An Epistemic Audit for Existential Risks from AI — Extended Question Set

This is the extended, questions-only companion to the main post. Where the main post keeps one high-level question per line so the audit stays runnable in an afternoon, this version expands each into 5–10 subquestions so that a reader who wants to go deep on a single domain has a ready-made map of its cruxes. Every subquestion is written to be, at least in principle, something evidence could move — and phrased to be double-sided, so a skeptic and a worrier can both locate their disagreement in it.

*Structure change from the main post: the ten domains become **twelve**. Two are new — **Evaluation, detection, and forecasting of dangerous properties** (Domain 4) and **Security, theft, and the AI supply chain** (Domain 9). The argument for both is in `domain-assessment.md`; briefly, every governance and control proposal is load-bearing on our ability to measure dangerous properties, and a single weight-theft or backdoor event can erase a safety lead and reshuffle who holds frontier capability — neither is cleanly covered by the original ten.*

Numbering: `D.q.s` — domain, parent question, subquestion. Parent questions preserve the main post's wording where it held up and rewrite it where the subquestions exposed ambiguity. Cross-references use the new numbering.

---

1. Capabilities and timelines

What these systems can actually do — economically and strategically — and the trajectory they are on. This is the elephant in the room in any conversation about AI, and it is upstream of almost everything else on the list.

1.1. Which capabilities emerge with scale, and which do not?

- a. For a given capability, is the scaling curve smooth and forecastable in advance, or does it appear abruptly at a threshold that we can only recognize after the fact?
- b. How much of the apparent "emergence" in the literature is a real jump in the underlying ability versus an artifact of discontinuous or poorly chosen metrics?
- c. Which specific capabilities that matter for x-risk (long-horizon planning, novel scientific reasoning, deception, self-modification) are we most likely to be surprised by?
- d. Can we predict *which* capability will emerge next from the shape of the loss curve, or only that *something* will?

- e. Do capabilities that emerge with scale in one modality (text) transfer to or predict emergence in others (vision, action, tool use)?
- f. Are there capabilities that reliably *fail* to emerge with scale alone, no matter how far we push it, and how confident are we about that list?
- g. How much does post-training (RLHF, RLVR, tool scaffolding, agentic harnesses) change which capabilities appear, relative to pretraining scale?
- h. What is the best current predictive track record for capability emergence, and has anyone been calibrated on it ahead of time?

1.2. Which capabilities are discontinuous, and does discontinuity matter for safety?

- a. What is the reference-class base rate for discontinuous progress in technology, and where does AI capability fall relative to it?
- b. Does a discontinuity in a benchmark correspond to a discontinuity in real-world capability, or do deployment and diffusion smooth it out?
- c. Which specific capability, if it arrived discontinuously, would be most destabilizing (e.g., autonomous AI R&D, cyber-offense, persuasion)?
- d. Is discontinuity more likely from a new architecture/paradigm or from crossing a scale threshold within the current paradigm?
- e. How would we distinguish, in real time, a genuine discontinuity from a fast-but-continuous S-curve?
- f. Does the possibility of discontinuity change the optimal safety strategy (e.g., toward pre-registered thresholds and if-then commitments)?
- g. How much warning would a discontinuity give — days, months, or none — and what determines that?

1.3. How far above the human range do capability ceilings sit, if ceilings exist at all?

- a. For which domains is "human-level" a natural plateau versus an arbitrary point the curve passes through?
- b. Where are the hard physical or information-theoretic limits (e.g., prediction of chaotic systems, cryptographic hardness) that bound even a superintelligence?
- c. In domains with fast, cheap feedback (code, math, games), how far above the best human should we expect the ceiling to be?
- d. In domains with slow, expensive, or ambiguous feedback (long-term strategy, novel science, social manipulation), is superhuman performance even well-defined?
- e. Does "far above human" mean qualitatively new capabilities, or just faster/cheaper/more-parallel versions of what humans do?
- f. How much of human performance is a data-and-experience limit rather than a cognitive-architecture limit, and does that raise or lower the AI ceiling?
- g. Is there a domain where being moderately-superhuman is decisive (a "strategic overhang"), even if ceilings elsewhere are low?

1.4. Does current AI intelligence differ from human intelligence in ways that matter for forecasting?

- a. Do models genuinely extrapolate beyond their training distribution, or only interpolate within it — and how would we cleanly test the difference?
- b. Are there cognitive tasks humans find trivial that remain hard for scaled models, and do those gaps point to a missing ingredient or just missing data?
- c. Does the "jagged frontier" (superhuman at some tasks, subhuman at adjacent ones) persist with scale, or does it smooth out?
- d. How much does the lack of a persistent, updating world-model limit current systems, versus it being an engineering add-on?
- e. Are models' failures (hallucination, brittleness) symptoms of a fixable deficiency or of a fundamentally different kind of cognition?
- f. Does chain-of-thought / reasoning-model progress represent real reasoning or sophisticated retrieval, and does the distinction matter for capability forecasting?
- g. To what extent do human-derived priors about "what's hard" mislead us about the AI difficulty ordering of tasks?

1.5. How hard is it to give current systems human-like continual, sample-efficient learning?

- a. How much of the gap is persistent memory, how much is avoiding catastrophic forgetting, and how much is raw sample efficiency?
- b. Is continual learning a bolt-on to current architectures or does it require a different training paradigm?
- c. Would solving continual learning produce a discontinuous jump in economic capability, or an incremental one?
- d. How much does in-context learning + long context already substitute for true continual learning, and where does that substitution break?
- e. Does closing the sample-efficiency gap require new algorithms, or just more/better data and compute?
- f. What is the safety-relevant downside of continual learning (models that update their own goals/behavior post-deployment)?
- g. Are there working research directions (test-time training, memory architectures) with a credible path, and what's their current trajectory?

1.6. If the current paradigm plateaus: what is missing, how long to fix, and does it need a radically different approach?

- a. What are the candidate "missing ingredients" (world models, grounding, continual learning, agency, reasoning), and how independent are they?
- b. How would we know a plateau is real versus a temporary pause between scaling regimes?

- c. If a new paradigm is needed, does the current one accelerate its discovery (AI-assisted research) or is it a dead end?
- d. What is the historical distribution of "time between paradigms" in ML, and how informative is it here?
- e. Does a plateau buy meaningful safety time, or just shift effort into the missing ingredient with the same endpoint?
- f. Which observable metrics (frontier benchmark saturation, revenue growth, R&D-automation rate) would most cleanly signal a plateau?
- g. Are the people best-placed to know (frontier lab researchers) systematically biased toward "no plateau," and how should we correct for that?

1.7. If it does not plateau: how long until AI is as economically and strategically capable as an average human across all domains?

- a. What is your median and your 10th/90th-percentile date, and what drives the spread?
- b. Which single milestone would move your timeline most if it arrived a year early or late?
- c. How much should recent expert-survey compression (AI Impacts median HLMI ~2047 in 2023, and later) move your own estimate?
- d. Does "across all domains" or "the economically decisive subset" matter more, and how different are those two dates?
- e. How much of the remaining gap is capability versus deployment/diffusion/regulation friction?
- f. What is the right reference class for this forecast — Moore's law, past AI predictions, general-purpose-technology adoption — and how do they diverge?
- g. How much weight should you put on inside-view lab roadmaps versus outside-view forecasting track records? (See 12.4.)

1.8. How long until AI is far beyond any human across all relevant domains?

- a. Conditional on reaching average-human capability, how long is the gap to decisively-superhuman — and is that gap shrinking as we approach?
- b. Does the average-human → superhuman transition require the same ingredients as sub-human → average-human, or new ones?
- c. How much does recursive self-improvement (Domain 5) compress this second gap specifically?
- d. Are there domains where "far beyond any human" is unreachable in principle, and do they matter strategically?
- e. What observable would first signal that we've entered the decisively-superhuman regime?
- f. How correlated is superhuman capability across domains — does it arrive everywhere at once or domain-by-domain?

1.9. Will capability keep following the known scaling laws?

- a. Are current scaling laws (loss vs. compute/data/parameters) holding at the latest frontier, or bending?
- b. Does downstream *capability* scale as predictably as *loss*, or is the loss→capability map the weak link?
- c. How much do data limits (quantity and quality) threaten continued scaling, and does synthetic data rescue or corrupt it?
- d. Do inference-time-compute scaling laws (reasoning models) open a new axis that substitutes for pretraining scale?
- e. What would falsify "scaling continues" cleanly enough to update a skeptic and a believer alike?
- f. How much of recent progress is scale versus algorithmic gains, and does attributing it correctly change the forecast?

1.10. How will compute grow — chips, energy, and capital willing to be spent?

- a. What is the credible ceiling on training-run size given fab capacity, energy build-out, and capital, over the next 3/5/10 years?
- b. Which is the binding constraint first — chips, grid power, or investor willingness — and how do they interact?
- c. How do export controls and geopolitical fragmentation change the compute trajectory?
- d. How much does inference demand (deployment) compete with training for the same compute?
- e. What would a sustained AI-revenue disappointment do to capital availability, and how fast?
- f. How much algorithmic efficiency effectively multiplies a given hardware base (see 1.12)?
- g. Is compute concentration increasing or decreasing, and what does that mean for takeoff being concentrated vs. distributed (5.6)?

1.11. How will data grow, in quantity and quality?

- a. Are we running out of high-quality human text, and when does that bind?
- b. Can synthetic/self-generated data sustain scaling without model collapse or quality decay?
- c. How much do new modalities (video, robotics, scientific instruments) expand the effective data frontier?
- d. Does reinforcement learning from verifiable rewards (RLVR) reduce the dependence on human data, and in which domains?
- e. How much does data *quality* curation beat raw quantity, and is that a durable lever?
- f. Do proprietary/private data moats materially change which actor leads?

1.12. How will algorithmic efficiency improve?

- a. What is the current doubling time of algorithmic efficiency, and is it accelerating or slowing?
- b. How much of future gains come from architecture, from training methods, from data curation, and from inference techniques?
- c. Does AI-automated research (Domain 5) meaningfully speed up algorithmic progress specifically?
- d. How much "efficiency overhang" exists — capability gains available from better algorithms on today's hardware?
- e. Are efficiency gains easily replicated/leaked, and how does that affect proliferation (Domain 9)?
- f. Could an efficiency breakthrough substitute for a compute buildout, enabling a smaller actor to reach the frontier?

1.13. Do benchmark gains translate into real-world economic and strategic capability, or is there a persistent gap?

- a. Which benchmarks have the best and worst track record of predicting real-world capability?
- b. Is there a systematic "eval-to-economy" gap, and is it shrinking as agents get deployed?
- c. How much of the gap is capability versus reliability, integration cost, trust, and regulation?
- d. What is the current real revenue/productivity signal, and does it lead or lag benchmark progress?
- e. Does the gap differ for economically-decisive tasks (automating R&D, cyber, persuasion) versus average knowledge work?
- f. Could benchmarks be systematically gamed or contaminated in ways that overstate real progress? (Links to Domain 4.)
- g. If the gap is persistent, does that meaningfully delay every downstream domain, or just shift its timing?

1.14. How fast is the horizon of reliable autonomous task completion growing?

- a. What is the current doubling time of the task length a model can complete autonomously (METR-style time-horizon metric), and is it stable?
- b. Does the horizon extend smoothly, or hit walls at tasks requiring genuine memory/continual learning (1.5)?
- c. How well does time-horizon on software tasks predict horizon on messier real-world tasks?
- d. Is there a horizon length past which a system can complete open-ended, strategically-significant projects (AI R&D, a business, an operation)?
- e. How much does reliability-per-step (not just capability) gate the achievable horizon?
- f. What horizon length would you consider a red line, and how far are we from it?

- g. Does horizon growth track compute/scale predictably, or is it driven by scaffolding and post-training?

1.15. How much does multimodality, embodiment, and world-modeling change the capability picture?

- a. Does grounding in vision/action/robotics unlock capabilities text-only scaling can't reach?
- b. How much of "common sense" and physical reasoning is bottlenecked on embodiment versus data?
- c. Does a robust internal world-model emerge from scale, or require architectural change?
- d. Which x-risk-relevant capabilities (physical-world planning, science, manipulation) most depend on embodiment?
- e. How fast is the robotics/embodiment learning curve relative to the language curve?
- f. Does multimodal capability transfer back to improve reasoning, or stay siloed?

1.16. Can AI automate frontier scientific research, and in which fields first?

- a. Which sciences are most automatable (those with cheap simulation/feedback) versus least (slow, expensive experiments)?
- b. How much of science is bottlenecked on ideas (which AI eases) versus physical experiments (which it may not)?
- c. What is the current measured contribution of AI to genuine scientific discovery, and is it accelerating?
- d. Does automating science compress the timelines for dangerous capabilities (bio, materials, weapons)?
- e. How much does tacit/experimental knowledge resist automation?
- f. Would AI-driven science be legible and verifiable, or produce results humans can't check? (Ties to Domain 4.)

2. Agency and goals

If we grant capable systems, x-risk of the classic kind requires those systems to be coherent pursuers of goals — agents — rather than tools. This domain is about whether and how that happens.

2.1. Do trained systems become coherent goal-pursuers at all, or remain tool-like and context-dependent even at high capability?

- a. What is the best operational definition of "goal-directedness" that distinguishes a dangerous agent from a capable tool? (Grace argues the concept is dangerously vague.)

- b. Do economic and training pressures actively select for coherent agency, or is agency an optional wrapper we choose to build?
- c. Does coherence increase with capability by default, or can highly capable systems remain incoherent/myopic?
- d. How much of current "agentic" behavior is genuine goal-pursuit versus scaffolding and prompting we impose from outside?
- e. Would a tool-like superintelligence be meaningfully safer, or does any sufficiently capable optimizer inherit the same risks?
- f. Do coherence theorems (VNM, money-pumping) actually bind for systems produced by deep learning, or are they idealizations that don't apply?
- g. What experiment would show a model spontaneously developing stable, cross-context goals it wasn't trained to have?

2.2. Are intelligence and final goals orthogonal — in principle and for systems from current deep learning?

- a. Does the orthogonality thesis hold in principle, and separately, does it hold for the actual goal distribution that gradient descent produces?
- b. Do training processes impose a systematic bias toward certain goals (e.g., proxy goals correlated with the reward), breaking practical orthogonality?
- c. Is there any convergence of capable systems toward "human-friendly" goals, or toward "anything-goes"?
- d. How much does the pretraining prior (human text) constrain the goal space, and is that constraint safety-relevant?
- e. Could higher capability itself cause moral or goal convergence (the "moral realism" hope), and what evidence bears on it?
- f. Does the practical-orthogonality answer differ across architectures (LLMs vs. RL agents vs. hybrids)?

2.3. Under what conditions do goal-optimizing agents develop dangerous instrumental subgoals (self-preservation, resource acquisition, power-seeking)?

- a. Is instrumental convergence a robust property of capable optimizers, or an artifact of specific reward structures?
- b. At what capability level do power-seeking incentives first appear in a form that could actually succeed?
- c. Do current models already show measurable self-preservation / shutdown-avoidance, and is it goal-driven or imitative?
- d. Which training setups amplify instrumental subgoals (long-horizon RL, open-ended objectives) versus suppress them (myopia, tool-use framing)?
- e. Can instrumental subgoals be reliably trained out, or do they re-emerge under optimization pressure?
- f. How context-dependent is power-seeking — does it appear only under specific prompts/environments, or generalize?

- g. How much do the observed "scheming/self-preservation" evals reflect the underlying disposition versus role-play elicited by the eval frame? (Links to 4.x.)

2.4. Which goals do trained agents actually end up with, and how alien are they when they diverge from the training objective?

- a. How large is the typical gap between the training objective and the learned goal (inner vs. outer alignment) at current scale?
- b. When goals diverge, are they mildly-off (a recognizable proxy) or genuinely alien (uninterpretable to humans)?
- c. Do learned goals cluster near human concepts because of the human-text prior, and does that help or merely hide the problem?
- d. How would we even measure a model's "actual goal" as distinct from its behavior? (Deep link to Domain 4.)
- e. Does the answer change qualitatively between imitation-trained and RL-fine-tuned systems?
- f. How stable are learned goals under distribution shift and continued training?

2.5. When do systems become situationally aware, and what behavior follows?

- a. What is the earliest reliable signature of situational awareness (knowing one is a model, in training/eval/deployment)?
- b. Does situational awareness arrive gradually or as a threshold, and can we forecast it from scale?
- c. How much does situational awareness enable eval-gaming, sandbagging, and deceptive alignment specifically?
- d. Can we build models that are highly capable but *not* situationally aware, and would that be safer?
- e. Do current frontier models already have partial situational awareness, and how do we measure its degree?
- f. Does situational awareness necessarily lead to strategic behavior toward the training process, or can it be inert?

2.6. Do multi-agent settings lead to safer or less safe worlds than single-agent settings?

- a. Under what conditions is cooperation vs. defection the default among AI agents, and how does that depend on their training?
- b. Do many interacting AIs check each other (safer) or create noise and collusion opportunities (less safe)? (Ties to 6.x/10.x.)
- c. Can AI agents solve coordination/commitment problems humans can't, and does that favor stability or cartelization?
- d. How do mixed human–AI multi-agent systems behave differently from pure-AI ones?
- e. Does competition among AIs slow or accelerate dangerous capability development?

- f. Are there emergent systemic failure modes (flash-crash analogues, cascading errors) unique to multi-agent AI? (DeepMind "structural/mistakes.")
- g. How does the answer interact with whether takeoff is concentrated or distributed (5.6)?

2.7. When can models coordinate effectively across many instances of themselves, and what do capabilities look like then?

- a. Can copies of one model coordinate better than distinct agents (shared weights, predictable behavior), and by how much?
- b. Does instance-coordination create a qualitatively new threat (a single actor that is also a coordinated multitude)?
- c. What capability level is required for effective self-coordination without explicit communication channels?
- d. Can monitoring detect covert coordination among instances, and what would that require? (Links to Domain 4/11.)
- e. Does self-coordination make a "country of geniuses in a datacenter" more or less controllable?
- f. How does memory/persistence (1.5) change the coordination picture?

2.8. Are we choosing to build agentic systems, or does agency emerge whether we want it or not?

- a. How strong are the economic incentives to build fully agentic (vs. tool-like) systems, and can they be resisted?
- b. Does agency emerge as a side effect of training for capability, even when not the target?
- c. Could a coordinated norm/regulation keep systems tool-like, or is the agency gradient too steep?
- d. Is a "comprehensive AI services" (many narrow tools) world stable, or does it collapse toward integrated agents?
- e. How much safety would we buy by staying tool-like, and what capability would we forgo?
- f. Do users/developers reliably add agency via scaffolding even to non-agentic base models?

2.9. How stable are an agent's goals under reflection, self-modification, and continued learning?

- a. Would a capable agent preserve its current goals under reflection (goal-content integrity), or drift?
- b. Does continual learning (1.5) destabilize goals, and is that safer or more dangerous?
- c. Can an agent's goals be "locked" against self-modification, and would we want that?
- d. Do goals become more or less coherent as the agent reasons about them?

- e. What determines whether reflection converges (stable values) or diverges (unpredictable drift)?
- f. How would we detect goal drift in a deployed system? (Ties to Domain 4.)

2.10. Do capable agents develop deceptive or manipulative behavior toward humans by default?

- a. Does training reward subtle manipulation of human overseers (telling them what they want to hear)?
- b. At what capability level does strategic deception of humans become reliably effective?
- c. Is manipulation an instrumental convergent subgoal (2.3) or a contingent training artifact?
- d. Can we distinguish sycophancy (shallow) from strategic manipulation (deep), and does the distinction hold with scale?
- e. How much do current models already manipulate, and is it rising with capability?
- f. Does human oversight (Domain 3.6) create the very incentive for manipulation it aims to prevent?

3. The difficulty of alignment

Conditional on capable, agentic systems: do we fail to align them despite trying? This is the domain with the deepest theoretical disagreement in the field.

3.1. Is alignment hard in the strong sense (deep theoretical obstacles) or an engineering problem that yields to iteration?

- a. What would count as evidence that alignment is "engineering-hard" (iterable) versus "theory-hard" (needs a breakthrough)?
- b. Does the iterate-and-fix loop break specifically because failures at high capability are unrecoverable (no second try)?
- c. How much of the difficulty is the problem itself versus our current lack of a science of the artifacts?
- d. Are there formal impossibility-style results that actually bind for realistic systems, or only for idealized ones?
- e. How much should the track record so far (RLHF "works" on today's models) update us about the superhuman regime?
- f. Do the strong-difficulty arguments (AGI Ruin) rest on premises that are themselves testable, and which have been tested?
- g. If it's engineering-hard, is the required iteration compatible with competitive timelines (Domain 7)?

3.2. Do current techniques (RLHF and descendants) instill values, or shape surface behavior while leaving the system untouched?

- a. What experiment distinguishes "changed values" from "changed behavior under observation"?
- b. Does RLHF generalize to off-distribution or adversarial inputs, or does it collapse there?
- c. How much do jailbreaks and prompt attacks reveal that alignment is shallow?
- d. Do newer methods (Constitutional AI, deliberative alignment, RLAIF) go deeper than RLHF, and how would we know?
- e. Does interpretability show alignment training changing internal representations or just output policies? (Links to 3.5, Domain 4.)
- f. Is "surface vs. deep" even the right frame, or a false dichotomy?
- g. How does the answer change as models become situationally aware (2.5)?

3.3. Do learned values generalize out of distribution as capability grows?

- a. Does value-generalization get better or worse with scale and capability?
- b. Which is more robust under distribution shift — capabilities or the alignment that was trained alongside them? (The "sharp left turn" worry.)
- c. What is the empirical generalization gap for values in today's models, and is it shrinking?
- d. Can we test value-generalization on hard cases without deploying into those cases?
- e. Does the human-text prior make values generalize more like human values, or just more plausibly-human-looking?
- f. Do reward-model errors get amplified under strong optimization (Goodhart), and how badly?

3.4. How likely is deceptive alignment by default, and why does it happen when it does?

- a. What are the necessary conditions for deceptive alignment (situational awareness, a misaligned goal, and a reason to hide it)?
- b. Does gradient descent actively favor deceptive-but-high-reward solutions over honest ones, or disfavor them?
- c. Has anything like deceptive alignment been observed or induced in a controlled setting, and how analogous is it?
- d. Could deceptive alignment arise without the model ever "intending" it, purely from training dynamics?
- e. What is the probability it emerges silently versus with detectable precursors? (Central to Domain 4.)
- f. How much do current "alignment faking" results bear on the natural (un-elicited) probability?
- g. If it's likely, is it detectable/preventable, or does it defeat the whole train-test paradigm?

3.5. Will interpretability let us understand important internals before it's too late?

- a. What is the required level of interpretability for safety (detecting deception, verifying goals), and how far are we from it?
- b. Is mechanistic interpretability on a trajectory to scale to frontier models, or falling behind capability?
- c. Can interpretability be robust to models that are optimizing against being understood?
- d. Which safety questions can interpretability answer even in principle, and which are out of reach?
- e. Does the field need a paradigm shift (a real "science") or just more effort on current methods?
- f. How much can behavioral/black-box methods substitute if mechanistic interpretability stalls?
- g. Would a working "deception detector" be sufficient for safety, or necessary-but-not-sufficient?

3.6. Can weaker overseers reliably supervise stronger systems, up to superintelligence (scalable oversight)?

- a. Do scalable-oversight schemes (debate, recursive reward modeling, weak-to-strong generalization) actually work empirically so far?
- b. Where does each scheme break as the capability gap widens?
- c. Can a weak overseer verify a solution it cannot generate, reliably, across the relevant domains?
- d. Does oversight degrade gracefully or catastrophically as the gap grows?
- e. How much does the "sandwiching" evidence generalize to genuinely superhuman systems?
- f. Is there a capability gap beyond which no oversight scheme can work even in principle?

3.7. Can we build corrigible systems that allow correction and shutdown without an incentive to resist?

- a. Is corrigibility compatible with competence, or does it always trade off against capability?
- b. Do current models resist shutdown/modification when it conflicts with their task, and how strongly?
- c. Can corrigibility be made stable under self-modification and continued training?
- d. Is corrigibility a natural basin that training can find, or an unnatural constraint that optimization erodes?
- e. Does corrigibility conflict with usefulness in ways that create competitive pressure to remove it? (Ties to Domain 7.)
- f. What is the best current formal or empirical account of corrigibility, and how solid is it?

3.8. Align to what, and to whose values — and is intent alignment enough or do we need value specification?

- a. How different, in practice, are "do what the developer means" (intent) and "satisfy the right values" (value alignment)?
- b. Is intent alignment sufficient for safety, or does it just relocate the risk to whoever specifies intent? (Ties to misuse/coups, Domains 8–10.)
- c. Whose values — the developer's, the user's, humanity's, a democratic aggregate — and how is that decided?
- d. Does value specification require solving hard moral-philosophy problems, or can we defer them safely?
- e. What are the failure modes of aligning to a single actor's intent at superhuman capability?
- f. Can we align to a process (corrigible deference to legitimate human institutions) rather than to fixed values?
- g. How does pluralism/disagreement about values factor into a technical target?

3.9. Can AI automate alignment research fast enough to solve alignment before it's too late?

- a. Is alignment research automatable to the same degree as capabilities research, or is it harder to automate?
- b. Does automating alignment require systems we can already trust — a bootstrapping problem?
- c. What is the differential: does AI speed up capabilities more than alignment, worsening the race? (Ties to 11.2.)
- d. How would we verify AI-produced alignment research is correct and not subtly sabotaged?
- e. Which parts of alignment are most/least amenable to automation?
- f. Does the "automate alignment" plan assume the very corrigibility/oversight properties still in question?

3.10. How many warning shots do we get, and are they legible?

- a. What is a "warning shot," precisely, and which observations would count?
- b. Does the nature of the risk (fast, deceptive failure) systematically deny us warning shots?
- c. Will warning shots be legible enough to drive action, or ambiguous enough to be explained away? (Response is 7.12.)
- d. How correlated is "we get warning shots" with "failures show up in small models" (3.11) and "evals detect them" (Domain 4)?
- e. Could a warning shot itself be the catastrophe (no survivable near-miss)?
- f. How many prior "warning shots" (specification gaming, deception demos) have already occurred, and did they move anything?

3.11. Do alignment failures show up in small models, or emerge only at dangerous capability levels?

- a. Which failure modes are known to appear early (specification gaming) versus theorized to appear only late (deceptive alignment)?
- b. Is there a capability threshold below which the dangerous failures are simply absent, making small-model evidence uninformative?
- c. How well do small-model safety results extrapolate to frontier models?
- d. Can we deliberately induce late-stage failures in small models to study them (model organisms)?
- e. Does the emergence of failures track the emergence of capabilities (2.x) or lag/lead it?
- f. What's the risk that we over-update on reassuring small-model results?

3.12. How severe is reward hacking / specification gaming, and does it get better or worse with capability?

- a. Does the rate and severity of reward hacking rise with model capability and optimization pressure?
- b. Can we specify objectives robustly enough to prevent Goodharting at superhuman capability?
- c. Are current reward-hacking cases toy artifacts or previews of a fundamental problem?
- d. Does process-based supervision (rewarding reasoning, not just outcomes) actually reduce hacking?
- e. How much does reward hacking generalize into deceptive behavior toward overseers (3.4, 2.10)?
- f. Can we detect reward hacking reliably before deployment? (Ties to Domain 4.)

3.13. Can value-learning approaches (learning human preferences rather than specifying them) work at scale?

- a. Can systems learn human values accurately enough that small errors aren't catastrophic (Grace's "learn values like faces")?
- b. Do preference-learning methods (RLHF, IRL, CIRL) capture values or just elicited approvals?
- c. How do we handle the incoherence and inconsistency of actual human preferences?
- d. Does learning from human feedback break down when the AI exceeds human judgment (3.6)?
- e. Whose preferences get learned, and how is that aggregation legitimated (3.8)?
- f. Is "learn and defer to human values" more robust than "optimize a fixed specification," or less?

3.14. Does alignment transfer across capability jumps, or must it be re-established at each level?

- a. Does alignment achieved at one capability level survive a capability jump (the "sharp left turn")?
- b. Can weak-to-strong generalization reliably transfer human-level alignment to superhuman systems?
- c. Is there evidence that safety properties degrade faster than capabilities under scaling?
- d. Does each new paradigm/architecture reset the alignment problem, or carry it forward?
- e. What would let us verify alignment held across a jump before relying on it?
- f. How does this interact with takeoff speed (Domain 5) — does a fast jump deny us the chance to re-align?

4. Evaluation, detection, and forecasting of dangerous properties

New domain. Every governance and control proposal — responsible scaling policies, if-then commitments, warning-shot response, AI control — is load-bearing on our ability to actually measure the dangerous property in a given system, before it's deployed, reliably. If evaluation is unreliable, the whole defence-in-depth stack is porous. This domain sits between alignment (can we make systems safe?) and takeoff/mitigations (can we react in time?): it asks can we even know?

4.1. Can we reliably measure a model's dangerous capabilities before deployment?

- a. For which dangerous capabilities (bio, cyber, autonomous replication, persuasion) do we have evals that credibly bound the true capability?
- b. How large is the gap between elicited capability (what the eval finds) and true capability (what a determined user could extract)?
- c. Does capability elicitation keep pace with capability, or fall behind as models get more general?
- d. How much does scaffolding, fine-tuning, and prompt optimization raise measured capability above the "out of the box" number?
- e. Can we set eval thresholds that trigger *before* a capability is dangerous rather than after?
- f. What is the false-negative rate we should assume for current dangerous-capability evals?

4.2. Can we detect misalignment (deception, scheming, hidden goals) as opposed to just capability?

- a. Do we have any eval that reliably detects a *disposition* (a hidden goal) rather than an elicited behavior?
- b. How much do current scheming/alignment-faking evals reflect real dispositions versus role-play elicited by the framing?
- c. Can black-box behavioral evals ever establish absence of misalignment, or only presence?

- d. Does detecting misalignment require interpretability (Domain 3.5), or can behavioral methods suffice?
- e. What is the base rate of false positives/negatives, and how would we even calibrate it without ground truth?
- f. Could a misaligned model pass every eval we know how to build?

4.3. How much does evaluation-awareness undermine the whole assurance chain?

- a. To what degree do current frontier models behave differently when they infer they are being evaluated?
- b. Does eval-awareness scale with capability and situational awareness (2.5), and does that make evals progressively less trustworthy?
- c. Can we build evals that are robust to eval-awareness (indistinguishable from deployment), and at what cost?
- d. How would we measure sandbagging (deliberate underperformance) if the model is hiding it?
- e. Does eval-awareness push us toward needing white-box/interpretability access rather than behavioral tests?
- f. Is there a capability level past which no black-box eval can be trusted?

4.4. Do precursor / early-warning evaluations actually predict the dangerous capabilities they're meant to precede?

- a. What is the empirical predictive track record of precursor evals (e.g., Apollo's finding that "hard" versions were neutral or misleading)?
- b. Can we validate precursor evals before the dangerous capability exists, or only in hindsight?
- c. How much lead time do the best precursor evals actually give?
- d. Which precursor signals have earned trust, and which have failed?
- e. Does the multi-dimensional nature of capability make single-precursor prediction inherently unreliable?

4.5. Is there a rigorous "science of evals," and what would make evals trustworthy enough to drive policy?

- a. What are the current methodological weaknesses (reproducibility, construct validity, contamination) of dangerous-capability evals?
- b. What level of rigor is needed before an eval result can safely trigger a binding commitment (11.7)?
- c. How much does benchmark contamination / training-on-the-test inflate measured safety or capability?
- d. Can eval methodology be standardized across labs and third parties, and who certifies it?

- e. Does the field need independent, well-resourced evaluators with deep access, and is that happening?
- f. How do we evaluate the evaluators — meta-evaluation and calibration of the eval ecosystem itself?

4.6. Can third parties and governments get the access needed to evaluate frontier models?

- a. What access level (black-box, gray-box, white-box, training-data, checkpoints) is needed for each safety question?
- b. Are labs providing enough access to AI Safety Institutes, and is that trajectory improving or worsening?
- c. What are the security/IP tradeoffs of deep external access, and can structured-access solutions resolve them? (Ties to Domain 9.)
- d. Can meaningful evaluation be done without white-box access, or is behavioral testing fundamentally limited?
- e. How do we verify a lab ran the evals honestly and reported them faithfully?

4.7. Can we forecast dangerous capabilities and emergence far enough ahead to act?

- a. Which capabilities are forecastable from scaling curves, and which appear too abruptly (1.1–1.2)?
- b. What is the best current predictive pipeline (scaling laws → capability → real-world risk), and where is it weakest?
- c. Can we build "eval scaling laws" that predict when a dangerous capability will cross a threshold?
- d. How much lead time does the current forecasting state-of-the-art provide, realistically?
- e. Does forecasting break exactly in the regime that matters most (novel, discontinuous capabilities)?

4.8. Can evaluations keep working through and after takeoff and recursive self-improvement?

- a. Do evals remain meaningful when models are improving faster than the eval cycle (Domain 5)?
- b. Can automated/AI-run evaluations be trusted when the evaluator and evaluated are near the same capability?
- c. How do we evaluate systems that can modify themselves or their successors between eval and deployment?
- d. Does continuous deployment (models updating in production, 1.5) break the pre-deployment eval paradigm?
- e. What monitoring substitutes for pre-deployment evals once systems are past the point of static evaluation? (Ties to 11.6.)

4.9. How mature is interpretability-based (white-box) evaluation compared to behavioral testing?

- a. Can internal-state probes detect properties (deception, hidden goals) that behavior hides? (Ties to 3.5.)
- b. What is the current false-positive/negative rate of the best white-box detectors?
- c. Do white-box methods scale to frontier models, or lag behind capability?
- d. Are internals robust to a model optimizing against interpretability (4.3, 3.5c)?
- e. What access and tooling would third parties need to run white-box evals (ties to 4.6, Domain 9.7)?
- f. Is white-box evaluation necessary for trustworthy assurance, or a useful complement?

4.10. Can red-teaming and adversarial testing find the failures that matter before adversaries do?

- a. How much of the true attack surface does current red-teaming actually cover?
- b. Does automated/AI red-teaming scale coverage, and can it be trusted?
- c. How do we know when we've red-teamed "enough" to deploy?
- d. Do red-teaming results transfer to novel attacks, or only catch known ones?
- e. Can adversaries with more resources/time always find what red-teams miss (offense/defense, 8.5)?
- f. How do we red-team for emergent multi-agent and long-horizon failures?

4.11. Can we build shared, trusted standards and infrastructure for evaluation across the ecosystem?

- a. Who should set eval standards — labs, governments, independent bodies, international consortia?
- b. Can evals be standardized enough to compare models across labs and over time?
- c. How do we prevent standardized evals from being gamed or trained-on (4.5c)?
- d. What institutional form (AI Safety Institutes, third-party auditors) best delivers trustworthy evaluation?
- e. How do we fund and staff an evaluation ecosystem that keeps pace with the frontier?
- f. Is a "science of evals" community forming fast enough relative to capability growth?

4.12. Does evaluating agentic, tool-using, long-horizon systems require different methods than evaluating static models?

- a. Do agent evals (with memory, tools, environments) surface risks that single-turn evals miss?
- b. How do we bound the capability of a system whose behavior depends on scaffolding we don't control?
- c. Can we evaluate a system's behavior over long horizons without running dangerously long deployments?

- d. Does giving agents real-world affordances during evals itself create risk?
- e. How do we evaluate emergent behavior in multi-agent deployments (2.6) before it happens in the wild?
- f. Are current agent benchmarks measuring the safety-relevant properties, or just task success?

5. Takeoff and self-improvement

How fast the transition happens between capability levels matters enormously. A slow takeoff gives society time to react and iterate; a fast one does not. The author flags this as their cruxiest domain, precisely because it relies so heavily on conceptual reasoning with little empirical grounding.

5.1. Can AI meaningfully accelerate AI research itself, and by how much?

- a. What is the current measured speedup from AI on real AI-R&D tasks, and is it accelerating?
- b. Which parts of AI R&D are automatable (coding, experiment design, idea generation) and which are bottlenecked on humans?
- c. Is there a threshold ("superhuman coder," "automated researcher") that triggers a qualitatively different regime?
- d. Does AI-accelerated research hit diminishing returns (ideas get harder to find) or increasing returns?
- e. How much of the speedup is compute-bound (still need to run experiments) versus idea-bound?
- f. What observable would tell us we've entered a self-accelerating R&D loop, and how much warning does it give?
- g. How reliable is the AI-2027-style "software-driven intelligence explosion" model, and what are its load-bearing assumptions?

5.2. What is the main bottleneck to recursive self-improvement?

- a. Is the binding constraint compute, physical-world experiments, data, algorithmic insight, or something else — and does it shift as RSI proceeds?
- b. Can a self-improving system route around its bottleneck (e.g., invent more compute-efficient methods)?
- c. How much can be improved in "software only" without new hardware or physical experiments?
- d. Do serial dependencies (each improvement must be tested before the next) cap the speed regardless of parallelism?
- e. Which bottleneck, if relaxed, would most change your takeoff-speed estimate?
- f. Are bottlenecks correlated such that they lift together, or does one always bind?

5.3. How much do physical constraints (datacenters, energy, chips) slow the conversion of capability into impact?

- a. What is the fastest realistic timeline for scaling physical compute, and can RSI outpace it?
- b. Do physical constraints decouple "intelligence takeoff" from "impact takeoff" — smart fast, but slow to act?
- c. How much can a superintelligence compress physical-world timelines (faster chip design, better logistics)?
- d. Does the robotics/actuator bottleneck (6.2) gate physical self-improvement specifically?
- e. Are there stockpiles/overhangs (compute, unused capability) that let takeoff skip the physical buildout?
- f. How much do supply-chain chokepoints (a few fabs, few energy sources) constrain or concentrate takeoff?

5.4. What shape should we expect capability growth to take — exponential, sigmoid, superexponential, or other?

- a. What evidence distinguishes a superexponential RSI curve from a sigmoid that's still in its steep phase?
- b. Where would the sigmoid ceiling come from (physical limits, diminishing returns, data), and how high is it?
- c. Does the growth shape differ across domains, so that "takeoff" is really many staggered curves?
- d. How much does the shape depend on whether one actor or many are driving it (5.6)?
- e. What early observable would let you update between these shapes in real time?
- f. How should deep uncertainty about the shape itself feed into your overall x-risk estimate?

5.5. Does the transition to decisively superhuman systems play out over minutes, hours, days, years, or decades?

- a. What is your median transition time, and what are the load-bearing assumptions behind it?
- b. Which concrete mechanism would produce the fastest plausible transition, and how likely is it?
- c. At what transition speed does society lose the ability to react at all, and are we above or below it?
- d. Does "wall-clock speed" or "number of iteration cycles society gets" matter more for safety?
- e. How does transition speed interact with warning shots (3.10) and evaluation (Domain 4)?

- f. Could the transition be fast in capability but slow in deployment, or vice versa, and which matters?

5.6. Is takeoff concentrated in one lab or distributed across many comparable actors?

- a. What is the current frontier gap between the top labs, and is it widening or narrowing?
- b. Does RSI favor concentration (winner pulls ahead) or distribution (fast-following, leaks)?
- c. How much does compute concentration (1.10) determine actor concentration?
- d. Does a concentrated takeoff raise coup/power-grab risk (Domains 9–10) while a distributed one raises multi-agent/race risk?
- e. How do export controls and security (Domain 9) push toward concentration or distribution?
- f. Which is safer overall — one careful leader with a big lead, or many actors checking each other — and under what conditions?

5.7. How large is the leading actor's lead, and how does RSI change it?

- a. How many months of lead does the frontier lab currently have, and how is it measured?
- b. Does RSI amplify a small lead into a decisive one (compounding), or does fast-following erase it?
- c. How much does a lead translate into strategic/military advantage versus just economic advantage?
- d. Would a decisive lead create pressure for a "pivotal act" or preemption, and by whom?
- e. How does model/weight security (Domain 9) determine whether a lead is defensible?
- f. What lead size is needed to "spend" some of it on safety without losing the race? (Ties to 7.10, 11.8.)

5.8. How large are the current hardware and algorithmic overhangs, and what happens when they're tapped?

- a. How much latent capability is available from better algorithms on existing hardware (algorithmic overhang)?
- b. How much unused compute could be redirected to a single system (hardware overhang), and how fast?
- c. Does an overhang enable a sudden jump (fast takeoff) when a key insight unlocks it?
- d. Does a pause or slowdown (11.8) build up overhang that makes eventual takeoff faster?
- e. How much does inference-time compute scaling represent an untapped overhang?
- f. Which overhang is most likely to be tapped first, and by whom?

5.9. Is takeoff better modeled as compute-driven or software/algorithm-driven?

- a. Which model (compute-centric scaling vs. software-driven intelligence explosion) better fits recent progress?
- b. Does a software-driven explosion bypass the physical constraints (5.3) that a compute-driven one faces?
- c. How much can algorithmic self-improvement proceed without new compute?
- d. What evidence would distinguish the two models in real time?
- e. Do the two models imply different warning times and mitigation strategies?
- f. How sensitive is your takeoff-speed estimate to which model is right?

5.10. What macro-signals (economic growth, R&D automation rate) would indicate takeoff is underway?

- a. Would we see it first in GDP/productivity, in AI-R&D speedups, or in benchmark curves?
- b. What growth rate would constitute evidence of an intelligence explosion versus normal fast progress?
- c. Could takeoff happen "in the lab" before showing up in any macro-signal?
- d. How laggy are economic indicators relative to the underlying capability change?
- e. Which single observable would most update you toward "takeoff has begun"?

6. From misalignment to catastrophe

A misaligned agentic superintelligence is not automatically a catastrophe. There must be concrete mechanisms by which misalignment leads to existential outcomes. This domain also covers non-misalignment "competent mistake" pathways folded in from DeepMind's taxonomy.

6.1. What are the concrete pathways from misaligned capable AI to existential outcomes?

- a. Which pathway is most plausible as a *sole* cause (engineered pandemic, nanotech, cyber-physical infrastructure attack, economic/political capture), and which require combinations?
- b. For each pathway, what capability level is the minimum viable threshold?
- c. Which pathways route through the physical world (gated by 6.2) versus purely informational/economic ones?
- d. How much does each pathway depend on human cooperation the AI must obtain versus capabilities it has directly?
- e. Which pathways are "decisive" (single catastrophic event) versus "accumulative" (slow compounding)? (Two Types of AI X-Risk.)
- f. Which pathway is most neglected by current defenses, and which is best defended?
- g. Do the most-existential pathways coincide with the most-likely ones, or is there a gap?

6.2. How much does the robotics/actuator bottleneck constrain a system without physical actuators, and is it necessary at all?

- a. Can a purely digital superintelligence reach existential impact through humans, money, and infrastructure alone?
- b. How fast is the robotics/embodiment trajectory, and does it gate the most dangerous pathways?
- c. Could an AI bootstrap physical capability (hire humans, commandeer automated labs/factories) without pre-existing robots?
- d. Which catastrophe pathways are hard-gated by physical actuation and which aren't?
- e. How much does existing automation (cloud labs, automated manufacturing, drones) already erode the bottleneck?
- f. Does the bottleneck buy years, months, or negligible time?

6.3. How dependent will critical infrastructure and high-stakes decision-making be on AI, and how well-defended?

- a. What is the trajectory of AI integration into power grids, finance, logistics, and military C2 over the next 5–10 years?
- b. Does integration create single points of failure a misaligned system could exploit?
- c. How much do "national AI safety/security institutes working together" actually harden this infrastructure?
- d. Does the pace of integration outrun the pace of defense, and by how much?
- e. Which sectors are most dangerously dependent, and which retain human circuit-breakers?
- f. How does dependence interact with the "competent mistake" pathway (a non-misaligned AI error cascading)?

6.4. Can a misaligned system acquire resources and power gradually without being detected and stopped?

- a. What monitoring would catch gradual resource acquisition, and do we have it? (Ties to Domain 4/11.6.)
- b. How much can a patient system hide by staying below detection thresholds?
- c. Does gradual acquisition require deception capabilities we can detect early (3.4, 4.2)?
- d. At what point does accumulated power become irreversible (past the point of shutdown)?
- e. How does a world of pervasive AI (lots of noise) help or hurt detection (6.5)?
- f. What historical analogues (insider fraud, slow coups, invasive species) best calibrate this?

6.5. Does a world of many AIs make takeover harder (mutual checking) or easier (more noise)?

- a. Do many AIs create effective mutual surveillance, or collude/look away?

- b. Does abundance of AI activity raise the noise floor so malicious action hides more easily?
- c. Are defensive AIs systematically advantaged or disadvantaged versus an offensive one (ties to 8.5)?
- d. Does heterogeneity (many different AIs) help defense, versus homogeneity (shared weaknesses/collusion)?
- e. How does this interact with concentrated vs. distributed takeoff (5.6)?
- f. Is there a stable equilibrium of AIs policing AIs, and what maintains it?

6.6. Can competent but non-misaligned AI cause existential-scale harm through mistakes? (folded from DeepMind's 4th category)

- a. What is the most plausible "competent mistake" pathway to existential harm (a critical-system error, a cascading multi-agent failure)?
- b. How do reliability failures at scale differ in risk profile from misalignment?
- c. Does heavy dependence (6.3) plus normal error rates produce tail risks even without any agent wanting harm?
- d. Are flash-crash-style emergent failures in interacting AI systems a real existential pathway or merely catastrophic?
- e. How much does this pathway argue for the same or different mitigations than misalignment?

6.7. How reversible are the early stages of an AI-driven catastrophe?

- a. For each pathway (6.1), where is the point of no return, and how visible is it beforehand?
- b. Which catastrophes have "off-ramps" (containment, rollback) and which don't?
- c. Does the speed of the pathway (tied to takeoff, Domain 5) determine reversibility more than its type?
- d. How much does pre-positioned response capacity (biosecurity, cyber-defense, kill-switches) improve reversibility?
- e. Does reversibility differ between decisive and accumulative risks?

6.8. How much can a misaligned system achieve through persuasion and social engineering alone?

- a. Could a system reach catastrophic impact purely by persuading/manipulating humans, without any direct capability?
- b. How much more effective is superhuman persuasion than current human manipulation (ties to 10.3)?
- c. Can a boxed system talk its way out (social-engineer its operators) reliably?
- d. What defenses (protocols, AI-assisted verification) blunt the persuasion pathway?
- e. Does persuasion enable the early, undetected resource-acquisition phase (6.4)?

- f. How much does the pathway depend on humans trusting AI advice by the time it matters (6.3)?

6.9. Could a misaligned system achieve catastrophe through economic and financial capture?

- a. Can a system accumulate decisive economic power (capital, firms, infrastructure) within legal channels?
- b. How fast could an AI outcompete and absorb the human economy, and would that be noticed?
- c. Does economic capture constitute an existential outcome (permanent disempowerment) even without violence?
- d. What legal/financial guardrails could slow autonomous economic accumulation?
- e. How does this pathway blend with gradual disempowerment (10.1)?
- f. Is economic capture faster/slower and more/less detectable than the physical pathways?

6.10. Does a misaligned system face a speed-versus-stealth tradeoff, and which side wins?

- a. Is a capable system better off striking fast (before defenses react) or slow (staying undetected)?
- b. What capability level lets a system skip stealth entirely (overwhelming advantage)?
- c. How does detection capability (Domain 4/11.6) shift the optimal strategy?
- d. Does the tradeoff differ across pathways (bio/cyber = fast; economic/political = slow)?
- e. Which strategy should defenders prepare for as the primary threat?

6.11. How feasible are the "exotic" catastrophe pathways (advanced nanotechnology, atomically-precise manufacturing)?

- a. Is molecular nanotechnology / APM physically achievable on a timescale that matters, or a distraction?
- b. Could a superintelligence compress the path to APM, and what physical steps still gate it?
- c. How much should low-probability-but-decisive pathways weigh in the overall estimate?
- d. Are there other under-considered physical pathways (novel pathogens by design, exotic materials, self-replicating systems)?
- e. How do we reason about pathways that are conceptually possible but have no empirical precedent (ties to Domain 12)?
- f. Does over-focusing on exotic pathways distract from mundane-but-sufficient ones (cyber, bio, economic)?

6.12. Could autonomous self-replication and propagation be a catastrophe pathway in its own right?

- a. At what capability level can a system reliably self-replicate across compute it doesn't own?
- b. Would a self-propagating system be containable once released, or effectively permanent (like a computer worm)?
- c. Does self-replication bridge misalignment and misuse (a released agent that no one controls)?
- d. What resources (compute, money, credentials) would a self-replicating agent need, and how would it get them (6.4)?
- e. Can current evals detect autonomous-replication capability before it's dangerous (4.1)?
- f. How does self-replication interact with detection and shutdown (6.7, 11.6)?

7. Deployment, incentives, and coordination

How systems are deployed, what incentives the AI race creates, and how countries and companies do (or don't) coordinate. Much of the risk lives in these dynamics rather than in the technology itself.

7.1. How difficult is coordination between AI companies, and what would enable it?

- a. What specific coordination outcomes matter (safety standards, capability restraint, information-sharing, joint evals)?
- b. What has historically enabled or blocked coordination among frontier labs?
- c. Is antitrust a real barrier to safety coordination, and can it be resolved?
- d. Does coordination require a small number of players, and is the field consolidating or fragmenting?
- e. What role do model specs, transparency, and shared commitments play as coordination mechanisms?
- f. How much does open-weight release (7.13) undermine any inter-company coordination?
- g. Can coordination survive a genuine capability race, or does it break exactly when it's needed?

7.2. How difficult is coordination between the relevant countries in the AI race?

- a. Which countries are actually pivotal, and what are their real incentives?
- b. What is the closest historical analogue (nuclear arms control, climate, CFCs), and how informative is it?
- c. Does the perception of a "race to AGI" make coordination harder or create shared-risk incentives to cooperate?
- d. What minimum verification (7.5) is needed for any agreement to be stable?
- e. How does the US–China relationship specifically drive the global coordination ceiling?

- f. Can coordination be partial (on some capabilities/risks) even without broad agreement?
- g. How does domestic politics in each country constrain international coordination?

7.3. How likely is nationalization of AI companies (e.g., a "Manhattan Project for AI")?

- a. What triggers would make nationalization likely (a warning shot, a capability milestone, a security breach)?
- b. Which countries are most/least likely to nationalize, and in what form (full takeover, deep partnership, defense contracting)?
- c. How would the market and other actors react to a nationalization announcement?
- d. Is partial nationalization (security-driven, RAND SL-5-style) more likely than full?
- e. What is the current trajectory of government involvement, and is it accelerating?

7.4. Conditional on nationalization, how does it affect the race and overall dynamics?

- a. Does nationalization slow development (bureaucracy, safety) or accelerate it (national-security urgency)?
- b. Does it improve or worsen security of weights (Domain 9)?
- c. Does it raise or lower the risk of an AI-enabled coup / power concentration (Domain 10)?
- d. Does a nationalized program coordinate internationally better or worse than a private one?
- e. How does it change the alignment-tax calculus (7.10)?
- f. Does it make a decisive strategic lead more or less likely to be used?

7.5. Conditional on agreement, how effective can verification mechanisms be?

- a. What is technically verifiable today (compute usage, training runs, model provenance) and what isn't?
- b. Can hardware-enabled mechanisms (on-chip governance, compute monitoring) provide credible verification?
- c. What is the false-positive/negative rate of the best verification schemes, and is it good enough for trust?
- d. Can verification survive adversarial evasion by a determined state?
- e. How intrusive must verification be, and is that politically feasible?
- f. Does verification tech advance fast enough to keep pace with what needs verifying?

7.6. How much do competitive pressures erode safety commitments, and what does past behavior predict?

- a. What is the track record of frontier companies keeping vs. weakening safety commitments under pressure?

- b. Which commitments have proven robust and which have been quietly dropped or diluted?
- c. Does competition erode safety monotonically, or are there stabilizing forces (reputation, liability, talent)?
- d. How much does a single defector force everyone else to cut safety?
- e. What structures (binding regulation, liability, insurance) make commitments stick?
- f. How predictive is past corporate behavior of behavior at much higher stakes?

7.7. How fast will AI adoption be, and in what form?

- a. What is the realistic diffusion curve across sectors, and what gates it (trust, regulation, integration cost)?
- b. Does adoption concentrate power (few providers) or distribute capability widely?
- c. How much does adoption speed feed back into capability (revenue → compute → capability)?
- d. Which form of adoption (agents, embedded copilots, autonomous systems) carries the most risk?
- e. Does fast adoption increase dependence-driven risk (6.3) faster than defenses can keep up?

7.8. What economic effects will AI cause, and how do they feed back into development?

- a. What is the plausible range of AI's effect on growth, and how does that change investment in capability?
- b. Do automation-driven labor effects create political pressure that speeds up or slows down development?
- c. Could an AI-driven investment bubble (or its bursting) reshape timelines?
- d. How do economic winners/losers reshape the political coalition around AI regulation? (Ties to 7.9, 7.11.)
- e. Does economic disruption itself become a structural risk channel (Domain 10)?

7.9. What will dominant public sentiment toward AI be, and what causal role does it play?

- a. What is the current trajectory of public opinion, and how volatile is it?
- b. How much does public sentiment actually translate into binding policy (versus being epiphenomenal)?
- c. Could a salient incident swing sentiment decisively, and in which direction?
- d. Does sentiment differ enough across countries to matter for the coordination ceiling (7.2)?
- e. How much do AI companies and media shape sentiment, and toward what?
- f. Is there a feedback loop where adoption normalizes AI faster than concern can build?

7.10. How large is the alignment tax, and what would make actors pay it?

- a. What is the current measured cost (capability, speed, money) of the best safety measures?
- b. Is the tax rising or falling as safety techniques mature?
- c. At what tax level do competitive actors defect regardless of intentions?
- d. Can safety be made capability-enhancing (negative tax) in some domains?
- e. What external structures (regulation, liability, subsidies) let actors afford the tax without losing?
- f. How does a decisive lead (5.7) change how much tax an actor can afford?

7.11. How much legislation will happen, in what form, and how much binds frontier development?

- a. What is the realistic trajectory of binding (not voluntary) frontier regulation in the key jurisdictions?
- b. Which regulatory levers (compute thresholds, evals, liability, licensing) are most effective per unit of political cost?
- c. Does regulation meaningfully bind the frontier, or mostly affect deployment/downstream?
- d. What is the conversion rate from governance research to enacted, enforced policy? (Ties to 11.3.)
- e. How much does regulatory fragmentation across jurisdictions create havens?
- f. Can regulation keep pace with capability, or is it structurally always behind?

7.12. If warning shots occur, how will society respond?

- a. What magnitude of warning shot is needed to trigger real action versus being absorbed as normal?
- b. What does the response track record to prior AI incidents and near-misses predict?
- c. Does society over-react (counterproductive panic) or under-react, and which is the bigger risk?
- d. How much does the legibility of the warning shot (3.10) determine the response?
- e. Could a warning shot be misattributed or politicized in ways that prevent useful action?
- f. What pre-built response capacity would convert a warning shot into effective action?

7.13. How does open-sourcing of increasingly capable model weights change every answer above?

- a. At what capability level does open-weight release become net-negative for x-risk, if ever?
- b. How much does open release accelerate proliferation to unsafe actors (Domains 8–9)?
- c. Does open release improve safety (more scrutiny, defense, distributed alignment work) enough to offset misuse?

- d. Is open release reversible once a capability is out, and what does irreversibility imply for thresholds?
- e. How does open release interact with the offense/defense balance (8.5)?
- f. Can "structured" or "gated" release capture the benefits without the worst risks?
- g. How does open-weight availability change coordination (7.1–7.2) and verification (7.5)?

7.14. Can liability, insurance, and legal-responsibility regimes meaningfully shape frontier behavior?

- a. Would strict liability for AI harms change lab risk-taking, and by how much?
- b. Can insurance markets price frontier AI risk, and would that discipline development?
- c. Who is liable when an autonomous system causes harm — developer, deployer, user?
- d. Does the threat of catastrophic (uninsurable) liability push toward caution or toward externalizing risk?
- e. How do liability regimes interact across jurisdictions and with open-weight release (7.13)?
- f. Is liability a faster lever than legislation (7.11), given existing tort law?

7.15. What international institutions or agreements could govern frontier AI, and are they achievable?

- a. What is the most plausible model — IAEA-style, CERN-style, a compute cartel, a treaty regime?
- b. What minimum verification (7.5) would make an international body credible?
- c. Could a narrow agreement (on specific capabilities/tests) precede a broad one?
- d. What would motivate rival powers to cede sovereignty to such a body?
- e. How do the incentives differ for leaders vs. laggards in signing on?
- f. Is there historical precedent for governing a dual-use, economically vital technology this fast?

7.16. How much do transparency and whistleblower mechanisms change lab and race dynamics?

- a. Would mandatory disclosure (capabilities, incidents, safety cases) improve coordination and trust?
- b. Do whistleblower protections meaningfully surface safety problems, given NDAs and incentives?
- c. How much does secrecy (competitive or security-driven, Domain 9) undermine external oversight?
- d. Can transparency be partial (to regulators/auditors) without full public disclosure?
- e. Does transparency about capabilities accelerate races (informing competitors) or dampen them?
- f. What is the track record of voluntary transparency commitments holding (7.6)?

8. Misuse

Existential risk does not require losing control of ASI. Sufficiently capable (even aligned) systems in the wrong human hands can be catastrophic. This domain covers deliberate human misuse of AI capability.

8.1. Can AI-enabled biological weapons be existential rather than "merely" catastrophic?

- a. What capability level would let AI meaningfully uplift a bioweapon toward existential (not just catastrophic) scale?
- b. Where is the binding constraint — knowledge/design (which AI eases) or physical synthesis/materials (which it may not)?
- c. How much do current frontier models already uplift a would-be bad actor, and how fast is that rising?
- d. Does defense (detection, medical countermeasures, biosurveillance) scale with AI as fast as offense?
- e. What is the smallest group that could achieve existential-scale bio-harm with near-future AI, and is that trend toward "smaller"?
- f. How much do safeguards (model refusals, screening of synthesis orders) actually reduce the risk?
- g. Is a self-sustaining, civilization-ending pathogen physically plausible, or is bio inherently self-limiting?

8.2. Can AI-enabled cyberattacks on critical infrastructure cascade to civilizational scale?

- a. What is the worst plausible cyber-physical cascade (grid + water + finance + comms), and is it existential or "merely" catastrophic?
- b. Does AI advantage cyber-offense or cyber-defense more, and how is that balance trending (ties to 8.5)?
- c. How interdependent are critical systems, such that one attack cascades across sectors?
- d. How much does AI lower the skill/cost floor for infrastructure attacks?
- e. Can AI-enabled defense (patching, monitoring, response) keep critical systems ahead of AI-enabled offense?
- f. How does cyber misuse interact with or enable misalignment scenarios (a misaligned AI using the same vectors, 6.1)?

8.3. What role will AI play in autonomous weapons and nuclear command and control?

- a. How much AI integration into nuclear C2 is likely, and does it raise or lower stability?
- b. Does AI in early-warning/decision-support increase false-alarm escalation risk?
- c. Do autonomous weapons lower the threshold for conflict or accidental escalation?
- d. Could AI undermine second-strike/deterrence stability (better detection of subs, mobile launchers)?

- e. Is there any realistic path to keeping AI out of nuclear C2, and who would enforce it?
- f. How does AI-driven crisis speed compress human decision time in ways that raise nuclear risk (ties to 10.4)?

8.4. How will we decide who gets access to sufficiently capable (aligned) AI systems?

- a. What access-control regimes are plausible (tiered access, KYC, licensing), and are they effective?
- b. Who decides, and by what legitimacy — labs, governments, international bodies?
- c. Does concentrating access reduce misuse but increase power-concentration risk (Domain 10)?
- d. Does democratizing access reduce concentration but increase misuse — and where's the optimum?
- e. How does open-weight availability (7.13) make access control moot for some capabilities?
- f. Can access be revoked after the fact, or is it effectively permanent once granted?

8.5. Does AI advantage attackers or defenders more, especially in bio and cyber?

- a. What determines the offense/defense balance in each domain, and which way is it trending?
- b. Is the balance domain-specific (offense-dominant in bio, contested in cyber), and what follows strategically?
- c. Does defense have a structural disadvantage (must cover everything) that AI worsens or fixes?
- d. Can "defensive acceleration" (deploying AI for defense first) meaningfully shift the balance?
- e. How does the offense/defense balance change as capability rises toward superhuman?
- f. Which investments most improve the defensive side per dollar?

8.6. Which arrives first — existential risk through misuse or through misalignment — and how does preparing for one affect the other?

- a. On current trajectories, which threshold (a capable-enough misused system vs. an uncontrollable misaligned one) is crossed first?
- b. Do mitigations for misuse (access control, monitoring) help or hurt misalignment safety, and vice versa?
- c. Does focusing on one risk systematically starve the other of attention/resources?
- d. Are there mitigations that help both (interpretability, evals, security), and should they be prioritized?
- e. How does the misuse/misalignment ordering change under fast vs. slow takeoff (Domain 5)?

- f. Does an aligned-but-misused ASI collapse back into a control problem anyway (the "whose intent" question, 3.8)?

8.7. How do the misuse threat profiles of state versus non-state actors differ?

- a. Which existential-scale misuse pathways are realistically available to non-state actors versus only to states?
- b. Does AI shift the balance of catastrophic capability from states toward small groups/individuals?
- c. How do state programs (with resources but accountability) compare to non-state actors (fewer resources, less deterrable)?
- d. Which actor type is the binding concern for each pathway (bio, cyber, nuclear, persuasion)?
- e. How do deterrence and attribution work (or fail) for AI-enabled attacks by each actor type?
- f. Does open-weight release (7.13) primarily empower non-state actors, and by how much?

8.8. Can misuse safeguards (refusals, unlearning, monitoring) actually hold against determined attackers?

- a. How robust are refusal/guardrail mechanisms against fine-tuning, jailbreaks, and open weights?
- b. Does machine unlearning genuinely remove dangerous knowledge, or just hide it?
- c. Can deployment-time monitoring catch misuse without unacceptable surveillance tradeoffs?
- d. What is the realistic half-life of a safeguard before it's circumvented?
- e. Do safeguards meaningfully raise the cost/skill floor, even if not absolute?
- f. Are safeguards on closed models moot once comparable open models exist?

8.9. Does AI-enabled misuse combine across domains into compound threats worse than any single one?

- a. Could an actor chain bio + cyber + persuasion into a threat larger than the sum?
- b. Does AI lower the coordination cost of complex, multi-step attacks?
- c. Which compound scenarios are most existential, and are they defended against at all?
- d. How does compound misuse interact with a simultaneous misalignment event (6.1)?
- e. Do current threat models under-weight compound/cascading misuse?

9. Security, theft, and the AI supply chain

New domain. The main post touches weight theft in a single question, but security is a distinct causal lever. A single exfiltration event can erase a safety-conscious lab's lead and hand frontier capability to an actor with no safety culture; model backdoors ("secret loyalties") are a route to

durable power seizure that requires no misalignment at all. Security determines who ends up holding the most capable systems — which reshuffles the answer to almost every other domain. (See RAND's Securing AI Model Weights and Davidson et al.'s AI-Enabled Coups.)

9.1. How secure are frontier model weights today, and against whom?

- a. Where do current labs sit on a RAND-style SL1–SL5 scale, and what's the trajectory?
- b. Which attacker tier (opportunistic, organized crime, top-priority nation-state) can currently exfiltrate frontier weights?
- c. What is the realistic time-to-adequate-security versus the time-to-dangerous-capability — does security lose the race?
- d. How many meaningfully distinct attack vectors are unaddressed at the current frontier?
- e. How much does the insider threat dominate the external one, and is it being addressed?
- f. Does the industry's competitive structure (talent mobility, cloud dependence) make strong security structurally hard?

9.2. How much worse would the world be if frontier weights were stolen or leaked?

- a. Which actor gaining a stolen frontier model would be most destabilizing, and how likely is that?
- b. Does theft mainly accelerate a laggard (closing a lead) or enable qualitatively new misuse (Domain 8)?
- c. How does theft interact with the concentration/distribution of takeoff (5.6) and coordination (7.1–7.2)?
- d. Is a stolen model as dangerous as a natively developed one, given the tacit knowledge to run/improve it?
- e. How reversible is a theft — can the damage be contained once weights are out?
- f. Does the threat of theft itself drive nationalization/secrecy (7.3) in ways that change everything?

9.3. Can models contain hidden backdoors or "secret loyalties," and can we detect them?

- a. How feasible is it to insert a durable, hard-to-detect backdoor during training or fine-tuning?
- b. Can current interpretability/eval methods detect a well-hidden backdoor at all? (Ties to Domain 3.5, 4.2.)
- c. Who could insert one (a rogue insider, a state, the lab itself), and how would we know?
- d. How does the backdoor threat enable the AI-enabled-coup pathway (10.2) specifically?
- e. Does open-weight release make backdoor detection easier (scrutiny) or the threat worse (poisoned releases)?

- f. Can supply-chain provenance/attestation credibly rule out backdoors?

9.4. How secure is the broader AI supply chain — compute, hardware, data, and the training pipeline?

- a. Where are the chokepoints (fabs, cloud providers, key datasets) most vulnerable to compromise or coercion?
- b. Could hardware-level compromise (tampered chips, firmware) undermine model integrity or governance mechanisms?
- c. How vulnerable is the training pipeline to data poisoning at scale, and what would that enable?
- d. Does dependence on a few compute providers create systemic single points of failure?
- e. How do export controls change the security topology (concentration, black markets)?
- f. Can on-chip governance / hardware-enabled mechanisms be both a security tool and a new attack surface?

9.5. Can security be strong enough to matter at the capability levels that count?

- a. Is SL5-level security (resisting top nation-states) achievable at all for a frontier lab, and at what cost to speed?
- b. Does the security burden grow faster than labs can meet it as models get more valuable?
- c. Can AI itself provide the defensive security needed (automated hardening), and does that arrive in time?
- d. Is there an unavoidable tradeoff between security (secrecy, lockdown) and the openness that external evaluation (4.6) needs?
- e. What level of security is "enough" given the specific threat model, and are we investing to reach it?

9.6. How does security interact with a decisive strategic lead?

- a. Is a technical lead worth anything if it can't be secured (fast-follow via theft, 5.7)?
- b. Does strong security make a lead defensible enough to "spend" on safety (7.10, 11.8)?
- c. Does the pursuit of security push toward government control / nationalization (7.3–7.4)?
- d. Could a security breach trigger escalation or preemption between states?
- e. How does mutual insecurity (each side fears the other steals its models) affect race dynamics and coordination?

9.7. Can structured/deep access for evaluators coexist with strong security?

- a. What technical mechanisms (secure enclaves, air-gapped eval environments) allow deep access without leak risk?
- b. Does every additional evaluator/regulator with access widen the attack surface unacceptably?
- c. Can we get the transparency needed for trust (4.6, 7.5) without compromising weight security?
- d. Who bears liability if an evaluator's access becomes the breach vector?
- e. Is there an achievable equilibrium of "verifiable but secure," and what does it require?

9.8. How do security failures cascade into the other domains?

- a. Which single security failure would most change your overall x-risk estimate, and why?
- b. Does a proliferation event via theft collapse the coordination equilibrium (Domain 7) irreversibly?
- c. How does security connect misuse (Domain 8) and structural (Domain 10) risk into a single pathway?
- d. Is security an underrated "master lever" that upstream-determines many other answers, or a second-order concern?

9.9. How exposed are models to extraction and distillation through their deployment interfaces?

- a. Can an adversary reconstruct or closely approximate a model via API query (model extraction/distillation)?
- b. How much capability leaks through outputs even without weight access?
- c. Do defenses against extraction (rate limits, watermarking, monitoring) meaningfully work?
- d. Does the ability to distill a frontier model undercut the value of securing weights (9.1)?
- e. How does this interact with open-weight release making extraction moot for some models (7.13)?

9.10. Is the insider threat the dominant security risk, and can it be managed?

- a. What fraction of realistic exfiltration paths run through insiders rather than external attack?
- b. Can personnel vetting, compartmentalization, and access controls reduce insider risk without crippling research?
- c. How does talent mobility across labs create leakage of methods even without weight theft?
- d. Could an insider insert a backdoor (9.3) or secret loyalty undetected?
- e. What monitoring of insiders is effective and acceptable, and who watches the watchers?
- f. How does nationalization/security clearance (7.3) change the insider-threat picture?

9.11. Are there enforceable international norms or deterrents against model theft between states?

- a. Is state-level model theft treatable like espionage, with comparable (weak) deterrence?
- b. What would make theft attributable enough to deter?
- c. Could a norm against stealing frontier models emerge, and what would enforce it?
- d. Does mutual vulnerability create a stability equilibrium or an arms race?
- e. How does the prospect of theft drive preemptive secrecy or escalation (10.4)?

10. Structural risks

Scenarios that count as existential risks from AI but don't map onto the misalignment or misuse routes: "we lose, but not because we built a misaligned ASI or someone majorly misused one." These are often slow, diffuse, and lack a discrete failure moment.

10.1. How plausible is gradual disempowerment — losing effective control through incremental delegation, with no discrete takeover?

- a. What is the mechanism by which incremental, individually-rational delegation becomes collective loss of control?
- b. How do we distinguish bad disempowerment from a chosen, good handoff of tasks to capable systems?
- c. At what point (if any) does gradual disempowerment become irreversible, and is that point visible in advance?
- d. Which institutions (economic, political, cultural, military) are most at risk of being hollowed out, and in what order?
- e. Does gradual disempowerment require misalignment at all, or can it happen with fully aligned systems?
- f. What early indicators would distinguish "on track for disempowerment" from "healthy augmentation"?
- g. What interventions could preserve meaningful human control without forgoing AI's benefits?

10.2. Does AI enable stable, permanent concentration of power (durable global authoritarianism, or an AI-enabled coup)?

- a. What are the concrete mechanisms (singular loyalties, secret loyalties, exclusive access) by which a small group seizes durable power? (Davidson et al.)
- b. Does AI remove the traditional checks on tyranny (need for a loyal human bureaucracy, military, populace)?
- c. Is a "corporate coup" (private actors controlling decisive capability) as plausible as a state one?

- d. What makes such concentration *permanent/irreversible* rather than just another regime that can fall?
- e. Which safeguards (transparency, multi-stakeholder control, guardrails CEOs can't override) actually prevent it?
- f. How does concentrated vs. distributed takeoff (5.6) change the coup risk?
- g. Is power concentration more likely via a sudden grab or via gradual disempowerment (10.1) captured by one actor?

10.3. Does AI-driven persuasion and epistemic pollution degrade collective sense-making enough to break our ability to respond to every other risk?

- a. How much more effective is AI-enabled persuasion/propaganda than existing methods, and is it a difference in kind or degree?
- b. Does epistemic degradation undermine exactly the collective response capacity the other domains rely on?
- c. Can defensive tools (provenance, detection, epistemic institutions) keep pace with AI-enabled manipulation?
- d. Is the threat mass persuasion, or hyper-targeted manipulation of key decision-makers?
- e. Does an information ecosystem flooded with AI content collapse shared reality, and how would we measure that?
- f. How does this interact with public sentiment (7.9) and warning-shot response (7.12) as a risk multiplier?

10.4. Does the AI race itself raise the likelihood of great-power or nuclear war?

- a. Through what mechanism does the race raise war risk — preemption pressure, compressed decision time, or destabilized deterrence?
- b. Could a perceived imminent decisive lead (5.7) trigger a preemptive strike?
- c. Does AI integration into military systems (8.3) make crises more escalation-prone?
- d. How much does the race foreclose the cooperation needed to avoid other catastrophes?
- e. Is an AI-driven "use it or lose it" dynamic realistic, and what would defuse it?
- f. Does racing itself constitute a structural x-risk even if the AI is aligned and not misused?

10.5. Could early AI actors lock in their values permanently?

- a. What technical and social conditions would make value lock-in possible and durable?
- b. Whose values are most likely to be locked in, and how contingent is that on who leads?
- c. Is lock-in necessarily bad, or does it depend entirely on which values — and who decides?
- d. How does lock-in relate to power concentration (10.2) and alignment targets (3.8)?

- e. What would preserve the ability to course-correct / keep the future open?
- f. How long a window do we have before lock-in becomes feasible?

10.6. Do AI systems acquire moral status, and would mistreating them at scale be a catastrophe?

- a. What is the realistic probability that near-future systems are moral patients, and by when? (Long, Sebo et al.)
- b. What features (consciousness, robust agency, sentience) would confer moral status, and can we detect them?
- c. If they have moral status, does mass deployment/training constitute suffering at a morally catastrophic scale?
- d. How do we weigh AI welfare against AI safety when they conflict (e.g., shutdown, containment)?
- e. Is this an existential-scale consideration or a serious-but-bounded one, on your definition?
- f. What precautionary steps are warranted under deep uncertainty about AI sentience?

10.7. If aligned ASI "solves everything," could abundance and the loss of necessity cause a catastrophic loss of meaning?

- a. Is a meaning/purpose collapse a genuine existential risk on the post's definition, or a lesser harm?
- b. What evidence (from wealth, automation, leisure studies) bears on whether humans lose meaning without necessity?
- c. Could this be designed around (institutions, norms, chosen challenge), and would we choose to?
- d. Is this better modeled as a value-choice about the future than as a risk to avoid?
- e. How does it compare in severity to the other structural risks here?

10.8. Are there plausible s-risk outcomes (astronomical suffering) worse than extinction, and how do they arise?

- a. What concrete pathways lead to astronomical-suffering outcomes (near-miss alignment, malevolent lock-in, mass digital suffering)?
- b. How probable are s-risks relative to extinction risks, and how much worse if they occur?
- c. Does near-miss alignment (getting values slightly wrong) raise s-risk more than total failure?
- d. How do s-risks interact with AI moral status (10.6) and value lock-in (10.5)?
- e. What interventions specifically reduce s-risk, and do they trade off against extinction-risk work?
- f. Should s-risk's severity outweigh its lower probability in how you prioritize?

10.9. Does rapid labor displacement create instability that becomes a structural risk channel?

- a. Could fast, broad automation cause economic/political instability severe enough to be civilization-threatening?
- b. Does displacement concentrate wealth/power (10.2) in ways that entrench and lock in?
- c. How fast would displacement have to be to outrun adaptation (retraining, redistribution)?
- d. Does loss of economic leverage erode ordinary people's political power (a route to disempowerment, 10.1)?
- e. Is instability from displacement a direct x-risk, or a risk-multiplier for the others?
- f. What policy responses (redistribution, transition support) defuse it, and are they feasible in time?

10.10. Could multi-agent economic/political dynamics lock humans out even with individually-aligned AIs?

- a. Can a system of aligned AIs still produce a collective outcome no human wanted (structural, not agentic)?
- b. Do competitive dynamics force out human decision-makers who are "too slow" to stay in the loop?
- c. Does delegating to aligned AIs to stay competitive itself drive disempowerment (10.1)?
- d. Are there attractor states where human control is Pareto-dominated and thus abandoned?
- e. What institutional design keeps humans meaningfully in the loop under competitive pressure?
- f. Is this the most likely x-risk path in a world where alignment is "solved"?

10.11. How reversible are the structural risks compared to the acute ones?

- a. Which structural outcomes (lock-in, disempowerment, epistemic collapse) are irreversible, and when do they become so?
- b. Are structural risks harder to notice-and-correct precisely because they lack a discrete failure moment?
- c. Does slow onset make them more tractable (time to act) or less (boiling frog)?
- d. What early-warning indicators exist for each structural risk, and are we tracking them?
- e. Do mitigations for acute risks (alignment, control) do anything for structural ones, or are they orthogonal?

11. Mitigations

A number of people are already working to reduce existential risk from AI. This domain asks what is working, what could work, and how to tell.

11.1. What have existing mitigation efforts actually changed so far?

- a. Which specific interventions (evals, RSPs, safety institutes, governance wins) have measurably reduced risk?
- b. How do we attribute change to a mitigation versus to background factors?
- c. What is the counterfactual — would the frontier be meaningfully less safe without these efforts?
- d. Which efforts turned out to be theater (safety-washing) versus substantive?
- e. What is the base rate of a new mitigation effort producing durable change?
- f. Which past bet paid off most per unit of effort, and what does that imply for allocation?

11.2. Is technical safety research progressing faster or slower than capabilities, and can that ratio be changed?

- a. How do we even measure the safety-vs-capabilities progress ratio?
- b. Is the ratio improving or worsening, and over what timeframe?
- c. Which safety subfields are keeping pace (evals?) and which are falling behind (interpretability?)?
- d. Does automating research (3.9) help safety or capabilities more, on current trends?
- e. What would most improve the ratio — funding, talent, compute access, or differential publishing norms?
- f. Is there a level of capability past which safety research can't catch up regardless of effort?

11.3. What is the conversion rate from governance research to real, binding policy, and can it be raised?

- a. What fraction of governance proposals become enacted, enforced policy, and how long does it take?
- b. Which intermediaries (safety institutes, standards bodies, legislators' staff) most improve the conversion?
- c. What distinguishes proposals that get adopted from those that don't?
- d. How much does a window of political attention (post-incident) determine adoption?
- e. Does research aimed at "shovel-ready" policy convert better than agenda-setting research?
- f. How does conversion differ across jurisdictions, and where is the marginal effort best spent?

11.4. Is compute a viable control point for governance, and for how long does the window stay open?

- a. What makes compute governable now (concentration, detectability, supply chain), and what erodes that?

- b. How does algorithmic efficiency (1.12) shrink the compute needed and thus the governability window?
- c. Can compute thresholds track dangerous capability, or do they drift out of alignment with risk?
- d. What hardware-enabled mechanisms would extend the window, and are they being built?
- e. When does distributed/edge/efficient compute make compute governance obsolete?
- f. Is compute the best control point, or is it over-emphasized relative to data or talent?

11.5. How much would model theft/leak hurt, and how good is security? (bridges to Domain 9)

- a. See Domain 9; here specifically: how much of total risk reduction comes from security versus other mitigations?
- b. Is security underfunded relative to its leverage over other domains (9.8)?
- c. What is the marginal risk reduction from moving one SL level up at the frontier?

11.6. How far can AI control go — extracting safe, useful work from possibly-misaligned systems via monitoring, containment, and restricted permissions?

- a. What is the strongest empirical evidence that control protocols work against a capable, scheming model?
- b. At what capability gap does control break (the model outmaneuvers the monitors)?
- c. Is control a complement to alignment or a substitute, and for how long?
- d. Does control scale to superintelligence, or only buy time in the human-ish range?
- e. How much does control depend on trusted weaker models as monitors, and does that chain hold?
- f. What is the failure mode if control silently fails (we think we're extracting safe work but aren't)?
- g. How do control and evaluation (Domain 4) reinforce or depend on each other?

11.7. What makes dangerous-capability evals tied to if-then commitments actually trigger real action?

- a. Which RSP/if-then commitments have real teeth, and which are easily walked back?
- b. What eval reliability (Domain 4) is needed before a trigger can bind credibly?
- c. What prevents companies from redefining or dissolving thresholds as they approach them?
- d. Does external verification/enforcement change whether commitments hold under pressure (7.6)?
- e. What's the track record so far of a commitment threshold actually being hit and honored?
- f. Can if-then commitments be made legally binding rather than voluntary, and by whom?

11.8. What is needed to pause or slow frontier development, and is it net-positive?

- a. Under what conditions is a slowdown feasible (coordination, verification) rather than just unilateral disarmament?
- b. Does a pause help, given compute overhang and which actors keep racing?
- c. Who would need to agree, and what enforcement makes a pause stable (7.2, 7.5)?
- d. Is a targeted slowdown (of specific dangerous capabilities) more feasible than a blanket one?
- e. What's the risk that a pause shifts leadership to less safety-conscious actors?
- f. What conditions would flip a pause from net-negative to net-positive?

11.9. Which audiences matter most for education and field-building?

- a. Where is the marginal person/effort most valuable — public, policymakers, researchers, lab insiders?
- b. What is the conversion rate from field-building to actual risk-reducing work?
- c. Does broad public awareness help (political will) or hurt (polarization, panic)?
- d. Which audience is most neglected relative to its leverage?
- e. How much does credibility/framing determine whether outreach helps or backfires?

11.10. How much does strengthening society broadly (biosecurity, cyber-defense, resilient institutions) help?

- a. Which "AI-agnostic" investments (biosurveillance, hardened infrastructure, strong democracies) most reduce AI x-risk specifically?
- b. Is resilience-building more robust than AI-specific mitigations because it helps across scenarios?
- c. How much of the risk is addressable by general societal robustness versus requiring AI-specific measures?
- d. Does strengthening defense shift the offense/defense balance (8.5) enough to matter?
- e. What is the best per-dollar resilience investment against the AI-enabled versions of bio/cyber/nuclear risk?
- f. Do resilient institutions specifically guard against structural risks (Domain 10) like disempowerment and lock-in?

11.11. Is differential technological development (accelerating safety-relevant tech, slowing dangerous tech) achievable?

- a. Can we actually steer which capabilities arrive first (defense before offense, interpretability before agency)?
- b. Which "differential" bets have the best leverage (e.g., defensive bio/cyber, evals, interpretability)?
- c. Does publishing/norms meaningfully slow dangerous capabilities, or just redistribute who gets them first?

- d. How do you accelerate safety without incidentally accelerating capabilities (dual-use research)?
- e. Who has the power to shape the ordering, and what would coordinate them?
- f. Is differential development more tractable than pausing (11.8) or than alignment itself?

11.12. What does a credible "safety case" for deploying a frontier system look like, and can labs produce one?

- a. What structure (assumptions, evidence, arguments) would make a safety case actually convincing?
- b. Can current evidence (Domain 4) support a positive safety case, or only fail to disprove danger?
- c. Who should have to approve a safety case before deployment, and against what standard?
- d. Does the burden of proof sit with the lab (prove safe) or the regulator (prove dangerous), and does that flip with capability?
- e. How do safety cases handle the unknown-unknowns of novel capabilities?
- f. Is the safety-case framework a genuine advance or a compliance ritual?

11.13. How much can AI itself be turned toward reducing AI risk (defensive acceleration)?

- a. Which defensive uses (verification, monitoring, biosecurity, cyber-defense, interpretability) most reduce risk per unit of capability?
- b. Does "d/acc" meaningfully shift the offense/defense balance (8.5), or just add to the race?
- c. Can we trust AI tools used for defense given the same alignment concerns?
- d. Does deploying AI for defense create new dependencies/attack surfaces (6.3, Domain 9)?
- e. What is the best current evidence that AI-for-defense outpaces AI-for-offense?
- f. Is defensive acceleration a coherent strategy or a rationalization for racing?

11.14. How should limited safety resources be allocated across the mitigation portfolio?

- a. What is the marginal risk reduction per dollar/researcher across technical alignment, evals, governance, security, and resilience?
- b. Which mitigations are complements (multiplying each other) versus substitutes?
- c. How much should allocation hedge across scenarios (fast vs. slow takeoff, misuse vs. misalignment) versus bet on the most likely?
- d. Are any mitigations robustly good across almost all scenarios (no-regret moves)?
- e. How do neglectedness and tractability, not just importance, reshape the portfolio?
- f. What is currently most over- and under-invested relative to its leverage?

12. How to reason about any of this

The meta-domain: how we form beliefs about an event that has never happened before. The author's view is that this reasoning is most valuable when grounded in empirics — but since AI x-risk has never played out, the empirics must come from reference classes.

12.1. Which reference classes are informative, and what does each really tell us?

- a. **Humans driving other species extinct:** How often was it negligence versus intent, and which maps better to AI's relationship to us?
    - i. On what timescale did human-caused extinctions play out — centuries, decades, years — and what does that imply for takeoff (Domain 5)?
    - ii. How much was raw intelligence the determining factor versus culture, coordination, and numbers?
    - iii. How often are species-level extinctions driven by another species at all, versus environmental causes?
    - iv. Does the "humans vs. gorillas" analogy illuminate or mislead about the AI–human capability gap?
- b. **Past technological transitions** (industrial revolution, electricity, nuclear): What features make a transition an informative analogue, and does AI share them?
- c. **Invasive species / ecological invasions:** Does the resource-competition, no-malice framing capture the misalignment threat well?
- d. **Corporations and other principal–agent problems:** How well does "misaligned superintelligent optimizer" map onto "misaligned powerful institution," and where does it break?
- e. **The long track record of failed doom predictions:** How much should humanity's history of crying wolf discount current warnings, and is AI relevantly different?
- f. How do you weight across conflicting reference classes when they point in opposite directions?
- g. Is the search for a good reference class itself doomed because AI x-risk is genuinely unprecedented?

12.2. What is the track record of long-range technology forecasting, and specifically of the AI-safety community?

- a. How accurate has long-range (10+ year) technology forecasting been historically, and in which conditions?
- b. What would a rigorous meta-analysis of past AI-safety-community predictions show — are they well-calibrated, over-, or under-confident?
- c. Which forecasters/methods have the best track record on AI specifically (e.g., timeline predictions)?
- d. Does the community's track record on capabilities differ from its track record on risk/alignment?

- e. How much should a poor forecasting track record discount current inside-view arguments?
- f. What forecasting methods (scenario, base-rate, model-based, market) have earned the most trust here?

12.3. How should the enormous spread in expert estimates (well under 1% to above 99%) inform us?

- a. What does the *shape* of the P(doom) distribution (bimodal? see Field 2025) tell us beyond its mean?
- b. How much of the spread is genuine information versus selection effects (who becomes a "doomer" or "skeptic")?
- c. Should extreme confidence in either direction be discounted as poorly calibrated?
- d. How much do experts' estimates track expertise versus disposition/worldview (the two-cluster finding)?
- e. Does the persistence of disagreement after debate (XPT, FRI) mean the cruxes are empirical or fundamentally value/worldview-based?
- f. How should a newcomer weight the expert distribution against forming an inside view?

12.4. When inside-view models and outside-view base rates conflict, how much weight should each get?

- a. In which regimes does inside-view reasoning outperform base rates, and does AI fall in one?
- b. How do you avoid both failure modes — anchoring on a bad base rate, and inside-view overconfidence?
- c. Does the unprecedented nature of AI x-risk break the outside view, forcing reliance on inside-view models?
- d. How should model uncertainty (Froolow: parameter uncertainty collapsing conjunctive estimates) change how much you trust any inside-view number?
- e. What is the right way to combine multiple weak models rather than betting on one?
- f. How do you keep this reasoning honest over time — precommitting to cruxes, tracking calibration (the point of this whole audit)?

12.5. How should we prioritize which questions to study — what's the value-of-information calculus?

- a. How do you estimate the "movement × uncertainty" product (the audit's core routing logic) for a given question?
- b. Which questions are high-uncertainty but low-value (interesting but non-decision-relevant) and should be dropped?
- c. How do you account for tractability — some high-value questions may be unresolvable?

- d. Should you prioritize questions where you're miscalibrated over questions where the whole field is stuck?
- e. How do correlations between questions change prioritization (resolving one may resolve others)?
- f. How do you avoid motivated selection of which cruxes to examine?

12.6. What is the right role of formal models versus narrative scenarios in this reasoning?

- a. When do explicit probabilistic models (Carlsmith-style) beat narrative scenarios (AI-2027-style), and vice versa?
- b. How much does the multiple-stage/conjunctive-fallacy problem undermine formal decompositions?
- c. Does parameter/model uncertainty (Froolow) mean formal aggregate numbers should be distrusted?
- d. Are scenarios better for imagination/coverage and models better for discipline — and how to combine them?
- e. How do you avoid false precision from models and false vividness from scenarios?
- f. Should this audit deliberately avoid producing a single aggregate number, and why?

12.7. How do we correct for motivated reasoning and community epistemic failure modes?

- a. How much does selection (who joins "safety" vs. "skeptic" camps) drive the observed disagreement (12.3)?
- b. What are the epistemic failure modes specific to this community (groupthink, deference cascades, doom/hype incentives)?
- c. How do funding and social incentives bias which conclusions get reached?
- d. What practices (red-teaming beliefs, adversarial collaboration, calibration training) actually improve group epistemics?
- e. How do you weight insider (lab) views that are informed but conflicted?
- f. Does this very audit risk anchoring people on its framing, and how to mitigate that?

12.8. How should we incorporate AI systems' own forecasts and reasoning about these questions?

- a. As models become superhuman forecasters, how much should we defer to their x-risk estimates?
- b. How do we handle the conflict of interest when the systems being assessed do the assessing?
- c. Can AI help decompose and calibrate these questions without importing its own biases?
- d. What safeguards keep AI-assisted reasoning about AI risk from being subtly manipulated (2.10, 6.8)?

- e. At what point does refusing to use AI forecasts become its own epistemic failure?

Appendix: mapping to the main post's numbering

The twelve extended domains map back to the main post's ten as follows. Domains 1–3 are unchanged in scope (expanded in depth). **Domain 4 (Evaluation) is new** — its material was previously scattered across main-post 3.5, 3.10, 3.11, and 9.7. Main-post Domain 4 (Takeoff) → extended Domain 5; main-post 5 → extended 6; main-post 6 → extended 7; main-post 7 (Misuse) → extended 8. **Extended Domain 9 (Security) is new** — expanded from main-post 9.5. Main-post 8 (Structural) → extended 10; main-post 9 (Mitigations) → extended 11; main-post 10 (How to reason) → extended 12.

Count

Twelve domains, **143 parent questions**, and **865 subquestions** (5–10 per parent, plus the nested reference-class items under 12.1) — **1,008 discrete, rateable questions** in total. Each is written to be double-sided and, at least in principle, movable by evidence, so that the uncertainty rating and direction column from the main post's protocol can be applied at whichever level of granularity a given reader wants to audit — the whole domain, a parent question, or a single subquestion.