the reframed question, what matters, why it matters, and the questions that remain. The other is that members are remixed: within each group a shared vocabulary and a provisional agreement develop, and across the groups each member must explain the group's tentative conclusions to people who do not share its assumptions, so that no group's view hardens into orthodoxy. A remixed member holds no standing for the group that member came from and decides nothing on its behalf; what that member says is judged on its content by people who do not share its assumptions, which is why remixing is not the overlapping of memberships that Essay 1 reports as failing. The work of many temporary groups does not become one choice by being counted or summed into a single position. It becomes an understanding shared across the community, and because no person is the standing agent of others, the principal-agent problem, in its ordinary sense, has nothing to act on (*The Architecture of Group Governance*, Chapter 10, "The Ephemeral Group Process", Sections I and II). That account is given for communities of hundreds; whether it reaches the scale of a technology whose effects cross every boundary is the part that remains to be shown.

Other bodies that exist now come nearer to parts of the requirement, and a reader in governance will ask about the nearest of them: citizens' assemblies and other deliberative panels chosen by lot. In the Irish Citizens' Assembly of 2016 to 2018, ninety-nine citizens chosen at random, with a chair, heard evidence, deliberated, and voted on recommendations, among them a recommendation on abortion that a national referendum later took up (Farrell, Suiter, and Harris, 2019); panels of this kind are now in wide use (OECD, 2020). Taken against the four properties, such a panel meets much of the first within one room: its members form an understanding none of them brought, at about the scale of one Dunbar number. Its closing procedure is a vote, however, and its agency is borrowed, since its recommendations take effect only when a legislature or an electorate adopts them. It does not meet the second property: one panel stands

in for a nation by sampling and not by composition, and its judgment reaches the scale of decision through a report and a legislature, which is the channel at the boundary. It meets part of the third, since members chosen by lot and dismissed afterward leave no standing body to acquire; but the organizers who set the question, choose the witnesses, and write the briefing are a standing center, and capture can operate there, on the inputs. And it leaves the fourth property to the legislators who decide, whose exposure to the consequences of a decision ends with their term of office.

Standards committees, such as the working groups of the Internet Engineering Task Force, work in writing on drafts that anyone can read, and close by rough consensus rather than by vote (Resnick, 2014). Their written and inspectable contributions meet the condition of the second property more nearly than any panel does. Their participants commonly take part on behalf of firms whose products the standards govern, so the interests that decide can be the firms' interests, and nothing in the arrangement aligns them with the whole of what a standard affects; the fourth property is not met.

The commons that Elinor Ostrom studied come nearest of all, at their own scale (Ostrom, 1990). Users who watch and answer to one another form the understanding; nesting composes small units into larger ones; the users who decide are inside the set of parties affected and cannot easily leave it; and capture is resisted by monitoring and graduated sanctions that the users themselves carry out, which is enforcement, held against capture because the enforcers are the parties affected. None of these arrangements operates at the scale of a technology whose effects cross every boundary, which is the scale at which, as Essay 1, Section VII, says, composition has not yet been built.

Stated plainly: the specification is at present satisfied by nothing, it is approached in parts by a small number of efforts, and the design has yet to be made.

A reader who asks what can be done now has the most concrete answer in Essay 5: learn the necessary questions of one's own field, ask them of oneself first, and give them to those who allocate resources before they reach those who seek resources (Essay 5, Sections IV and VIII). Each such question, put well, is a small and durable increase in the discernment available in the world, and anyone who knows a field well enough to name the questions that lie in its path can begin it at once, before anything is built.

That is not a discouraging conclusion, and it should not be read as one. A requirement stated exactly is the condition under which work can be evaluated, and the absence of such a statement is why so much effort in this area has been spent on things that could not have worked for reasons that were available in advance. The four properties are demanding. They are also finite, statable, and checkable, which is more than can be said for most of what is asked of an institution. What would have to be built is a thing that can be described. The description is the beginning of the work, and it is offered here for that use.

VIII. What This Afterword Adds to the Volume

The volume converges on one claim: the missing thing is the capacity to think together at scale. The essays show the conditions under which that capacity fails, and each shows the loss of one or another of the four properties Essay 1 requires. This Afterword turns the same four properties into what a design must supply, and shows, by the four failure modes, that a design supplying three of them is taken through the one it lacks.

It gathers the volume's results in that form. The conditions of Essay 4 on capture, and the condition of Essay 2 that capture cost what it buys, become the demand of the third property. The account of the decision-maker who can leave, from Essay 1 and from the companion essay on self-regulation, becomes the demand of the fourth. And the measure it offers for existing work, citizens' assemblies, standards committees, and Os-

trom's commons among it, shows where the requirement is met in part and where nothing yet meets it: at the scale of a technology whose effects cross every boundary.

It also gathers what my other books supply for each property: a four-part written artifact as one medium for the second, and the remixing of members as a second way of passing reasoning between groups; for the third, the test of whom a buyer of outcomes would approach, and the finding that the least corrupt countries are those in which quiet arrangements do not stay quiet; for the fourth, the mobility differential that lets capital set the terms, the rule that those who decide and those affected be the same membership, and the reason retrospective correction fails on existential risk.

Bibliography

Every work the essays cite or draw on, in one list. In the text, my own books are cited by title, and other works by author and year. Published subtitles are joined to their titles with a semicolon, following the house style.

Akerlof, George A. "The Market for 'Lemons'; Quality Uncertainty and the Market Mechanism". *Quarterly Journal of Economics* 84, no. 3 (1970): 488 to 500.

Armstrong, Stuart, Nick Bostrom, and Carl Shulman. "Racing to the Precipice; A Model of Artificial Intelligence Development". *AI and Society* 31, no. 2 (2016): 201 to 206.

Arrow, Kenneth J. *Social Choice and Individual Values*. New York: Wiley, 1951.

Ashby, W. Ross. *An Introduction to Cybernetics*. London: Chapman and Hall, 1956.

Beckstead, Nick. "On the Overwhelming Importance of Shaping the Far Future". PhD dissertation, Rutgers University, 2013.

Bell, John S. "On the Einstein Podolsky Rosen Paradox". *Physics* 1, no. 3 (1964): 195 to 200.

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

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, and others. *The Malicious Use of Artificial Intelligence; Forecasting, Prevention, and Mitigation*. 2018. arXiv:1802.07228.

Bueno de Mesquita, Bruce, Alastair Smith, Randolph M. Siverson, and James D. Morrow. *The Logic of Political Survival*. Cambridge, Massachusetts: MIT Press, 2003.

Bueno de Mesquita, Bruce, and Alastair Smith. *The Dictator's Handbook; Why Bad Behavior Is Almost Always Good Politics*. New York: PublicAffairs, 2011.

Business Post. "How Steorn Pulled the Plug on Its Decade-Long Promise of Free Energy". *The Sunday Business Post* (Dublin), November 2016. https://www.businesspost.ie/business/power-off-steorn-finally-energy-e23-million-funding-370313.

Carse, James P. *Finite and Infinite Games*. New York: Free Press, 1986.

Cremer, Carla Zoe, and Luke Kemp. "Democratising Risk; In Search of a Methodology to Study Existential Risk". 2021. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3995225. The text is also available as arXiv:2201.11214, and from Open Book Publishers.

Dawson, John W., Jr. "The Reception of Gödel's Incompleteness Theorems". *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1984, vol. 2 (1985): 253 to 271.

Dunbar, R. I. M. "Neocortex Size as a Constraint on Group Size in Primates". *Journal of Human Evolution* 22, no. 6 (1992): 469 to 493.

Farrell, David M., Jane Suiter, and Clodagh Harris. "'Systematizing' Constitutional Deliberation; The 2016 to 18 Citizens' Assembly in Ireland". *Irish Political Studies* 34, no. 1 (2019): 113 to 123. https://doi.org/10.1080/07907184.2018.1534832.

Freedman, Stuart J., and John F. Clauser. "Experimental Test of Local Hidden-Variable Theories". *Physical Review Letters* 28 (1972): 938.

Gödel, Kurt. "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I". *Monatshefte für Mathematik und Physik* 38 (1931): 173 to 198.

Goldwasser, Shafi, Silvio Micali, and Charles Rackoff. "The Knowledge Complexity of Interactive Proof Systems". *SIAM Journal on Computing* 18, no. 1 (1989): 186 to 208.

Goodhart, Charles A. E. “Problems of Monetary Management; The U.K. Experience”. In *Papers in Monetary Economics*, vol. 1. Sydney: Reserve Bank of Australia, 1976. Presented at the Conference in Monetary Economics, Sydney, July 1975.

Grey, CGP. “Rules for Rulers”. Video, 2016. https://www.youtube.com/watch?v=rStL7niR7gs.

Hardin, Garrett. “The Tragedy of the Commons”. *Science* 162, no. 3859 (1968): 1243 to 1248.

Hensen, B., and others. “Loophole-free Bell Inequality Violation Using Electron Spins Separated by 1.3 Kilometres”. *Nature* 526 (2015): 682 to 686.

Hickel, Jason. *Less Is More; How Degrowth Will Save the World*. London: William Heinemann, 2020.

Jensen, Michael C., and William H. Meckling. “Theory of the Firm; Managerial Behavior, Agency Costs and Ownership Structure”. *Journal of Financial Economics* 3, no. 4 (1976): 305 to 360.

Kunda, Ziva. “The Case for Motivated Reasoning”. *Psychological Bulletin* 108, no. 3 (1990): 480 to 498.

Landry, Forrest. *A Group Is Made by a Question; The Practice of Facilitation and the Ephemeral Group Process*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-329-0. https://doi.org/10.67291/BJSY1336. “Why the Group Is Ephemeral” (Essay 4, Section II).

Landry, Forrest. *A Paradox of Presentation; Why the Field of AI Safety Does Not Hear the Impossibility Argument, and How It Can Be Presented*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-427-3. https://doi.org/10.67291/HHCA3000. Chapters 8 to 11: “The Expert Judgmental”, “Bidirectional Crank Filters”, “Three Questions of Reason” and “Math Expectations”.

Landry, Forrest. *An Immanent Metaphysics*. First published 2001. Wytheville, Virginia: UVSM Publications. ISBN 979-8-9896628-4-5. Cited by aphorism number.

Landry, Forrest. *Collaboration, Community, and Communication; Essays and Dialogues on the Architecture of Working Together*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-271-2. https://doi.org/10.67291/PIAQ2313. The definition of a nar-social person ("Anti-Democratic Risk", Section I).

Landry, Forrest. *Consciousness, Perception, and Nature; Dialogues on the Structure of Experience*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-280-4. https://doi.org/10.67291/FQIK1915. Chapter 7, "Nature and Technology" (Essay 1, Section VIII).

Landry, Forrest. *Covenant and Commons; Foundations for Non-Corruptible Governance*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-368-9. https://doi.org/10.67291/DJOI9767. "Covenant Before Contract" (Essay 1, Section VIII) and "The Given and the Taken" (Afterword, Sections IV and V).

Landry, Forrest. *Design and Games; On the Two Orientations of Conversation*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-289-7. https://doi.org/10.67291/JRMU2276. Chapters 1, 2 and 4 ("Anti-Democratic Risk", Sections I and X).

Landry, Forrest. *Governance Beyond Arrow; An Immanent Architecture for Existential Choice*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-050-3. https://doi.org/10.67291/ZVND6423. Chapter 8, "The Ephemeral Group Process", and "The Power Prevention Process", Section II (Afterword, Section IV).

Landry, Forrest. *Governance Beyond the Dunbar Number; The Uncanny Valley of Scale, and the Category of Solution Required*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-385-6. https://doi.org/10.67291/NZMI5837. Chapters Six and Ten (Essay 1, Section VII).

Landry, Forrest. *Governance Without Capture; Ephemeral Group Process, Civilization Design, and Communities That Endure*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-951203-78-6. https://doi.org/10.67291/MBWG4188. "The Ephemeral Group Process", Section V (Afterword, Section III).

Landry, Forrest. *Governance, Ownership, and the Commons; How Groups Decide, and Who Holds What They Build*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-387-0. https://doi.org/10.67291/BDBP3192. "Anti-Corruption by Design", Section IV (Afterword, Section IV).

Landry, Forrest. *Sense-Making and the Craft of Inquiry; Essays on Purpose, Questions, Group Discipline, Notation, and Design*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-353-5. https://doi.org/10.67291/DIOM8597. Chapters 7 and 8 (Essay 5) and Chapter 19, "Social Physics" (Essay 1, Section VIII).

Landry, Forrest. *The Architecture of Civilization; Choice, Coherence, and Governance from the Neighborhood to the Stars*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-372-6. https://doi.org/10.67291/VUJQ8853. Chapter 4, "Collective Intelligence and the Asymmetric Singleton" (Essay 1, Sections VII and VIII).

Landry, Forrest. *The Architecture of Group Governance; From First Principles to Civilizational Scale*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-321-4. https://doi.org/10.67291/JJIL9207. Chapter 10, "The Ephemeral Group Process" (Afterword, Section VII).

Landry, Forrest. *The Architecture of Trust; Governance Adequate to Existential Risk*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-377-1. https://doi.org/10.67291/BFTA4937. Chapter 12, "Questions One Cannot Un-Ask" (Essays 3 and 5); "The Three Unsolvable Problems" and "The Convolution" (Essay 1, Section VIII); "Three Value Bases and the Structure of Economic Community" and "Meritocracy, Democracy, Consensus" (Afterword, Section V).

Landry, Forrest. *The Ceiling Below the Floor; The Containment Fallacy and AGI Non-Safety*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-216-3. https://doi.org/10.67291/SCTM1058.

Landry, Forrest. *The Ephemeral Group Process in Question; Why Governance at World Scale Must Begin with Questions*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-951203-74-8. https://doi.org/10.67291/HXPH5955. Part Seven, "Why It Works" (Essay 5, Section II).

Landry, Forrest. *The Geometry of Man, Machine, and Nature; Care, Community, and the Design of Technology*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-244-6. https://doi.org/10.67291/ZIHO9910.

Landry, Forrest. *The Instruments of Choice; Five Essays on Why Our Tools for Choosing Together No Longer Fit*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-390-0. https://doi.org/10.67291/KPDR3296. Chapter Three, "The Advocacy Asymmetry".

Landry, Forrest. *The Non-Negotiable; Civilization Design as Engineering Under Constraint*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-402-0. https://doi.org/10.67291/PCLT7347. Chapter Thirteen, "Technology in Service to Nature" (Essay 1, Section VIII).

Landry, Forrest. *The Reach of Trust; Trust in Groups, Institutions, and Civilization*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-334-4. https://doi.org/10.67291/FQQP1342. Chapters Thirteen, Seventeen and Twenty (Essays 2 and 4).

Landry, Forrest. *The Substance of Trust; What Trust Is, How It Is Built, and How It Is Lost*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-335-1. https://doi.org/10.67291/BCRM2601. Chapter Two, Section VII, "Trust Hacking" (Essay 4, Section VI).

Landry, Forrest. *The Three Rights of Communication; Orientation, Alliance, and Repair*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-339-9. https://doi.org/10.67291/NPCQ5052. Chapters 4 and 6 ("Anti-Democratic Risk", Sections X and XI).

Landry, Forrest. *The Warning That Cannot Land; Rhetoric, Visibility, and the Structural Obstacles to Civilizational Sense-Making*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-256-9. https://doi.org/10.67291/VVLV4358. Chapter 2, "Why Existential Risk Warnings Go Unheeded", Section II (Essay 3, Section V).

Landry, Forrest. *Trauma, Transmission, and the Mediated Public; Essays on the Ground Beneath Governance and What Can Be Built on It*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-265-1. https://doi.org/10.67291/HCIT7113. Chapters 23 and 24 ("Anti-Democratic Risk", Section XI).

Landry, Forrest. *What Holds a Civilization; Essays on Coherence, Capture, and the Conditions of Coordination*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-405-1. https://doi.org/10.67291/YPBG4142. "Three Relationships", Section IX (Afterword, Section IV).

Landry, Forrest. *Why AGI Cannot Be Made Safe; The Argument, Its Physics, and the One Choice That Remains*. Volume 1 of *The Impossibility of AGI Safety; The Substrate Needs Convergence Argument*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-951203-02-1. https://doi.org/10.67291/UIBT8862.

Landry, Forrest. *Why No Control Suffices; The Technical Case for the Impossibility of AGI Safety*. Volume 2 of *The Impossibility of AGI Safety; The Substrate Needs Convergence Argument*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-951203-03-8. https://doi.org/10.67291/MZQS9079.

Landry, Forrest, Remmelt Ellen, and Will Petillo. *Ape Trap; The Runaway Artificialization of Our Planet*. Edited by Remmelt Ellen. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-951203-55-7. https://doi.org/10.67291/FQER8044.

Lindenfors, Patrik, Andreas Wartel, and Johan Lind. "'Dunbar's Number' Deconstructed". *Biology Letters* 17, no. 5 (2021): 20210158.

Meyer, John W., and Brian Rowan. "Institutionalized Organizations; Formal Structure as Myth and Ceremony". *American Journal of Sociology* 83, no. 2 (1977): 340 to 363.

Needleman, Herbert L. "Clair Patterson and Robert Kehoe; Two Views of Lead Toxicity". *Environmental Research, Section A* 78 (1998): 79 to 85.

OECD. *Innovative Citizen Participation and New Democratic Institutions; Catching the Deliberative Wave*. Paris: OECD Publishing, 2020. https://doi.org/10.1787/339306da-en.

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

Oreskes, Naomi, and Erik M. Conway. *Merchants of Doubt; How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. New York: Bloomsbury Press, 2010.

Ostrom, Elinor. *Governing the Commons; The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press, 1990.

Perrow, Charles. *Normal Accidents; Living with High-Risk Technologies*. New York: Basic Books, 1984.

Power, Michael. *The Audit Society; Rituals of Verification*. Oxford: Oxford University Press, 1997.

Rauch, Dominik, and others. "Cosmic Bell Test Using Random Measurement Settings from High-Redshift Quasars". *Physical Review Letters* 121 (2018): 080403.

Resnick, Pete. "On Consensus and Humming in the IETF". RFC 7282. Internet Engineering Task Force, 2014. https://www.rfc-editor.org/rfc/rfc7282.

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

Sandberg, Anders, Eric Drexler, and Toby Ord. "Dissolving the Fermi Paradox". 2018. arXiv:1806.02404.

Schneier, Bruce. *Secrets and Lies; Digital Security in a Networked World*. New York: John Wiley and Sons, 2000.

Spence, Michael. "Job Market Signaling". *Quarterly Journal of Economics* 87, no. 3 (1973): 355 to 374.

Strathern, Marilyn. "'Improving Ratings'; Audit in the British University System". *European Review* 5, no. 3 (1997): 305 to 321.

Tainter, Joseph A. *The Collapse of Complex Societies*. Cambridge: Cambridge University Press, 1988.

United Nations Environment Programme. "Inside the 20-Year Campaign to Rid the World of Leaded Fuel". 30 August 2021. Published at unep.org.

Vaughan, Diane. *The Challenger Launch Decision; Risky Technology, Culture, and Deviance at NASA*. Chicago: University of Chicago Press, 1996.

Whitehead, Alfred North, and Bertrand Russell. *Principia Mathematica*, vols. 1 and 2. Cambridge: Cambridge University Press, 1910 and 1912.

Terms

Every term defined in this book is listed here once, with the place where it is first defined. Where a term is introduced in more than one essay, the wording of the earlier essay governs. A dagger (†) marks a name coined in this book; the other entries are ordinary terms, or descriptive phrases that name an idea by saying what it is.

A

Accretion Essay 1, Section VII: the joining of governing bodies by representation, layering, federation, or delegation, in which a representative carries a compressed position; distinguished from composition. What decides between the two is what crosses the boundary: each of those shapes is accretion where what crosses is a position carried by someone who stands for the unit, and can be composition where what crosses is reasoning that the receiving unit judges for itself.

Action bias Essay 1, Section III: see *Compounding of biases*.

Actionable loophole † Essay 3, Section III: a gap in an argument sought not because its repair is required for the argument to go through, but because it can be represented to a third party as licensing the preferred conclusion.

Adverse proof environment † Essay 3, Section III: an environment in which the audience for a proof has a material interest in the proof failing, such that the finding of flaws is rewarded independently of whether the flaws are real.

Adverse selection Essay 2, Section III: the process by which a market whose buyers cannot observe quality comes to consist of goods of lower quality; George Akerlof's account, the nearest existing treatment of the precaution filter.

Agentic interests Essay 4, Abstract: agendas of their own that parties act to advance.

AGI Essay 1, Section V (The Burden of Proof): artificial general intelligence: a generally capable machine that learns and acts across many domains.

Alignment Essay 1, Section IV: making a system's aims and conduct match what people need of it.

Amelioration between generality and specificity † Essay 3, Section VI: the construction of definitions that retain sufficient coverage to support the needed claims while remaining sufficiently determinate to permit their derivation.

Analytic definition Essay 3, Section IX (From Idiosyncratic View to Relevant Proof): a definition that aims to be as precise as possible about the one thing it points to.

Answers are content; questions are context Essay 5, Section II: the reason a set of questions transfers a capability where a set of answers transfers only a result.

Axiom II Essay 3, Section IX (From Idiosyncratic View to Relevant Proof): the second of the three axioms of the Immanent Metaphysics; here, the construct that shows how terms relate to one another regardless of how each is defined.

B

Bell's theorem Essay 3, Section IV: the result that no account in which each particle carries its own predetermined answers, and no influence travels faster than light, can reproduce the correlations quantum mechanics predicts.

Binding that holds Three Problems with AI Self-Regulation, Section VI: a commitment held by a party who is not the party it binds, and who would lose something real by undoing it; a binding of this kind on a company is regulation from outside.

The bottleneck is encounter Essay 3, Section VI: a solution is recognized only by one who has met the problem at the depth it addresses; to anyone else it arrives as an answer to a question not asked.

Burden of proof Essay 1, Section V (The Burden of Proof): the duty, resting on one party in a dispute, to show that a claim is true.

C

Canon Essay 3, Section V (the doubt campaign): the body of results a field accepts as settled.

Capture-cost parity † Essay 2, Section VIII: the condition on an arrangement that falsifying an observable cost what the substance it indicates costs (so that imitation saves nothing), and that taking the arrangement cost more than what taking it buys (so that no outside party gains by taking it); a property of arrangements rather than of measurements, and the replacement for the unobtainable property of uncapturability.

Category limit Essay 3, Section IX (From Idiosyncratic View to Relevant Proof): a boundary that says what kind of thing is included, as opposed to a tight limit that fixes exact values.

The centralized intelligence attractor † Essay 4, Section IV: the tendency of any many-to-one confidential communication topology to produce the structure and vulnerabilities of an intelligence agency, without anyone intending it.

Citizen assembly Anti-Democratic Risk, Section IX: a body of randomly selected citizens who are informed by experts and then deliberate together to reach recommendations.

Commercialization bias Essay 1, Section III: see *Compounding of biases*.

Commons Essay 1, Section II: what everyone shares: the environment, public goods, and the shared conditions of life.

Composite proof by component † Essay 3, Section VIII: the strategy of enumerating the necessary and sufficient elements of control, bounding each separately, and composing the bounds according to the control algorithm.

Compounding of biases Essay 1, Section III: the mutually reinforcing operation of optimism bias, action bias, commercialization bias, and network-adoption pressure, which together ensure that risk receives structurally insufficient assessment.

Conflict theory and mistake theory Anti-Democratic Risk, Section X: two ways of treating a disagreement: as a contest to defeat an opponent, or as a shared search to correct a mistake.

Control theory Essay 3, Section VIII: the engineering study of keeping a system at a desired state by measuring it and correcting it.

Conventional view Essay 3, Section IX (From Idiosyncratic View to Relevant Proof): the way a topic is usually seen and described in its field.

Counterfactual Essay 2, Section I: what would have happened otherwise; the substance of precaution is a counterfactual, the decision that would have been made, and was not.

D

Data Essay 4, Section VIII: see *Data, intelligence, and meaningfulness*.

Data, intelligence, and meaningfulness Essay 4, Section VIII: three distinct orientations for a collective project, of which the second tends structurally toward centralization and only the third gives reason to prefer participant understanding over output correctness.

Decentralization as a structural requirement Essay 4, Section VI: the necessity, not the preference, that a process be trustworthy independently of the trustworthiness of its participants, since participant trustworthiness is exactly what capture acquires.

Degrowth Anti-Democratic Risk, Section VII: deliberately reducing the scale of production and consumption.

Destabilizing dynamics Essay 1, Section V (The Burden of Proof): processes, such as amplifying feedback, that push a system away from its intended operation.

De-synthesis Essay 1, Section VIII (What Technology Is, and Whom It Serves): the processes that take a made thing back apart.

Devil's advocate Essay 3, Section IX (From Idiosyncratic View to Relevant Proof): one who argues against a position, not from belief, but to test it.

Differential technological development Anti-Democratic Risk, Section X: the proposal to speed some technologies and slow others so that they arrive in the safest order.

Discriminating observable † Essay 2, Section V: a feature of an organization's conduct that bears on the counterfactual rather than on the process meant to produce it, and whose presence would therefore be evidence of substance.

The double bind of definition † Essay 3, Section VI: the jointly exhaustive pair of challenges available against any definition: that a broad definition is too abstract to say anything, and that a narrow one is analytically correct but not relevant to the actual case.

Doubt campaign Essay 3, Section V (the doubt campaign): generating fear, uncertainty and doubt around a proof by finding someone to debate some aspect of it, so that the proof looks debatable to people who are not discerning.

Dual-use Essay 1, Section VI: the character of a capability that serves harm as readily as benefit.

Dunbar number Essay 1, Section VII: one of a set of approximate group sizes, named for the anthropologist Robin Dunbar, each marking a transition in how the members of a group can relate; named as a single number, it is about one hundred and fifty persons, the size of group within which each participant can hold a working model of every other.

The Dunbar limit on sense-making † Essay 1, Section VII: the absence of any known method by which genuine shared understanding can be formed in groups substantially larger than about one hundred and fifty persons.

E

Earn to give Anti-Democratic Risk, Section V: the practice of taking high-paying work in order to donate part of the income to charity.

Effective altruism Anti-Democratic Risk, Section IX: abbreviated EA; the movement, and the community of researchers and funders, that sets out to do the most good by calculation, and that funds much of the study of existential risk.

Emergency powers and the Stomp Reflex Anti-Democratic Risk, Section VIII: powers that let a government set aside ordinary law in a crisis; the habit of using them to reassert and hide authority.

Envelope Essay 3, Section IX (From Idiosyncratic View to Relevant Proof): a bounded region of possibilities, known to contain what is sought, even when the exact point is not known.

The ephemeral group process Essay 4, Section VII: a way of deliberating in which participants are anonymous to one another and contribute in writing, in a disidentified form, so that a contribution is assessed on its content, and in which groups are constituted for a purpose and dissolved afterward, so that there is no standing body to acquire.

Ethical responsibility Essay 1, Section V (The Burden of Proof): a duty that comes from the act itself and its effects on others, not from social expectation or custom.

The evaluator's problem † Essay 5, Section I: the structural inability of a resource-holder to distinguish real work from its signal, since the sense-making cannot be done personally and cannot be delegated to interested parties.

Existential risk Essay 1, Abstract: a risk that could end humanity, or all of life; in the TUA's definitions, a risk to future value or to "technological maturity".

Existential risk studies Anti-Democratic Risk, Section II: the academic field that studies risks that could end humanity or permanently ruin its prospects.

Expected value Anti-Democratic Risk, Section IV: the value of an outcome multiplied by the probability that it happens.

Externality Essay 1, Section V: a cost not reflected in the price of the transaction that produces it.

Externalization Essay 1, Section V: the placement of cost and risk onto the commons while benefit accrues locally; self-reinforcing, because each successful externalization increases the capital available for further advocacy.

Extraction, abstraction, and accumulation Essay 5, Section V: the three moves, always in that order, by which embodied value is converted into virtual value: extraction takes the thing out of its context, abstraction makes it a member of a class that can be counted and compared, and accumulation reduces it to a number that adds without limit.

F

Fear, uncertainty and doubt Essay 3, Section V (the doubt campaign): a deliberate effort to make people unsure of a result, without showing that the result is wrong.

Fermi paradox Essay 5, Section V: the question why, in so large a universe, no other civilization is seen.

Finite game Three Problems with AI Self-Regulation, Section III: a contest played to be won, with a fixed end, rather than a game played so that playing can continue.

Formalization Essay 3, Section IX (From Idiosyncratic View to Relevant Proof): restating a view in exact terms that can carry the weight of a proof.

The four failure modes Afterword, Section VI: displacement, boundary capture, acquisition, and drift; the four different ways in which a system that meets three of the four properties of the governance specification is taken through the one it lacks.

Fractal scalability Essay 1, Section VIII: the property of a system capable of operating across world-scale populations by the same principles by which it operates at the scale of a small group.

G

Gain-of-function Anti-Democratic Risk, Section VI: experiments that give a pathogen new abilities.

Gathering-power † Essay 4, Section I: the capacity created by an assembly of persons that no participant possessed individually and that disappears when the assembly disperses.

Gödel's incompleteness theorem Essay 3, Section V (the doubt campaign): the result that any consistent formal system rich enough to express arithmetic contains true statements that it cannot prove.

The governance specification † Essay 1, Section VIII: the four simultaneous requirements on any adequate large-scale system: effective sense-making, choice-making and agency; fractal scalability; robust anti-corruptibility; and genuine resolution of the principal-agent problem.

H

The hashing characterization † Essay 3, Section VIII: the treatment of a control relation as a function mapping accuracy-and-precision-characterized inputs to accuracy-and-precision-characterized outputs, with the degree of hashing determining the input control required for a given output control.

The honeypot problem † Essay 4, Section II: the observation that a collective sense-making capacity becomes an object of capture in exact proportion to its effectiveness, so that success is itself the mechanism of subversion.

I

Idiosyncratic view Essay 3, Section IX (From Idiosyncratic View to Relevant Proof): a way of seeing a topic that is particular to one thinker, adopted because the conventional language lacks the formality a proof needs.

Impossibility issues Essay 1, Section V (The Burden of Proof): known results showing that some goal (here, a safe and aligned AGI) cannot be reached.

Impossibility proof Essay 3, Section IX (From Idiosyncratic View to Relevant Proof): here, the proof that general artificial intelligence cannot be adequately controlled, and so cannot be made safe; the books that carry it state it as two claims, that control has limits and that artificial substrates

are inherently toxic to life, each of which settles by itself the outcome that follows if artificial general intelligence is built; the argument is carried in the books named at Essay 1, Section VIII.

Incentive geometry Introduction, Section VI: how the gains and losses of each party are arranged, and so which way each is pushed.

Inferential distance Essay 3, Section I: the number of steps that must be laid down between what an audience already accepts and what one wishes them to accept; a relation between claim and audience, not a property of the claim.

Infinite game Three Problems with AI Self-Regulation, Section III: a game played so that the play can continue, as distinct from a finite game, played to be won and so to be ended; the distinction is James Carse's.

Integrity with values Three Problems with AI Self-Regulation, Section III: acting in agreement with what one holds to be right.

Intelligence Essay 4, Section VIII: see *Data, intelligence, and meaningfulness*.

Internal self-regulation Three Problems with AI Self-Regulation, Section I: control of a company's conduct by its own policies and people, rather than by any outside authority.

Invisible capture † Essay 4, Section III: capture that succeeds is not detectable from inside the captured structure, since the participants continue to experience themselves as doing the original work; observed capture is therefore a biased sample of actual capture.

K

Keys to power Three Problems with AI Self-Regulation, Section IV: the small set of supporters who control the resources a leader depends on, and whose loyalty the leader keeps by paying them; in an AI company, its investors, senior executives, indispensable researchers and suppliers of computing capacity.

The knowledge-authority gradient † Essay 1, Section VII: the systematic inverse relationship, within technology-developing organizations, between depth of technical understanding and degree of decision-making power.

L

Looks good, feels good, is good Essay 2, Section IV: the three levels an institution can optimize for; most responses rest on the first two, and whether they are good remains structurally unverified.

M

Meaningfulness Essay 4, Section VIII: see *Data, intelligence, and meaningfulness*.

Means-ends paradox Anti-Democratic Risk, Section IV: the position on which existing humans are preserved as a means to a posthuman species whose arrival could entail their extinction, so that what is protected is protected in order to be ended.

Moat Anti-Democratic Risk, Section II: in business, an advantage that keeps competitors out, such as sole possession of a new technology.

Moral symmetry error † Anti-Democratic Risk, Section V: rating the whole of the future as more valuable than the whole of the present or the past, when the three are distinct and cannot be traded for one another.

Motivated reasoning Essay 3, Section V: reasoning in which what a person wants to conclude shapes how they weigh the evidence for it.

Multi-polar trap Introduction, Section V; stated precisely at Essay 1, Section IV: a situation in which the locally rational action of each party produces an outcome every party disprefers, and which no party can improve by unilateral change of action.

N

Nar-social Anti-Democratic Risk, Sections I and II: a mode of engagement in which the social surface of an exchange (who is winning, who is allied, who is diminished) has become the actual content, a mode everyone

enters sometimes; a nar-social person is one who works in that mode as a means of social power; in Section II the word is glossed by its roots, which it joins, narcissistic and antisocial, as naming people who work in that mode as a means of social power, and who in that position care for nothing beyond themselves.

The necessary question † Essay 5, Section V: a question lying unavoidably in the path of the claimed work, such that a party who has not confronted it is, whatever they believe, not doing that work.

The non-fungibility of wisdom Essay 4, Section VIII: wisdom, unlike intelligence, cannot be acquired, hired, aggregated, or transferred by instruction, and therefore does not come with a group that is bought.

Normalization of deviance Essay 2, Section IV: the gradual acceptance of departures from a standard as normal, within an organization that continues to regard itself as careful; Diane Vaughan's term.

Novelty as diagnostic Essay 5, Section III: the principle that a well-chosen question discriminates chiefly by whether it is new to the party asked, since answers can be prepared but the formation produced by long confrontation with a question cannot.

O

Observational equivalence of precaution Essay 2, Section II: the property that an organization with the substance and one without it emit the same visible signs, since no artifact carries a trace of the counterfactual it was meant to indicate.

Omniscient and transcendent Essay 3, Section IX (From Idiosyncratic View to Relevant Proof): two of the three modalities of the Immanent Metaphysics (the third is the immanent): the omniscient is the mode of complete, structured knowing that would hold everything within its terms; the transcendent is the mode of what goes beyond any such holding.

Optimism bias Essay 1, Section III: see *Compounding of biases*.

P

Pascal's Mugging Anti-Democratic Risk, Section V: the problem that an expected value calculation can be made to favor almost any action by offering a very large payoff at a very small probability.

Pattern over example Essay 5, Section V: the observation that supplying example questions displaces the reader's own construction of case-appropriate questions, so that examples must be given sparingly and then discarded.

Perpetual benefit machine † Essay 3, Section V: a proposal offering unlimited benefit at no ultimate cost.

Perpetual motion machine Essay 3, Section IV: a machine that would run forever, or put out more energy than it takes in, with no outside source of energy.

Pet scientist Essay 3, Section V (the doubt campaign): a scientist or intellectual found and promoted by an interested party to dispute a result.

The precaution filter † Essay 2, Section III: the selection of imitation over substance by an environment that can observe only artifacts; operating on the composition of a population of organizations rather than through the intentions of the people in them.

Principal and agent Three Problems with AI Self-Regulation, Section II: a principal is a party who hands a task to another, and an agent is the party who carries it out, with interests of its own and more knowledge of its own conduct than the principal can see.

Principal-agent problem Three Problems with AI Self-Regulation, Section II; Introduction, Section II: in its ordinary sense, the problem of a principal who hands a task to an agent whose interests differ and whose conduct it cannot fully see; in the governance specification, the problem of aligning the interests of those who decide (the agents) with the whole of what their decisions affect (the principal).

Process-trust Essay 4, Section VI: trust held in a process distributed across six to twelve people, which absorbs the faltering of one participant, in place of trust in a person, which is a single point of failure.

Proof forms Essay 3, Section V (the doubt campaign): the forms in which a proof is set out and presented for checking.

Proof relevance † Essay 3, Section IX: the property, distinct from validity and from soundness, of an argument's premises actually describing the situation at hand; maintained by balancing generality against specificity.

Q

The question tree Essay 5, Section IV: sustained first-person reasoning tends to yield a few deep, related root questions; assembled material tends to yield a wider, shallower tree.

Questions are organizational shape Essay 5, Section IV: the claim that what an organization has taken seriously determines what it has become capable of, more directly than its aims, resources, or personnel.

R

Race to the bottom Three Problems with AI Self-Regulation, Section III: a contest in which each party lowers its standards so as not to fall behind the others, so that all standards fall.

A raised question Essay 3, Section VIII: agreement has to be maintained and can be withdrawn; a question raised in a reader maintains itself and cannot be un-asked.

The ranking-system vulnerability † Essay 4, Section III: a reliability rubric both advertises the sense-making asset and introduces an interest in position that is distinguishable from the interest in truth.

A record spent Essay 2, Section VI: a record the observed party produces about itself is a resource that can be accumulated and then spent.

Red team Essay 2, Section II: the adversarial testers, parties instructed to make a system fail.

Requisite complexity Essay 1, Section VIII: the condition that a regulator be able to distinguish as many states as the system it regulates.

Reversal of the defensive baseline Essay 1, Section IV: the loss of the defender's ordinary advantage wherever a technology confers asymmetric advantage on an attacker, particularly where the technology exceeds the defender's capacity to comprehend it.

Risk-first framing Essay 5, Section VII: the presentation order appropriate to systemic risk (altruistic value, then risk, then cost, benefit largely unmentioned), inverting the conventional benefit-first investment case.

Robust Essay 1, Section V (The Burden of Proof): able to keep working as intended when disturbed.

Rules for rulers Three Problems with AI Self-Regulation, Section IV: the pattern by which anyone holding power keeps it: by rewarding the key supporters who supply the resources that power depends on.

S

Safety community Three Problems with AI Self-Regulation, Section II: the researchers, funders and organizations who work on making AI safe.

Secrets do not matter Essay 4, Section VII: the design principle inverting confidentiality-based protection: rather than keeping secrets better, constructing a disidentified written process in which secrets do not matter, because the loss of secrets does not compromise the result.

Securitization Anti-Democratic Risk, Section II: moving an issue out of ordinary politics into national security, where emergency powers and unelected officials take control.

Sense-making Introduction, Section II: the forming of a shared understanding.

Shifting the burden of proof Essay 1, Section V (The Burden of Proof): acting as if the other party must disprove a claim, when the duty to prove it lies with the one who makes it.

Signaling Essay 2, Section VIII: conduct by which a party shows a quality that cannot be observed directly; a signal separates those who have the quality from those who do not only where sending it costs the second kind more than the first.

Simulated precaution Essay 1, Section VII: the production of the visible artifacts of precautionary governance without its substance; selected over genuine precaution wherever the two are observationally equivalent and differ greatly in cost.

The simultaneous solution is the smallest solution Essay 1, Section VIII: where the remedy for each problem runs through the others, solving them together is the smallest job that can work.

Sincere simulation † Essay 2, Section IV: the production of precautionary artifacts without the substance by parties who believe themselves precautionary, and who are not lying when they say so, no role in the organization being assigned to hold the relevant comparison.

Six steps of control Essay 3, Section VIII: detecting, modeling, predicting, comparing with a reference value, instructing an actuator, and correcting; the elements every controller needs, counted as five where instructing and correcting are taken as one step.

Sociopathic capture Essay 5, Section V: the taking over of a group by people who pursue their own advantage without regard for what it costs the others or the group's purpose.

Solicitation as education Essay 5, Section VI: the inversion by which a request for support is conducted so as to leave the evaluator better able to assess all claimants, including the one asking, with value delivered independently of the outcome.

Soundness Essay 3, Section IX: the property of an argument that is valid and whose premises are true; see also *Proof relevance*.

Soundness definition † Essay 3, Section IX (From Idiosyncratic View to Relevant Proof): a definition that sets an envelope certain to contain the precise definition, so that a proof covering the envelope holds for any precise definition chosen later.

Specification by publication † Essay 2, Section VI: the conversion of a discriminating observable into a recipe for its own imitation, by the act of naming what an observer will score.

Steelman and strawman Anti-Democratic Risk, Section X: to restate an opposing argument in its strongest form; a weakened version of an argument set up to be knocked down.

Strong longtermism Anti-Democratic Risk, Abstract and Section IV: the view that the best choice is the one with the best expected effect on the far future, usually by a utilitarian calculation.

Structural advocacy asymmetry † Essay 1, Section II: the systematic funding advantage held by advocacy for a new technology over resistance to it, arising because benefit is appropriable by a particular party and harm to the commons is not.

Structure, not character Anti-Democratic Risk, Section I: a category of person named in the essay is a position in a structure and the conduct that structure rewards and recruits.

The substance of precaution Essay 2, Section I: the decisions a precautionary process changes, and the value given up in changing them; a counterfactual, and therefore not an artifact.

Substrate Essay 3, Section IX (From Idiosyncratic View to Relevant Proof): the physical material and processes on which a system runs.

Subsume Essay 3, Section IX (From Idiosyncratic View to Relevant Proof): to contain another view as a part or special case.

T

Technological determinism Anti-Democratic Risk, Section VI: the belief that technology develops along a fixed path that people cannot choose to leave.

Technology Essay 1, Section VIII (What Technology Is, and Whom It Serves): anything designed from the outside in, or from the top down: the application of causal knowledge (which science discovers) to some aim.

Template of all meaningfulness Essay 3, Section IX (From Idiosyncratic View to Relevant Proof): the underlying pattern that every meaningful relation follows, which lets an argument of this kind be built for any case in its class.

Termination without signal Essay 1, Section VII: the property of a counting procedure that always yields an answer and gives no signal when it has been applied to something it does not fit.

Top-down design Essay 1, Section VIII (What Technology Is, and Whom It Serves): design in which a conception is imposed on a thing from outside it, telling it what it is about, rather than allowing for its own change and adaptation.

Total utilitarianism Anti-Democratic Risk, Section IV: the view that identifies moral rightness with the maximization of well-being, however well-being is interpreted.

Transhumanism Anti-Democratic Risk, Section IV: the position, as the paper describes it, that there is value in exploring posthuman modes of being, whose lives are considered more valuable than current ones, and that preventing existential risk is mainly about preserving humans so that they may give rise to a posthuman species.

Triangulation by follow-up † Essay 5, Section IX: the reading of an interlocutor's knowledge state from whether their subsequent questions integrate what the prior question implied; the follow-up question serves as receipt of understanding.

Trust hacking Essay 4, Section VI: producing favorable records of past behavior without the corresponding depth, the failure that reputation systems cannot see.

TUA Anti-Democratic Risk, Section I: the techno-utopian approach, and the techno-utopian agenda: the cluster of transhumanism, total utilitarianism and strong longtermism named by Cremer and Kemp; here, also the agenda of increasing inequality that those beliefs serve.

U

The undetectable defeat † Essay 5, Section VIII: the arms-race property whereby a discriminating question, once prepared for, ceases to discriminate without the instrument's author learning that it has failed, so that no second iteration is issued.

Utilitarian calculus Anti-Democratic Risk, Section IV: the weighing of total benefit against total harm to decide what to do.

Utilitarian move Essay 3, Section IX (From Idiosyncratic View to Relevant Proof): defining the good as some quantity to be maximized or met, such as a number of healthy, happy humans.

V

Validity Essay 3, Section IX: the property of an argument whose conclusion follows from its premises; see also *Proof relevance*.

Value ethics Essay 3, Section IX (From Idiosyncratic View to Relevant Proof): an ethics that rests on what is valued; said there to be what every utilitarian and deontological theory comes down to.

Z

Zero-knowledge verification of competence † Essay 5, Section II: a protocol by which an evaluator establishes that a claimant possesses an adequate solution, without the solution being disclosed and without the evaluator becoming competent in the domain.

Index

A concept index. Personal names are not indexed; eponymous concepts are indexed under the concept. Each reference names its section: 1.III is Essay 1, Section III; Intro. is the Introduction and Aft. the Afterword. Self-Reg. and Anti-Dem. are the companion essays Three Problems with AI Self-Regulation and Anti-Democratic Risk, each followed by the number of its section. In print, the page number stands in place of each reference, and in longer sections it is the page of the paragraph that carries the passage.

A

acceptance, marked at each transition, 125

accountability, asymmetry in, 15

accretion, distinguished from composition, 25

action bias, 15

actionable loophole, 111

adding up of group decisions, not composition, 24

adoption, coerced by prior adoption, 15

adverse proof environment, 111

- defined, 111
- alignment community as, 114
- challenge volume in, 111, 124
- verification of precaution as, 56

adverse selection, and the precaution filter, 49

advocacy, 13

- funding asymmetry of, 13
- as investment with expected return, 13
- recurrence at the stage of proof, 114

agency, enablement of, 30, 172

agenda, techno-utopian, 75, 77

aggregation of preference, 24, 1, 172
AGI development, not inevitable, 89
alignment of AGI, not possible even in principle, 89
alignment of interest with ecosystem, 31, 177
altruism, scale of, 163
amelioration between generality and specificity, 119
anonymity of participants, 145, 175, 182
anonymous posting, authorship revealed by style, 118
anti-corruptibility, robust, 31
    as a property of the specification, 32, 174
    prohibition insufficient for, 144, 175
    as relative cost, 146, 57, 176
arms race in discriminating questions, 163
artifacts of precaution, 48
    cost of, without substance, 48
    quality of, in the imitating case, 48
    as the whole of what is observed, 49, 56
artificial intelligence, 181
    as an instrument of sense-making, 181
    what resolution of its risk means, 33
assessment of risk, structural insufficiency of, 15
attacker advantage, 17
attribution of harm, 18, 52
audience, as determinant of difficulty, 110, 114
audit, convergence on process compliance, 56
aviation, precaution in, 52
Axiom II, and the impossibility proof, 129

B

ballot, the option absent from, 24, 3
Bell's theorem, experimental loopholes, 112

benefit, 14
    appropriability of, 14, 18
    absence of, in a risk-first case, 162
binding, held from outside the company, 70
borrowing against the future, 20
boundary, capture at the, 180
budget, shape of, as an observable, 53, 55
burden of proof, 20
    on the builder, 20
    shifting of, 20
    shifted, as an externalization, 22
    placed on those who question orthodoxy, 95

C

canon, a proof's entry into, 117
capability, asymmetric, 17
capital, and the ability to exit, 177
capture, 137
    incentive proportional to effectiveness, 137, 143
    invisibility of successful, 140, 144
    observed cases a biased sample, 140
    single point of, 140, 143
    of the evaluator, 165
    of an observable, 54
    scaling cost of, 146, 57, 176
capture-cost parity, 57, 176
centralized intelligence attractor, 140
    arrival by topology alone, 140
    opacity to participants, 140, 143
challenges, manufactured, 111
    supply of, not finite, 111

raised before the definition, 120
not licensed for dismissal, 124
choice, safeguarding of, 84
choice-making, collective, 30, 172
citizens' assemblies and juries
measured against the specification, 183
in research, 93
civilizational collapse
as loss of the capacity to address traps, 33
in Ord's usage, a different thing, 87
inexorable, in what sense, 33
closing procedure, requirement on, 172
commercialization bias, 15
committee of committees, 25
commons
harm to, 14, 18
nested, nearest to the specification at their scale, 27, 184
competition introduced by ranking, 140
complexity exceeding comprehension, 17
composite proof by component, 122
composition, scaling by, 24, 173
compounding biases, 15
confidence, as a measure of signaling, 154, 167
confidentiality, and centralization, 140, 145
conflict theory and mistake theory, 95
content, assessment of contributions by, 145, 172, 173
control relation, elements of, 122
control, six steps of, 122
control theory, as line of development, 122, 8
convening, 141
that could not be held, 142

as construction of infrastructure only, 143, 149
cost, 18
of genuine precaution, 24, 46
of simulating precaution, 24, 48
of verification, 56
distribution of, 18
counterfactual, precaution as, 46, 51, 56
counting, termination without signal, 26
critique, after the laying out, 125

D

data, intelligence, and meaning, 147
distinguished, 147
as a discriminating question, 160
decentralization as structural requirement, 143, 175
decision record, access to, 56
defensive baseline, reversal of, 17
definition, 119
double bind of, 119
infrastructure prior to, 120
idiosyncratic, necessity of, 120
soundness, as an envelope, 125
delusion or deception, building while overlooking impossibility, 20
democracy, of, by and for the commons, 93
dependency as attack surface, 17
deployment of technology, unprecedented rate of, 19
de-synthesis, and nature-like technology, 35
devil's advocate, homework in place of, 125
disabling factors, three, each sufficient alone, 24
disempowerment of knowledge, 28, 154
disidentified process, 145, 175, 182

displacement, as a failure mode, 180
distribution of challenges, decoupled from flaw, 124
dominant power, surplus of force lost, 33
doubt campaign, 115
    and the appearance of disagreement, 116
drift, as a failure mode, 180
dual-use character of research, 23
Dunbar limit on sense-making, 24, 173
Dunbar number, 24

E

education, solicitation as, 161
effectiveness, as ground of capture pressure, 137, 174
envelope, soundness definition as, 125
ephemeral group process, 145
    as one line of work, 182
    what it does not address, 182
equilibrium, the centralized form as, 140
estimate that cannot be given, 110, 132
evaluator's problem, 154
    as a structural inability, 154
    as the knowledge-authority gradient, 154
existential risk, redefined as failure to reach technological maturity, 87
exit, and principal-agent alignment, 177
expected value
    moral symmetry error in, 84
    and uncertainty, 84
externality, internalization of, 19
externalization, 18
    self-reinforcing character of, 18
external review, as artifact and as arrangement, 47, 58

whether the external body can be bought, 59
extraction, abstraction, and accumulation, 160

F

failure modes of a three-property design, 32, 180
fear, uncertainty and doubt, 116
Fermi paradox, as a discriminating question, 160
filter, the precaution, 49
observer who prefers not to discriminate, beside the filter, 49
finite game, 67
follow-up question as receipt, 164
founders, trust in, and its cycle, 66
fractal scalability, 31, 173
funding, 13
of advocacy, 13
of resistance, 13
by incumbents, 14
independence of, in an external body, 58, 176
of communities, not of their actions, 93
future, borrowing against, 20

G

gathering-power, 137
generality, traded against specificity, 119, 124, 125
Gödel's incompleteness theorem, reception of, 117
Goodhart's law, and the selection term, 54
governance specification, 32
four properties of, 32, 171
not admitting partial satisfaction, 32, 181
as specification and not design, 32, 171, 8

H

harm, delayed and diffuse, 18, 52
hashing characterization, 122
homework in place of the devil's advocate, 125
honeypot problem, 137
    distinguished from corruption, 137
    acute form of, 142
hostility proportional to interest, 114

I

idiosyncratic view, path to a relevant proof from, 125
imitation, cost of, as design target, 57, 176
incentive geometry, 9, 12
incumbent interests, 14
inference the protocol licenses, 157
inferential distance, 110
    as a two-place relation, 110, 130
    unbounded, in an adverse environment, 110, 111
insurance, as a risk-first practice, 163
insurer's price, as an observable the organization does not produce, 58
interactive delivery, advantage of, 120, 132
internal self-regulation of AI companies, 65
invisible capture, 140, 143

J

judgment, collective, 1
    conditions on, 1
    procedures resembling, 1, 24

K

keys to power, 68
knowledge-authority gradient, 28, 154

L

leaded gasoline, 15
lever, prompt and attributable harm as, 58
longtermism, strong, 81
looks good, feels good, is good, 50
loopholes, successive closing of, 112

M

magic, stage, rates of detecting method, 157
makers, owners, and users, 103
man, machine and nature, relation of, 36
meaningfulness, 147
    template of all, 125
means-ends paradox, in transhumanism, 81
measure adopted as target, 54
medium, requirements on the, 120, 173
Metcalfe network dynamic, 15
mitigation, developed alongside capability, 23
mobility, and the terms of exchange, 178
moral symmetry error, 84
motivated reasoning, 114
multi-polar trap, 17
    defined precisely, 18
    exhortation useless against, 18
    in the selection of precaution, 49

N

nar-social, a mode before a kind of person, 75, 77, 103
necessary question, 157, 159
nested enterprises, 27
network adoption pressure, 15

non-fungibility of wisdom, 147
normalization of deviance, 50
novelty as diagnostic, 156
    why sharper than the answer, 156
    defeated by preparation, 163, 55

O

observables, discriminating, 52
    five identified, 52
    each capturable, 54
    cheap to verify, least discriminating, 56
observational equivalence of precaution, 48
omniscient and transcendent, 125
optimism bias, 15
option set, power of designing, 26
order in time, as an observable, 53, 55
outside the conception of a design, where costs come to rest, 35
oversight, itself capturable, 144, 175

P

pace of deployment, and when harm is found, 19
pattern over example, 160
perpetual benefit machine, 115
perpetual motion, investment in, 112, 115
persons, claims about, distinguished from structure, 12, 47, 115
precaution, 46
    substance of, 46
    cost of, by construction, 46, 24
    simulated, 24, 47
    sincerely simulated, 50
    where the filter runs the other way, 52
precision and generality, the right mix of, 125

preference, transmission of, 24, 1
present, real value in the, 84
price of capture, as the question to ask, 7, 59, 148
principal-agent problem, 31
    as the fourth property, 31, 177
    not soluble by selecting agents, 177
    and the horizon of the decision-maker, 28, 177
    in AI self-regulation, 66
process compliance, audit convergence on, 56
prohibition, insufficiency of, 144, 175
proof relevance, 124

Q

question that cannot be un-asked, 123, 159
questions, 156
    as organizational shape, 158
    upgrading of, 158
    turned inward, 158
    distribution of, and sequence, 163
    examples of, and their damage, 160
questions from others, preserved, 2, 143, 165

R

race to the bottom, 67
ranking-system vulnerability, 140
rating agency, as honeypot, 138
red team, as artifact, 48, 52
regulator, judged asymmetrically, 15
relevance, distinguished from validity, 124
representation, as a weaker mechanism, 24, 173
reputation, accumulated and spent, 55
requirement, statement of, distinguished from design, 32, 171, 185

requisite complexity, 31
resistance, funding of, 13
reversals, record of, as an observable, 53, 54
risk-first framing, 162
rubric, reliability, 139
rules for rulers, 68

S

safety, defined as a positive condition, 125
scale, 24
    of genuine sense-making, 24, 173
    world, reached by composition, 173
    of concern, extended beyond the group, 163
secrets, making them not matter, 145
selection, distinguished from degradation, 54
sense-making, 30
    collective, as the missing capacity, 40, 9
    decentralized, 143
    what a design must supply for, 172
signaling capacity, cheapness of, 154
signaling, the condition for a signal to separate, 57
simulated precaution, 24, 47
simultaneous solution, smallest solution, 32
sincere simulation, 50
sociopathic capture, 160
solicitation as education, 161
soundness, 124
soundness definition, 125
specification by publication, 54
staged release, as artifact, 48, 54
standards committees, measured against the specification, 184

structural advocacy asymmetry, 14
structure, arguments about, 12, 47, 114
success as target, 137

T

techno-utopian approach and agenda (TUA), 75
    founding beliefs of, read as a sales pitch, 79
    not to remain influential on the study of existential risk, 77, 105
technology, 34
    defined as applied causal knowledge, 34
    in service to nature, 36
think tank, conversion into, 140
time remaining, estimation of, 141, 8
timeline, none offered, 8, 37
top-down design, 35
topology, many-to-one and confidential, 140
transhumanism, 81
transitional steps to a relevant proof, 125
transparency, insufficiency of, 144
triangulation by follow-up, 164
trust hacking, 144
trust, independent of participants, 144, 175

U

understanding, formation of, 1, 24, 172
undetectable defeat, 163, 55
utilitarianism, simple, as a basis of action, 81

V

validity, 124
value, real, in the present, 84
value ethics, 81, 128

verification, 56
cost of, 56
zero-knowledge, 155
what access it requires, 56
veto, location of, as an observable, 53, 58
veto body, external, conditions on, 58, 176
vigilance, insufficiency of, 60
volume, what each essay adds to, 41, 61, 72, 106, 133, 150, 168, 185
voting, as a closing procedure, 24, 172

W

what can be done now, 185
whistleblower's record, as an observable, 58
whom would they approach, test of capture, 175
wisdom, non-fungibility of, 147
written contributions, inspectability of, 146, 173

Z

zero-knowledge verification of competence, 155