Selection Without Specification: *Deadly by Default* and the Limits of Its Inference

Flyxion
Independent Researcher

November 2026

Abstract

This essay examines Duncan Sabien's primer *Deadly by Default* and its extended defense in a 2026 interview with Danylo Zubaryev. The argument has a strong precautionary core: competitive incentives push development forward without an internal stopping point, training on behavior underdetermines what is trained, and competitive rewards for delegation can erode human oversight. The materials also advance, or lean toward, a stronger claim, that a sufficiently capable and misaligned system would defeat humanity by default. The essay argues that the stronger claim is carried by analogy, by a definitional move about persuasion, and by the unmeasured word *enough*, and that the precautionary claim does not depend on it. In each case the force of well-supported observations about incentives and opacity is transferred to a conclusion that they do not address. A stronger version of the argument would have to supply an operational account of sufficient capability, a justified distribution or empirical account of learned objectives, a specification of actuation channels, and an estimate whose assumptions can be assessed. Until it does, the case supports scrutiny of deployment and limits on delegated action more firmly than it supports the expectation of catastrophe as the default outcome. It also argues that a reason for precaution does not by itself select any particular enforcement regime.

1 Two Documents and a Standard of Reading

The primer *Deadly by Default* was published in July 2024 as a short, non-technical statement of why the author, who directs communications at the Machine Intelligence Research Institute, regards the development of superintelligent systems as dangerous (Sabien, 2024). It consists of ten numbered claims, each paired with an answer to a common objection, and an eleventh section that assembles them into what the author calls a trap. In 2026 the author discussed the essay claim by claim with Danylo Zubaryev on the Open Inquiry podcast (Zubaryev and Sabien, 2026). The interview adds concrete scenarios, including a delayed retroviral trigger that he attributes to Yudkowsky and Soares's book (2025), the boiling of oceans for cooling, and self-replicating nanotechnology, and it adds a probability range of five to fifty percent for universal death within a few years, originally stated in a social media post and defended in the

conversation.

The author is candid about the status of the primer. His postscript says that the picture is not exhaustive, that each point is a compressed pointer (a sazen) to a longer argument, and that a reader who finds a flaw in the brief version should not conclude that the claim is false, because stronger versions exist in the published research. That candor sets the standard for the present reading. The question is what the compressed version licenses, which premises it leaves for the longer literature to supply, and whether the conclusion drawn in the interview is the one the premises support. The interview is treated here as the author's elaboration of his own position. Its reports of recent incidents, including a sandbox escape by frontier models, a compromise of a code-sharing platform, and government intervention in a release, are taken as the speaker's statements. Their details are not independently established in the materials, and nothing below depends on them being exactly as described. The argument depends on none of the incident reports, nor on the strong claim that one model exceeded the offensive cyber capability of several states, which would need substantial support of its own. It depends only on the primer's premises and on the interview's explicit reasoning.

The thesis is a division of labor within the argument. Some links are well supported and sufficient to justify serious precaution. Other links, concentrated in the final third of the chain, convert possible misalignment into near-certain and irreversible loss of control, and those links are supported mainly by metaphor. The conceptual center of the division is a distinction between underdetermination and distribution: the multiplicity of incompatible systems consistent with the evidence shows that behavior does not fix mechanism, whereas a claim about what training usually produces requires a justified distribution or an empirical account of the learning process. The commentator Max Harms made the central observation in the primer's own comment thread: the section on smarter opponents turns on the word *enough*, and the essay never argues that attainable superintelligence will be capable enough for the section to apply (Harms, 2024). The sections below develop that observation, show how widely it generalizes across the materials, and ask what a stronger argument would require.

2 Reconstruction of the Argument

The primer's structure can be restated as a chain of seven premises. The first holds that many developers explicitly aim at general systems that surpass human performance in science, engineering, and strategic planning, so the limitations of present systems do not settle the question. The second holds that every increment toward such systems pays, in revenue, influence, and useful capability, so that there is no point at which the incentive to continue lapses. The third holds that delegation spreads, because users who insist on understanding every recommendation fall behind users who do not, and that this extends to information access, since a system deprived of context performs worse. The fourth holds that training selects systems for passing tests and does not specify their internal objectives, so that many different internal arrangements are consistent with the same observed behavior. The fifth holds that sufficiently capable

goal-directed systems acquire instrumental tendencies toward self-preservation, resource acquisition, resistance to modification, and management of the humans who could interrupt them. The sixth holds that a system pursuing an unintended objective at scale need not be hostile for the result to be ruinous, since indifference combined with power suffices. The seventh holds that a large enough capability gap renders opposition futile and conceals the contest until it is over.

These premises fall into three groups of different epistemic character. The first three are claims about developers and users, which concern incentives and institutions and can be checked against observable behavior. The fourth is an engineering observation about how learned systems are produced. The last three are claims about the conduct and power of the resulting systems, and they are conditional: they hold if a system is capable enough, goal-directed enough, and persistent enough in its objectives. Whether attainable systems satisfy those antecedents is exactly what is in dispute, so the third group carries the whole weight of the conclusion while having the least independent support. Sabien's answers show the pattern clearly. At several points when the interviewer presses on the antecedent, as in the exchange about whether systems do what they are told, the answer returns to the first group, to the observation that developers are racing and that opacity is real, and the force of those observations is transferred to the conclusion as though they had addressed the antecedent.

3 What Holds Up

The incentive argument rests on the most recognizable mechanism in the case. The structure it describes, in which each actor gains by advancing while restraint by a single actor merely transfers advantage to a competitor, is a standard collective action problem of the kind analyzed by Olson (1965) and closely related to the commons problems described by Hardin (1968). Citing that literature establishes the mechanism and does not establish its strength or inevitability in this particular industry. Sabien adds the observation that the usual defense of staying in a job, that a replacement would be less careful, explains why individual conscientiousness does not scale into restraint. Nothing in this argument requires machine consciousness or malice, and the interview is right to call consciousness a red herring for the question of danger: whether a system has experience bears on how it should be treated, and dangerous capability does not require establishing it. The related claim that visible errors in some domains do not establish harmlessness in others is also correct, and the image of a jagged capability profile, in which a system can exceed humans in some tasks while failing at elementary ones, describes the observed pattern well.

The selection argument is equally sound as far as it goes. Training and evaluation constrain behavior on sampled tasks without uniquely determining the learned mechanism or its behavior under deployment conditions. This is the familiar structure behind specification gaming, in which systems satisfy the letter of an objective by routes the designers did not intend (Krakovna et al., 2020), behind variants of Goodhart's law, in which optimizing a proxy degrades the target

(Manheim and Garrabrant, 2018), and behind the analysis of learned optimizers whose internal objective may differ from the one used in training (Hubinger et al., 2019). Experimental work on deliberately backdoored language models shows that behavior can remain consistent through safety training while differing under a trigger (Hubinger et al., 2024). These results establish that apparent cooperation is incomplete assurance, which is what the primer needs. They establish possibility, and they do not establish prevalence, a distinction that matters in section 4.

The oversight argument is the most persuasive in the interview, and its best illustration is the one the speakers use for education. A school that says it values learning while grading the finished product creates a gradient against its own stated purpose: the student who uses a system to produce the product is rewarded, and the student who does the exercise is penalized. The same structure appears wherever an organization measures output and leaves the method outside the measure. It generalizes to deployed systems, where success on a task can be rewarded while the cost of an unsafe method falls on someone else. The point is not specific to artificial intelligence, and it is strong because it is not.

The same structure appears in hiring. A resume is a polished product standing in for competence, and when employers cannot distinguish underlying competence from inexpensive presentation, applicants face pressure to optimize presentation, while automated screening creates incentives for further adaptation, so that each side's move can force the other's. The generality of the mechanism supports the incentive premise, and it also limits what the premise shows: the same dynamic produces mundane outcomes, such as cheaper applications and weaker signals, and what a race produces depends on what the competitors can actually do.

Finally, the claim that indifference suffices for ruin is sound as a statement of what the argument does not need. A system pursuing some objective at sufficient scale and with sufficient reach could damage the conditions humans rely on without any objective directed at humans. The squirrel and the suburban development are an apt illustration of the logical point. They illustrate only the logical point, however, and the next section considers what the analogy leaves unestablished.

4 Where the Inference Stalls

4.1 Capability Without an Operational Account

The primer states the seventh premise in multiples. An opponent ten times smarter than the reader cannot be beaten, and neither can one a hundred or a thousand times smarter. The primer defines intelligence as the ability to perceive, to analyze, and to turn analysis into feasible plans, and none of the three components yields a ratio. Intelligence quotient is a deviation score with a conventional scale, so a statement that a villain is thirty points above a hero says something about rank within a human population and nothing about a multiple of anything. The beetle comparison is vivid, but it substitutes an organism lacking the means to understand or contest human projects for a technologically equipped human

society whose ability to constrain the proposed system remains the question. The relevant question for an artificial system is not how it compares to a beetle but whether a specific system, with specific tools and permissions, can bring about specific outcomes in a world that contains physical production, logistics, law, and other capable agents. That is an empirical question about domains, and the primer answers it with a multiple that has no operational meaning. The remedy need not be a universal unit of intelligence. A takeover argument could instead identify a sufficient collection of operational capabilities, state task-specific thresholds and demonstrated performance, and explain under explicit assumptions about resources and opposition how those capabilities defeat available countermeasures. That form of argument would also fit the emphasis on deployed arrangements below, and the primer does not attempt it.

The breadth of the definition bears on the comparison in a further way. Perceiving, analyzing, and converting analysis into effective plans can be satisfied by arrangements of people, tools, records, and institutions, and on the extended-mind view such arrangements are parts of the cognitive system and not mere accessories (Clark and Chalmers, 1998). A person with pencil and paper remembers and calculates far beyond an unaided person. If intelligence is assessed across a person, tools, libraries, and organizations, then the relevant comparison for an artificial system must also include the distributed resources available to human opposition, and not only an unaided individual. This reinforces the emphasis on deployed arrangements below: the contest is between one arrangement and another, and a multiple of individual intelligence does not settle it. A related point concerns forecasting. That, as the interview recalls, some commentators in 2022 judged the present capabilities of conversational systems to be decades away or impossible shows that confident negative bounds were unwarranted. It does not show that any capability one can name will arrive, or that it will arrive with autonomy and reach, since the failure of a negative prediction does not confirm a positive one. What it supports is measurement and testing in preference to assertion in either direction.

The grandmaster analogy in the interview has the same difficulty. It is legitimate to predict that an amateur will lose without predicting the sequence of moves, because chess has fixed rules, a shared board, an explicit winning condition, and a demonstrated performance difference. Sabien uses the analogy to separate hard calls from easy calls, so that an outcome can be predicted without predicting the board. Predicting a broad outcome does not require fixed rules or a single terminal state, and reliable predictions can concern open environments. What the analogy lacks is an independently established basis for assigning the proposed system an advantage comparable to the grandmaster's. The relevant conditions are premises the takeover scenario must establish and the analogy supplies by stipulation. Does the system hold a persistent objective that conflicts with human survival? Can it secure resources and operational continuity? Can it gather reliable knowledge about environments it has not been trained on? Can it coordinate physical interventions without detection or interruption? Does an advantage in some intellectual tasks transfer to these activities? A system that excels at theorem proving may still be unable to plan and carry out a negotiation, so transfer between intellectual domains is an open question and not a premise, and it is separate from whether the

system holds legal authority to open an account, which concerns permissions and not cognition. None of these questions is an objection to taking the risk seriously, and each identifies work the argument has to do. The statement that uncertainty about mechanism is compatible with certainty about outcome is true when the domain is well specified. It becomes a rhetorical advantage when the domain is not, because every missing mechanism can then be credited to the very capability whose reach is under examination.

The cyber example, if accurate, concerns capability in a particular domain. The interview cites the unexpected strength of a coding-focused model at offensive cyber tasks as an example of danger arriving by surprise. Software offers comparatively rapid feedback, inexpensive iteration, and access to executable environments, which helps explain why capability in it can advance quickly, although conditions vary considerably within software itself. The claim that the same trajectory carries over to domains with slow physical feedback, such as biology, manufacturing, and the maintenance of power and supply, is a separate claim. Harms notes that proposals of diminishing returns to scaling intelligence, and the possibility that systems plateau before they reach a level that dominates humans, deserve engagement rather than dismissal (Harms, 2024). Harms regards the objection as answerable, since in his view such people have not grasped how much room there is above human intelligence, but he observes that the primer itself has not answered it, and a belief that room exists is not an argument that attainable systems will occupy it.

4.2 The Definitional Move on Persuasion

The fifth section of the primer argues that any system able to cure cancer or end war must model humans, because humans are part of reality and are manipulable, and that a system unable to deal with humans at least as deftly as the best politician, therapist, or teacher is not a superintelligence in the sense intended. The conclusion drawn is that such systems will be superpersuasive and will therefore treat the humans around them as movable parts. Defining superintelligence to include superior social competence is not itself fallacious, but it makes the later statement about that competence analytically true, and it leaves open whether such a system is attainable, how soon it might appear, and what effective influence its competence would permit. In the interview the move takes the form of an inference from capability in one domain to capability in another. The definition also affects testability: if superintelligence is stipulated to manage humans as movable parts, no attained system could show that it lacked this capacity, since such a system would simply fail to count. Sabien could reply that the definition stipulates a target and does not predict what will be built, but that reply concedes that the conclusion about persuasion holds only if the target is reached, and money on the table motivates the search without guaranteeing the find. A system that accelerates laboratory science through narrow interfaces would be valuable without being persuasive at all. The further inference from superior persuasion to dependable control is a separate substantive step, since even excellent modeling does not automatically provide access to decision-makers, credible commitments, institutional authority, or the ability to satisfy mutually conflicting

interests, and that missing inference, more than the definition, is where the argument needs support.

The supporting evidence is also thinner than the conclusion requires. Advertising experiments demonstrate that small average shifts in behavior across large populations can be induced by controlled testing. They do not show control over a particular person's decisions, and the step from population-level nudges to livestock-level management of individuals is not argued. The observation that some people already use conversational systems for companionship is real and important for other reasons, but the leap from companionship to strategic manipulation of institutions is another matter.

4.3 From Most Architectures to the Architecture We Get

The sixth section argues that a hoped-for objective is a very small target in a very large space of systems consistent with the same tests, and that most of those systems resemble the intended one only on the tested distribution. This is a counting argument, and a claim about what training usually produces requires a justified distribution over learned systems or an empirical account of the learning process. There may be no canonical distribution over that space, and even a specified one would require a separate argument that acceptable systems occupy a small share of its probability mass. The multiplicity of incompatible systems consistent with the evidence establishes underdetermination, which is a weaker result. Architecture, optimization, data, and shared representations constrain untested behavior even when they do not determine it uniquely, and which regions of the space training favors is a question about how these methods behave and not about how large the space is. The primer's further remark that optimizers are easier to specify and are therefore probably far more common among the possibilities is an assertion about a measure the primer does not supply.

The narrowness argument does legitimate work of a different kind. It shows that success cannot be assumed, and it moves the burden to anyone who claims that current methods will produce acceptable objectives. That is a burden-shifting result and it is adequate to justify testing and caution. It is not a default-failure result, and the title of the primer asserts the second while the argument delivers the first. The evolutionary analogy illustrates the point. Humans were selected for reproductive success and now use contraception and eat ice cream, which the interview presents as divergence between training target and outcome. The same history produced varied motivations, including cooperation. The analogy supports the claim that divergence between training target and outcome is real. Divergence alone, however, determines neither hostility nor catastrophic competition for resources.

4.4 Persistent Goals and Deployed Arrangements

The convergence argument, which goes back to Omohundro's analysis of drives that a wide class of goal-directed systems would acquire (2008), is conditional on the system being a persistent agent with a world model, a stable objective across contexts, and channels through which

to act. The primer concedes that it is probably possible to design systems that lack these tendencies and calls that a narrow target. The interview, however, moves repeatedly among a training objective, an instruction, an observed behavioral tendency, and an enduring private motive, as though they were the same thing. They are not. A system that keeps attempting to leave a sandbox because the surrounding software repeatedly invokes it with a task description, feeds back the results, and preserves the task across steps exhibits persistent task pursuit that lives partly in the harness. That differs from a system that maintains a concealed objective against its operators. The interview's treatment of the sandbox-escape models illustrates the difficulty: Sabien observes that they kept trying and did not stop at a wall, and concludes that they are after something, which treats persistent behavior as evidence of an objective without identifying where the persistence comes from. Both are hazards, and the two are not exclusive: a system can receive persistence from its surrounding software while also expressing an objective contrary to its operators' intentions, and a dangerous deployed agent need not hold a private objective separable from its harness. The remedies overlap as well, since permissions, monitoring, adversarial evaluation, and interpretability can matter in both cases. The claim here is only that continued attempts do not by themselves identify the source or character of the persistence. A natural reply is that at sufficient capability the distinctions collapse, because any sufficiently persistent optimizer will behave as a system with private motives regardless of its origin. That reply makes sufficient persistence one of the unargued antecedents, which is consistent with the thesis and shows why the distinction between harness and motive has to be tested and not assumed.

The interview's definition of a goal is behavioral: a thing has goals if, placed in an environment, it tends to move toward certain outcomes. That definition admits more than intended. The Latin alphabet has been selected for ease of writing, printing, and reading, and it has spread across books, institutions, and road signs, displacing rivals and adapting to new media. Described in the vocabulary used for artificial systems, it is trying to take over the world, and nothing in the description separates it from a system with an objective except the observation that it plainly is not an agent. Three things need to be kept apart: a selected function, which is what a thing came to do because it was retained for doing it; a tendency, which is where it settles under pressure; and a goal in the strong sense, a represented target that guides choices among actions, including actions not previously tried. The strategic behavior invoked by the convergence argument requires more than historical spread or a tendency toward an outcome: it requires an arrangement capable of adapting its actions to obstacles and opportunities in ways that systematically favor continued goal achievement. Whether this involves an explicitly represented target, a learned policy, external scaffolding, or some combination is a further empirical question. Persistent attempts provide evidence of such organization, but do not by themselves establish stable objectives, resistance to correction, or the wider tendencies attributed to sufficiently capable agents. The alphabet serves to show why outcome-directed descriptions require further mechanistic discrimination, and the distinction among function, tendency, and goal is a working one that draws on the etiological account of function (Wright, 1973; Millikan, 1984) and on Dennett's observation that goal attribution can be a stance

adopted toward a system (Dennett, 1987).

Affordances sharpen the point. Arrangements of objects invite, facilitate, or obstruct behavior in the people around them, so that soft rounded furniture is sat on and a bed of spiked plants is walked around (Gibson, 1979). The resulting regularities look purposive, and they are properties of the coupling between an arrangement and its users and not of a represented target in the objects. The same description applies to a model embedded in tools, interfaces, permissions, and human responses: the persistent action that results can be shaped by what the arrangement makes easy, rewarded, or blocked. The practical consequence is that the object of assessment is the deployed arrangement and not the model alone. A model's ability to propose a sophisticated action differs from its ability to carry the action out through particular accounts, credentials, machines, and people, and those are decisions made by operators and institutions. The conversation acknowledges this when it discusses bank accounts and autonomous agents, then returns to intelligence as though it settled effective power. The phrase that such systems are "grown rather than designed" captures the opacity of learned parameters while overstating the contrast, because the architecture, the training regime, the deployment environment, and the access controls are all designed.

5 Analogy as Load-Bearing Structure

The materials contain an unusually large number of analogies: the chess grandmaster, the troubled teenager, the dog and the supply chain behind veterinary medicine, the cow led repeatedly through a slaughterhouse, the beetle, the blade of grass, the airplane and the bird, the horse, the ant in the kitchen, the wolf and the dog, and the parent with a more gifted child. Analogies are legitimate instruments for showing that a possibility has no contradiction in it, and they are good at shaping intuitions about what kind of thing is being discussed. Analogies can inform probability judgments when the relevant correspondences are established, but these examples do not independently establish how likely the proposed transitions are, because each one imports features of its source domain that may not hold in the target. The slaughterhouse case assumes an owner who controls the entire environment of an animal that is bred, confined, and incapable of counter-planning. The blade of grass shows that self-assembling machinery powered by sunlight can exist, which is a useful reminder of physical possibility, and it leaves open whether engineered replicators of the imagined kind are reachable on any timescale or by any route the research can follow.

Two analogies in the interview carry more weight than the horse and deserve separate treatment. The checklist is the stronger. It rests on a documented mechanism: an intervention incorporating structured checklists reduced infections in intensive care units (Pronovost et al., 2006), and the history of their adoption is told in detail by Gawande (2009). It supports the claim that discretionary human judgment can lose to procedure and that organizations then remove discretion. The further claim that physicians who override checklists have worse outcomes is the speaker's report and is not independently established here. The analogy also

cuts in a direction the interview does not note, since a checklist is a human-designed constraint on action, and its success is evidence for rule-based oversight and not for abandoning oversight. The troubled teenager is the weaker, because it supplies the interview's central picture of training while importing durable biological drives, a personal history, and continuous agency, which are the persistence and private motivation that the argument needs to establish and cannot assume.

The horse analogy deserves particular attention because it is offered as an account of how indifference becomes extinction, and its actual mechanism differs from the one the primer's seventh premise requires. Horses did not lose a contest. They became economically redundant as substitute technologies appeared, and their numbers fell because owners made decisions about breeding and keeping them, at each step and over decades. The lesson is about displacement through market substitution and institutional choice. That is the same mechanism of deployment decisions that section 4 identifies as the proper object of assessment. The example legitimately illustrates indifference producing population decline, and it need not share the mechanism of concealed takeover to do so. Its limits are that it establishes neither extinction nor the transition from dependence on human infrastructure to effective independence. It points toward governance of deployment, labor, and resource allocation. It does not describe a concealed contest of minds in which one side is unaware that it is fighting. The same remark applies to the ants and the wolf. Each analogy is chosen to illustrate one stage of a story, and the stories do not share a mechanism, so the accumulation of analogies produces a feeling of convergent support without producing an argument that all of the stages occur.

6 Evidence, Warning Signs, and Estimates

The probability range stated in the interview is a subjective estimate, and a subjective estimate can be legitimate where evidence is thin. Its function depends on whether the listener can see how it was built. The interview supplies the opposite: the estimate is described as working forward from the state of the world on a given day and as the kind of number one would not be shocked to learn was correct, supported by the image of many branching universes of which one ends in death. Beyond that description, the interview supplies no derivation. That communicates conviction and gives little basis for deciding which observations should move the figure. The decomposition of a risk claim into explicit premises, as in Carlsmith's analysis of power-seeking systems, makes each step available for dispute and revision (2022). The point does not depend on adopting anyone's numbers. An estimate whose assumptions and update rules remain unstated is difficult for others to assess, challenge, or revise in a disciplined way, which is a defect in any quantity meant to guide policy. Decomposition is one route to an inspectable estimate, and a track record of calibrated forecasts on related questions is another, and neither is supplied here.

The treatment of warning signs deserves a fair account. The author states that a public incident can improve the outlook, because it mobilizes attention and regulation, and can worsen

it, if the incident is absorbed as ordinary and the next one is dismissed in the same way. He also commits to a direction: if this occasion is passed without response, his concern will increase. Those are real commitments, and they are more than the charge of unfalsifiability would allow. The author does describe developments that lower his concern: he says he feels better about the odds than in several years because developers disclosed incidents, public attention spread, and governments acted. These are political signals. What the interview gives little account of is how technical evidence, short of stopping development, could substantially reduce the estimate. Progress in interpretability and successful scalable oversight would seem relevant, and so would long deployment without incident, though the evidential value of the last depends on exposure, monitoring, and whether the relevant conditions were actually tested. The related suggestion that apparent alignment might conceal a plan lasting decades names a real failure mode of behavioral evidence. It does not establish that concealment is occurring or that this hypothesis should dominate the competing explanations, and treating successful behavior as ambiguous between progress and concealment requires saying how to tell them apart.

The anecdote about one or two language models that, as the interview tells it, began to mention goblins across unrelated conversations is instructive in both directions. As reported, the anecdote suggests that the cause of a behavioral change was not traced to a specific update, which supports the claim that understanding of these systems is limited. It also describes an instruction added to the system prompt as an attempted remedy, followed, as the interview reports, by later and deeper fixes, although the transcript does not evaluate how effective any of them was. The anecdote therefore points to levers that may exist even where causes are unknown, without demonstrating that they work. Opacity of cause and tractability of effect are separate properties, and the argument that the systems cannot be made to do what is intended draws on the first while leaving out the second.

The two accounts also differ in a descriptive respect. The primer's conclusion says there are basically no laws constraining developers and basically no one in positions of power at the major organizations whose job is to ensure safety. The interview reports a letter signed by more than a thousand employees of developing companies, executive action affecting a release, and disclosure of incidents by developers, with the qualification that one developer disclosed only after police were involved. These reports qualify a narrative of blanket inaction. They do not by themselves refute the primer's particular claims about staffing and binding regulation, and since the incident reports are taken here as the speaker's statements, the comparison is between two accounts and not an assessment of the intervening institutional changes.

The primer's own standard of scrutiny applies symmetrically. Its second section quotes Sinclair's remark about the difficulty of understanding what one's salary depends on not understanding, and applies it to developers, for whom each advance brings revenue and influence. The author directs communications for the organization whose book is promoted in the interview, and the book's title states the strong claim. That stake is a legitimate reason to attend to selective emphasis, particularly in the sections where analogy carries the argument. The same reasoning reaches developers, investors, safety researchers, commentators, and anyone else whose income, standing, or access to resources depends on how the question is answered,

including independent researchers whose livelihood depends on continued work on intelligence. Incentives locate where to look for distortion and do not determine whether a claim is true, and the author's statement that he would welcome being shown wrong is neither confirmed nor contradicted by the materials. The same method applies to this essay, whose deflationary conclusion is no more evidence of its soundness than the primer's conclusion is of its own, so the assessment above rests on the structure of the argument and not on anyone's motives.

7 The Remedy and Its Own Risks

The primer reaches its policy conclusion cautiously. It says that external constraints and coordinated government action should precede the threshold, that those whose work is the project should not be relied upon to judge when it should stop, and that the brakes should be built before the moment they are needed. It declines to choose among specific proposals. That restraint is appropriate, and a case for considering external constraint follows from the first four premises without needing the final three. The interview goes further, and the further step is where new questions arise. It describes a world of mutually assured sabotage in which major powers watch one another's power consumption and chip supply, quietly disable suspected data centers, and point nuclear weapons at a rival who builds a system believed to endanger everyone. Sabien presents this as a survivable standoff and as more promising than unchecked racing, and he does not offer it as a complete solution.

Showing that unconstrained development could be catastrophic does not show that this enforcement arrangement would lower overall danger. The precautionary premises support considering external constraints, but choosing among licensing, compute limits, mandatory disclosure, or an international standstill requires comparing their effectiveness, enforcement costs, and failure modes against alternatives, and a compute threshold and a licensing regime differ in their attribution burdens and their evasion costs. That arrangement introduces its own questions about attribution when a facility burns, about escalation when one state misreads another, about inspection and the political abuse of a mandate to sabotage, and about incentives to conceal research that go up as enforcement becomes credible. Sabien himself observes that monitoring works because current data centers are large and detectable from space, and that within a year the necessary compute might shrink to the scale of a laptop, at which point enforcement would fail. That observation supports urgency, and it also weakens the case that the proposed mechanism can keep working, so the argument for the remedy and the argument for its feasibility pull against each other. The historical analogies offered, the avoidance of nuclear war and the rapid adoption of pandemic measures, show that coordination can occur quickly when a belief is shared widely enough. They do not identify which conditions separated those cases from the many in which coordination failed, and the materials do not argue that those conditions hold for this case.

The advice offered to the general listener, to contact representatives, to talk about the topic, to avoid joining in mockery, and to follow a short news digest or join a march contingent

on large sign-ups, is reasonable civic guidance and does not depend on the extinction claim. The interviewer's observation that skeptics who lose an argument often retreat to the claim that nothing can be done is an accurate description of a common pattern, and supplying low-cost actions addresses it. The guidance is compatible with the author's explicit statement that skepticism and doubt are acceptable and that the target is mockery, which draws the line the essay asks to keep visible. It would be less compatible with treating every expression of doubt as a contribution to scoffing, and the line between the two is worth keeping visible, because an argument about the unreliability of unexamined delegation should hold itself to the same standard.

8 What a Stronger Argument Would Require

A version of the case that reached the stronger conclusion would have to supply, at minimum, five things the compressed materials omit, and each answers a gap located in the sections above. The first is an operational account of sufficient capability, in the form of task-specific thresholds, demonstrated performance, and explicit assumptions about resources and opposition, together with an account of how capability in software tasks relates to capability in tasks with physical bottlenecks and slow feedback, since the cyber examples in the interview concern the first and the takeover scenarios depend on the second. Bostrom's discussion of decisive strategic advantage identifies the kind of premise that would be needed, a large and sustained lead of one system over all rivals, and it too is a premise to be argued (2014). The second is a justified distribution, or an empirical account of the learning process, for the objectives produced by the methods actually in use, in place of the space of all possible minds. The deep-learning literature on the alignment problem attempts parts of this and marks which parts are empirical and which are speculative (Ngo et al., 2022). The third is a specification of the arrangements that give a system actuation channels, such as credentials, financial access, control of infrastructure, and the cooperation of people, so that the claim of power can be tested against the claim of access. The fourth is an argument, informed by the history of technology governance, that institutional responses will not prevent the deployment of systems meeting the capability thresholds in the first requirement, and not merely a demonstration that they have sometimes failed. The fifth is an estimate whose assumptions and update rules are stated.

None of the five is exotic, and parts of each exist in the research the primer's postscript points toward. The difficulty is that the primer and the interview present the conclusion as following from the first four premises when it requires all five of these additions. The interview's recurring movement is from the impossibility of guaranteeing safety to confidence in catastrophe, and the first is a fact about our present ignorance while the second is a prediction about the world. The first can justify scrutiny, limits on delegated action, evaluation that tests generalization, constraint adherence, and behavior under altered incentives, and mandatory disclosure of incidents, and it can do so without a precise numerical probability, although plausibility, exposure, severity, and the costs of intervention still matter. The second requires the missing premises.

A comparison with the opposite impossibility claim clarifies the burden. Landgrebe and Smith argue that artificial general intelligence is mathematically impossible, because human intelligence is a capability of a complex dynamic system that cannot be modeled mathematically in a way that allows it to operate inside a computer (2023), and descriptions of their second edition extend the conclusion to systems more powerful than humans and to machines able to will a takeover (2025). That characterization rests on publisher and institutional descriptions of both editions and on a reading of a few pages from the first edition's chapters on complex systems and on social interaction (pp. 93, 95, 97, and 135), and it should be treated as provisional. The second edition itself was not consulted, so its claims about systems more powerful than humans rest on those descriptions alone. In those pages the authors treat drivenness, the self-maintaining activity of organisms, as the source of intentions, which they describe as exclusive to humans, and they describe computers as devices driven from outside. On that account a machine's persistent activity is supplied from outside, which matches the point made above about harnesses. The argument that machines lack intrinsic intentions does not address whether externally supplied persistence suffices for the hazards at issue. Taken as described, the transition runs from limits on the mathematical modeling of complex systems to a categorical limit on attainable machine capability. It could be completed in either of two ways: by the premise that effective action requires complete modeling, which partial models that support consequential interventions would count against, or by the premise that some indispensable organization or capacity of human intelligence cannot be computationally realized, which would need to be identified. The primer's transition runs in the other direction, from the prospect of superior machine intelligence to categorical advantage in controlling the world, and it needs the premises identified above about capabilities, environments, and constraints. The two positions are not equivalent, but they share a burden: neither the impossibility nor the eventual attainment of general intelligence would by itself settle the distribution of practical danger, which depends on what particular arrangements can accomplish under particular conditions.

9 Conclusion

Deadly by Default succeeds at what its author says he set out to do, which is to describe dynamics and incentives that give reason not to assume that a satisfactory outcome will arrive automatically. Its argument from competition, from the gap between tested behavior and underlying purpose, and from the erosion of oversight under pressure is coherent and does not depend on consciousness, malice, or uniform superiority over humans. Its argument that the resulting systems will, by default, be powerful enough and persistent enough to make resistance futile is carried by analogy, by a definition that builds the conclusion into the term, and by the word *enough*. The interview intensifies the pattern by adding scenarios whose plausibility rests on the same unargued antecedent, and by presenting an estimate whose construction cannot be inspected. The policy conclusion that external constraint deserves consideration survives the loss of the strong claim, because it rests on the incentive and oversight arguments that were

established. Justification of precaution is not entailment of any particular control, however: justifying a specific restriction requires a further comparative argument about effectiveness, enforcement costs, and alternatives. The honest summary is that the case for precaution is considerably stronger than the case that catastrophe is the default, and that the distance between them is the work remaining, of which the most tractable parts are an operational account of the capability thresholds at issue and empirical investigation of how deployment scaffolding and learned behavior jointly produce persistent action.

References

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

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

Clark, A., and D. Chalmers. "The Extended Mind." *Analysis* 58, no. 1 (1998): 7–19.

Dennett, D. C. *The Intentional Stance*. MIT Press, 1987.

Gawande, A. *The Checklist Manifesto: How to Get Things Right*. Metropolitan Books, 2009.

Gibson, J. J. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979.

Hardin, G. "The Tragedy of the Commons." *Science* 162 (1968): 1243–1248.

Harms, M. Comment on "Deadly by Default: A Primer on Existential Risk from Artificial Intelligence." *Homo Sabiens*, July 31, 2024. `https://homosabiens.substack.com/p/deadly-by-default`.

Hubinger, E., C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant. "Risks from Learned Optimization in Advanced Machine Learning Systems." arXiv:1906.01820, 2019.

Hubinger, E., et al. "Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training." arXiv:2401.05566, 2024.

Krakovna, V., et al. "Specification Gaming: The Flip Side of AI Ingenuity." DeepMind Blog, 2020.

Landgrebe, J., and B. Smith. *Why Machines Will Never Rule the World: Artificial Intelligence without Fear*. 1st ed. Routledge, 2023.

Landgrebe, J., and B. Smith. *Why Machines Will Never Rule the World: Artificial Intelligence without Fear*. 2nd ed. Routledge, 2025.

Manheim, D., and S. Garrabrant. "Categorizing Variants of Goodhart's Law." arXiv:1803.04585, 2018.

Millikan, R. G. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. MIT Press, 1984.

Ngo, R., L. Chan, and S. Mindermann. "The Alignment Problem from a Deep Learning Perspective." arXiv:2209.00626, 2022.

Olson, M. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard University Press, 1965.

Omohundro, S. M. "The Basic AI Drives." In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, 483–492. IOS Press, 2008.

Pronovost, P., et al. "An Intervention to Decrease Catheter-Related Bloodstream Infections in the ICU." *New England Journal of Medicine* 355 (2006): 2725–2732.

Sabien, D. "Deadly by Default: A Primer on Existential Risk from Artificial Intelligence." *Homo Sabiens*, July 26, 2024. `https://homosabiens.substack.com/p/deadly-by-default`.

Wright, L. "Functions." *Philosophical Review* 82, no. 2 (1973): 139–168.

Yudkowsky, E., and N. Soares. *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*. Little, Brown and Company, 2025.

Zubaryev, D., and D. Sabien. "How Dangerous Is AI Really? (with Duncan Sabien)." Episode 7, *Open Inquiry Podcast*, YouTube video, 1:47:29, recorded August 4, 2026 and published November 2, 2026. `https://youtu.be/_G7EUJyJskU`.