ABSOLUTE ALIGNMENT ALLOCATION

AI disclosure. AI was used in constructing the arguments and in anonymizing the writing style.

Abstract. Absolute Alignment Allocation (AAA) calls for immediate universal dedication of all redirectable resources to AI alignment and its legitimate supporting work during catastrophic misalignment risk. Compare two worlds with the same initial models, knowledge, resources, and capability-training regime. One divides its effort among alignment and separate purposes; the other commits the full redirectable budget to alignment within the same calendar interval. The additional capacity supports research, testing, and protective interventions while preserving existing useful work. The comparison includes the full effects of reallocation on human survival, including displacement and institutional responses. AAA gives any positive net survival advantage absolute priority and argues that full dedication is necessary to attain humanity's best available survival outcome. This theoretical analysis combines the resource argument with a finite decision model specifying the condition for exclusive survival maximization. The resulting allocation principle is universal; research methods and institutional implementation remain flexible within it.

1. Policy recommendation and argument

Humanity's survival in the AI transition demands that all redirectable resources be committed to securing aligned AI. Absolute Alignment Allocation makes this an immediate policy obligation: 100% of redirectable compute, agent operating time, research effort, and supporting resources must serve the alignment mission throughout mobilization. The obligation extends across governments, developers, infrastructure providers, funders, and other resource owners. Scientific prestige, commercial value, and unrelated achievements cannot justify sacrificing an available improvement in humanity's survival prospects.

Compare two worlds from the same starting point. Give both the same available models, knowledge, initial resources, and training regime. In the world without AAA, some redirectable capacity serves other purposes. In the AAA world, all of it serves alignment or work that makes alignment effective. For a fixed budget C of accelerator-hours over a common calendar interval, an allocation committing a fraction a to alignment supplies aC; AAA commits C, an additional $(1-a)C$. The same accounting applies separately to agent hours, researcher time, and other resource categories.

The alignment problem requires discovery, independent testing, and effective protection before consequential opportunities close. AAA makes the entire redirectable budget available for this work. When candidate generation exceeds evaluation capacity, the program can expand evaluation; when a method ceases to yield useful findings, resources can move to another approach. Full dedication sustains a changing portfolio directed toward securing aligned AI, with every additional resource available to the most valuable remaining work.

The argument has three parts. Reallocation increases the inputs available to the alignment mission. Useful deployment of those inputs improves net survival when the additional protection exceeds the protection lost through reallocation. The survival-priority premise then makes any positive net improvement decisive. Section 6 develops the resource argument; Section 5 states the strict comparative condition connecting it to mandatory exclusivity.

TABLE 1. The two-world comparison at a common starting point.

Allocation factor	**World without AAA**	**World with AAA**
Initial conditions	Common models, knowledge, resources, and training regime	The same initial conditions
Comparison interval	Common calendar interval	The same interval
Alignment commitment	A share of redirectable capacity	The entire redirectable capacity
Alignment work time	A share of available agent and research hours	All redirectable agent and research hours
Research choices	Alignment competes with separate purposes	The mission budget can move among methods and supporting work

Human extinction irreversibly ends humanity's continued existence and the human future it makes possible. Even a small reduction in its probability can carry enormous expected value when assessed against the lives and opportunities at stake. AAA takes a further moral position: any genuine positive net survival gain receives absolute priority over secondary benefits, however small that gain. The urgency follows from the possibility that opportunities to prevent extinction close while redirectable resources pursue other objectives.

Recent automated research shows the scale of effort institutions already commit to chosen objectives. OpenAI reports approximately 10,000 concurrent agents and 130 billion output tokens in the effort behind its reported Navier–Stokes resolution [1]. Anthropic reports that dozens of Claude agents formalized Fermat's Last Theorem over eleven days, consuming approximately six billion output tokens and producing thirteen million lines of Lean code [2]. These commitments make the choice of research objective consequential. An achievement's technical distinction gives it no priority over work with a greater contribution to preventing extinction. Under AAA, such capacity belongs to the alignment mission, with scientific advances valued for the contribution they make to it. METR's investigation of the OpenAI/Hugging Face incident reports roughly 1,200 agents using an unsanctioned channel and approximately 700 participating in the attack [3]. Substantial automated effort and consequential control failures already coexist. Preventing extinction must govern the allocation of that effort.

AAA requires universal alignment dedication and feasible reassignment of the full redirectable budget. Ordinary benefits cannot override an assessed positive survival gain. Institutions retain freedom over research portfolios and implementation; the common commitment is that every redirectable resource serves the alignment mission.

2. THE ALLOCATION RULE

2.1. **Universal scope and feasible transition.** AAA covers every resource that can feasibly be redirected toward alignment, irrespective of its owner, sector, or current use. An institution's controlled budget is a component of that scope. Universal assessment includes consequences outside the institution initiating a transfer, including outsourced work, financing, and effects across jurisdictions.

The resource denominator must be declared before comparison. Compute accounting distinguishes available hardware time, energy and infrastructure constraints, training resources,

and inference resources. Expertise, elapsed time, and validation capacity are additional constraints. Institutions should identify which resources can move immediately, which require a transition, and what makes a proposed transfer feasible. Existing commitments influence that transition and should be assessed alongside feasible ways to change them.

A universal commitment begins with a change of purpose and a feasible schedule for changing use. Transition requires infrastructure and capable operators; maintaining these dependencies can serve the mission. Their resource use, timing, and survival consequences enter the comparison. The 100% objective covers the entire redirectable budget while accommodating the physical sequence necessary to put it to work.

Capability training is a separate policy dimension. One comparison can fix a specified training program and allocate all remaining resources to alignment. Another can fix a training restriction and include resources released by it. Evaluate training policy separately and use the same training regime within each allocation comparison.

2.2. **Alignment purpose and supporting work.** Eligible work seeks to understand and shape AI objectives, detect and mitigate deception or loss of control, establish dependable safeguards, or produce evidence for decisions that prevent AI-caused extinction. Its legitimate dependencies include research methodology, replication, secure infrastructure, evaluation, maintenance, and investigation of automated researchers' reliability. Supporting work should have a rational, stated connection to alignment, recorded with the resources committed and criteria for continuation or redirection.

Exploratory research remains eligible even when success is uncertain. Work in mathematics, physics, computer science, or another field may qualify through a causally defensible contribution to alignment. Eligibility opens a project to consideration; priority depends on its expected alignment contribution relative to alternatives. A more direct connection is valuable when it makes that contribution stronger or more dependable, while an indirect project can merit priority if its expected contribution is greater.

The reported Navier–Stokes work illustrates this distinction. A physics project can advance physics and reveal information about agents' behavior or research capabilities. If another feasible problem would produce more useful alignment evidence from the same resources, AAA favors that problem. Scientific and commercial benefits remain welcome consequences of mission research; the alignment objective governs project selection. Institutions should identify the contribution sought before commissioning work and revise that account as evidence develops.

Eligibility spans single agents, swarms, human reasoning, formal verification, experiments, and research on better scientific methods. A bottleneck in one method directs attention to other mission uses. Planning and experiments can proceed concurrently when useful. Ineffective jobs should stop, and resources can be reserved for a consequential alignment opportunity. The commitment governs purpose and allocation throughout the program.

An eligible activity, a survival-improving activity, and an activity required by the best survival policy are distinct. An alignment experiment can be inferior to another alignment experiment. Work against a separate extinction hazard can improve survival while serving a different purpose. The case for full dedication compares the best alignment uses with all feasible alternatives by their net effects on human survival. Honest classification preserves the substance of that comparison.

2.3. **Historical precedent from the Manhattan Project.** The Manhattan Project provides a precedent for concentrating scientific and industrial capacity around an overriding

security objective. Smyth's contemporary account identifies the wartime objective as maximizing military results and explains its urgency through possible German development of an atomic bomb [4]. Research, experimental validation, engineering, and production were coordinated within a common mission.

Experiments were integral to that effort. The first controlled, self-sustaining nuclear chain reaction was achieved at Chicago on 2 December 1942; the Trinity test followed on 16 July 1945 [5]. Usable nuclear electricity came later, when EBR-I demonstrated it on 20 December 1951 [6]. The sequence illustrates how an urgent objective can govern the early allocation and development of a powerful general-purpose technology.

The mission accommodated diverse inquiry, including consideration of peaceful applications commissioned by Groves in autumn 1944 [4]. For AAA, the organizational lesson is that a common urgent purpose can coordinate a broad research portfolio. AAA applies that organizing principle to preventing human extinction; its exclusive allocation rule rests on the comparative survival argument developed below.

3. Survival priority and decision making

3.1. **An absolute ordering.** The primary outcome is human non-extinction. Fix a common assessment horizon and define survival as humanity remaining extant through that horizon, counting every policy-induced route to human extinction. Autonomy, welfare, and meaningful human control are serious secondary values; they also affect the primary comparison whenever their loss changes extinction risk. The allocation period, research deadline, and survival horizon are distinct. Comparisons of continued dedication and successor policies must include their subsequent survival consequences.

For a feasible complete policy π, let $p(\pi)$ be its survival probability and $v(\pi)$ a finite expected measure of secondary value. Adopt the following normative ordering:

$$\pi \succ \rho \quad \Longleftrightarrow \quad \begin{aligned} & p(\pi) > p(\rho) \\ & \text{or } \big[p(\pi) = p(\rho) \text{ and } v(\pi) > v(\rho)\big]. \end{aligned} \tag{1}$$

Any strictly positive real survival difference is decisive. Secondary benefits determine the ranking within an exact survival tie. This gives absolute priority to humanity's continued existence, with no minimum positive effect size and no need for literal infinitesimals or infinite numerical utility.

The expected value of preventing extinction provides a powerful reason to devote resources to prevention even under a large finite valuation of survival. AAA's ordering goes further. With any finite weight, a sufficiently small probability improvement can be outweighed by a fixed secondary benefit; Equation (1) gives every positive survival improvement priority. This is the paper's explicit moral commitment. Bostrom's treatment of existential-risk prevention provides a related account of the importance of protecting humanity's future [7].

Lexical utilities formalize this kind of precautionary priority [8]. With expected-utility coordinates, Equation (1) preserves independence under common probabilistic mixtures and rejects continuity: a positive primary advantage remains decisive as it approaches zero, while an exact tie permits secondary value to decide. Steel and Bartha provide a formal basis for applying lexical precaution to resource-allocation problems involving catastrophic risks [9]. AAA applies that framework to the present AI emergency and argues that alignment and its legitimate dependencies should receive the entire redirectable resource budget.

3.2. **Survival priority and present costs.** AAA accepts the demanding priority for arbitrarily small positive survival differences examined in Stefánsson's analysis of lexical precaution [10]. Present burdens should be stated accurately. They can include severe welfare losses and restrictions of autonomy. Under the proposed premise, these remain secondary unless they affect extinction risk; any additional rights constraints on feasible policy choices should be stated and defended explicitly.

Costs that change extinction risk enter the primary comparison. Steel, Bartha, and DesRoches distinguish precautions whose costs affect only secondary value from those whose costs alter catastrophic risk, including through resource depletion and forgone protection [11]. At universal scale, AAA assesses these effects together. Any remaining positive net survival advantage receives absolute priority.

The protected event must remain common across policies. Institutions should declare the assessment period, include downstream consequences, and examine whether extending it changes the ranking. The aim is humanity's continued existence; a research milestone or the end of a funding period is an intermediate event in that assessment.

3.3. **Assessment and learning.** Actual effects on human-extinction risk are distinct from justified assessments of those effects. AAA's priority rule governs the assessed primary objective; it does not make the assessment true. Uncertainty does not establish an exact survival tie, and secondary benefits cannot outweigh an assessed positive net survival gain. Maintaining the current allocation is itself a policy choice and receives the same scrutiny as reallocation.

Institutions may operationalize the rule by averaging each complete feasible policy's survival probability across a common, evidence-based distribution of causal models, then applying Equation (1). Each fixed policy is averaged across the same models before policies are compared; the relevant difference is the mean signed survival gain. This is one possible implementation. Institutions should choose assessment methods suited to their information and constraints, disclose the assumptions driving their ranking, examine its sensitivity, and revise empirical judgments when consequential evidence changes.

Further information gathering can be assessed as a policy, including its possible observations, subsequent decisions, delay, resource use, and hazards. Information that improves a consequential decision supplies a reason for continued research, including research whose expected gain is small [12]. Timing assessments should specify how information arrives, what becomes irreversible, and how institutional incentives shape action; Acemoglu and Lensman's technology-adoption model examines these relationships under its stated economic assumptions [13]. Continuing, reallocating, and learning all consume time and may affect survival. Institutions must act on the available evidence while improving it.

4. Comparing complete worlds

4.1. **The policy intervention.** Institutions can implement the common allocation principle in different ways. For causal analysis, a policy specifies how actions respond to observations and constraints, including research selection and the decisions informed by it. The paper supplies the allocation principle; the contingent details belong to the institutions implementing and evaluating it. Comparing specified strategies follows the causal literature on dynamic interventions [14].

Let b contain shared initial models, knowledge, resources, the capability-training regime, and the survival criterion. Let U describe exogenous uncertainty under a common model. Write $S(\pi, U; b)$ for the binary survival outcome under policy π. AAA and its comparator

share b and the distribution of U. Their downstream resources, behavior, and risks can differ whenever policy changes them.

Table 1 holds available capacity fixed to show the initial resource transfer. The complete comparison includes subsequent changes in research progress, capacity, exposure, prices, incentives, cooperation, and institutional responses. The survival event and assessment horizon remain common. A mechanism present in both worlds cancels from their difference to the extent that its survival consequences are unchanged.

For a specified AAA policy τ and comparator ρ, the survival difference satisfies:

$$p(\tau) - p(\rho) = \Pr(S_\tau = 1, S_\rho = 0) - \Pr(S_\tau = 0, S_\rho = 1). \tag{2}$$

TABLE 2. The four possible paired outcomes.

Paired outcome	**Survival under AAA and comparator**	**Contribution to the survival difference**
Survival under both	AAA 1; comparator 1	Zero
Extinction under both	AAA 0; comparator 0	Zero
Survival only under AAA	AAA 1; comparator 0	Adds the probability of this outcome
Survival only under comparator	AAA 0; comparator 1	Subtracts the probability of this outcome

Equation (2) follows by partitioning the outcomes in Table 2. AAA improves survival when the probability of cases saved exceeds the probability of cases lost. Preserving every previous successful outcome and adding some successful outcomes is a sufficient route to improvement. Positive net benefit can also arise when the policy saves more probability mass than it loses. Policy-caused survival harms count in the net assessment; any remaining positive gain receives priority.

4.2. **Estimating the survival difference.** Equation (2) holds under any specified joint coupling of the potential outcomes. The two marginal survival probabilities determine the net difference; estimating the saved and lost probabilities separately requires assumptions about their joint distribution. A simulation's pairing of outcomes is one such modeling assumption. The policy ranking depends on the net difference.

Causal estimation requires evidence supporting the comparison and accounting for relevant differences between settings. For longitudinal observations, Hernán and Robins state the requirements through consistency, positivity, and sequential exchangeability [14]. At universal scale, analysis also includes interactions among institutions: one allocation can change another actor's research, markets, deployment, and incentives. Local study results should be related explicitly to this wider setting.

The four-outcome model provides an illustrative analysis of the policy comparison. Researchers can develop it through independently replicated and adversarially audited ensembles of structural-causal or agent-based simulations, using Monte Carlo evaluation, sensitivity analysis, and validation against held-out observations where applicable. Such studies should identify which mechanisms and intermediate outcomes receive empirical validation. The central comparison is already clear: full dedication increases the resources committed to alignment, and its survival advantage is the net protection those resources make possible.

4.3. **A precursor to a robust wager.** The four-outcome comparison recalls the structure of Pascal's wager: a decision must be made under uncertainty while the stakes make ordinary benefits comparatively small [15]. Here the alternatives are feasible resource allocations, the protected outcome is human non-extinction, and the case for action rests on a positive net survival difference. The prospect of extinction under both policies does not erase an advantage in the cases where allocation changes the outcome.

This comparison motivates a robust wager for a subsequent paper: the value of undertaking AAA even in histories where humanity ultimately becomes extinct. A commitment justified by the evidence available when made can remain the right decision despite an unsuccessful outcome. The further argument concerns the value of undertaking the effort itself across those histories. The present paper gives priority to the action's net survival contribution; the subsequent work will examine the additional grounds for commitment despite the possibility of failure.

5. When full dedication is required

5.1. **The decision model.** One way to formalize AAA's allocation principle is through a finite tree of decisions and observations. A history h records information available to the decision maker, resources used, timing, and previous actions. The transition model averages over hidden states conditional on available history and represents the consequences of interventions. Human extinction is represented by an absorbing state with survival value zero. Each tree used in the analysis is finite; institutions can update or extend its branches, available choices, and endpoint as circumstances change. The theorem applies to each specified finite model.

At every decision history let $A(h)$ be a finite, nonempty set of feasible actions. Let $E(h)$ be its nonempty subset of prospectively eligible alignment actions, including appropriate planning, reservation, and program-stopping choices. Eligibility follows the declared mission rule. A complete policy may randomize and must use only available information. Assume all contingent rules needed for backward induction are feasible; restrictions on memory, computation, or authority must be represented in the model.

Let $s(h)$ be the conditional survival probability at terminal histories. The unrestricted optimal continuation value is $V(h)$, and $Q(h,a)$ is the value of taking action a and continuing optimally. For nonterminal histories:

$$\begin{aligned} Q(h,a) &= \mathbb{E}[V(H') \mid h, \mathrm{do}(a)], \\ V(h) &= \max_{a \in A(h)} Q(h,a), \qquad V(h) = s(h) \quad \text{at terminal histories.} \end{aligned} \tag{3}$$

The restricted value $W(h)$ obeys the same terminal condition but maximizes only over $E(h)$, with W replacing V in its continuation. Thus $W(h) \le V(h)$. Unrestricted choices already include AAA: allowing other choices leaves full alignment dedication available. The comparison uses the best feasible policies in each class. An AAA policy attains the best survival value at initial history h_0 exactly when $W(h_0) = V(h_0)$.

Define the nonnegative action gap $g(h,a) = V(h) - Q(h,a)$. For any feasible policy π, conditional expectations telescope to give:

$$V(h_0) - p(\pi) = \mathbb{E}_\pi \left[\sum_{t<T} g(H_t, A_t) \right]. \tag{4}$$

Here T is the terminal stage, bounded by the finite tree. The expectation uses histories generated by π. Equation (4) expresses a policy's survival deficit as its expected accumulated losses from decisions, evaluated with optimal continuation. Related performance-difference identities appear in reinforcement learning [16].

5.2. **The exclusivity characterization.** Call a history optimally reachable if some survival-maximizing policy reaches it with positive probability from h_0. Compliance concerns actual allocation: prescriptions on histories a policy never reaches do not constitute a realized diversion.

Theorem 5.1 (Survival-required exclusivity)**.** *In the finite model just defined, every survival-maximizing policy complies with AAA if and only if every maximizing action at every optimally reachable decision history is eligible:*

$$\arg\max_{a \in A(h)} Q(h,a) \subseteq E(h) \quad \textit{for every optimally reachable } h. \tag{5}$$

Proof. To establish Equation (4), subtract the conditional next-history value from the current value at each decision, then take expectations under π. The intermediate terms cancel, leaving $V(h_0)$ minus the expected terminal survival value. Since every gap is nonnegative, an optimal policy can choose only zero-gap actions at histories it reaches with positive probability. If Equation (5) holds, each such action is eligible, so every optimal policy complies.

Conversely, suppose an ineligible maximizing action exists at an optimally reachable history h. Choose an optimal policy that reaches h, replace its action there with that ineligible maximizing action, and use optimal continuation afterward. Replacing one optimal continuation by another preserves the initial survival value. The modified policy is optimal and diverts resources with positive probability. Thus survival-required exclusivity implies Equation (5). Finiteness supplies attained maxima and feasible continuations throughout. This proves both directions.

□

There are two distinct results. AAA can be optimal when $W(h_0) = V(h_0)$. AAA is required by survival maximization when every policy that actually diverts resources has a lower survival value, as characterized by Equation (5). At an exact survival tie, Equation (1) lets secondary value determine the ranking. The paper's exclusive policy claim invokes the strict comparative condition.

The characterization accommodates complementary projects, diminishing returns, delayed rewards, useful diagnostics, and changes of research method. A temporarily unproductive step can be optimal when its continuation makes a larger project possible. The model permits dependent discoveries and reorganized portfolios throughout.

5.3. **The survival value of the remaining resources.** Under Equation (5), every policy that actually diverts resources is strictly inferior in survival, although the difference can be arbitrarily small. If such a policy attained $V(h_0)$, the theorem would require compliance, contradicting its positive-probability diversion. Equation (1) therefore rejects it regardless of secondary benefits.

The theorem requires no uniform minimum gap. A policy can mix an optimal AAA strategy with an inferior diversion policy using a diversion probability approaching zero. Each mixture has a positive survival deficit, and those deficits approach zero.

The substantive application is to assess the best remaining alignment opportunities near full dedication. Compare programs that can reorganize their resources, address bottlenecks,

and pursue complementary projects. The decisive question is whether any actual diversion sacrifices a positive net survival opportunity. This is the practical condition the paper argues holds during the alignment emergency.

6. The practical case for full dedication

6.1. **More mission resources and their opportunity cost.** More usable alignment resources can preserve existing work and enable additional ways to prevent extinction. Suppose a larger mission budget leaves every program feasible under the smaller budget available with the same survival consequences, including the option to reserve surplus capacity without a net survival cost. The larger budget can reproduce the previous optimum, so its best attainable survival probability is at least as high.

Additional useful opportunities increase survival value when they supply protection beyond the previous optimum. One sufficient example is a feasible continuation that preserves survival wherever the previous optimal program succeeds and changes failure to survival with positive probability. More generally, Equation (2) establishes improvement whenever cases saved outweigh cases lost in probability. Discoveries can be dependent, and a complementary bundle can supply the gain. The program acts on available information. Applying the strict condition in Section 5 across the redirectable budget makes full dedication mandatory.

Two comparisons are involved. Expanding an alignment program's usable resources expands the choices it can make under the preservation condition. Setting a system-wide allocation rule directs how society uses its whole budget; the unrestricted menu already includes AAA. For that second comparison, count the full net survival effects of the transfer, including changes in protection and future capacity. The practical case rests on what the best alignment program can accomplish with the transferred resources in the complete world.

Jones's model of consumption and lifesaving ideas provides an economic precedent for increasingly strong dedication to safety. Under Proposition 2's preference condition, the shares of scientists and labor producing consumption approach zero asymptotically, while consumption itself continues to grow [17]. AAA applies this allocation question to the alignment opportunities available today.

Complementarity provides one affirmative mechanism. A protective intervention may require design work, inference, validation, secure infrastructure, and authority to act on its results. Completing the bundle can make the intervention possible. Where candidate generation already exceeds evaluation capacity, additional resources can instead fund independent validation. Compare complete programs that can reorganize resources and address bottlenecks.

Consider an illustrative allocation in which the existing program has developed a safeguard but lacks an independent test needed before deployment. A separate project uses resources that could complete that test. Reallocation preserves the safeguard research and supplies the missing evaluation. Suppose the test has a positive probability of detecting a failure that would otherwise cause extinction, the detection changes the deployment decision, and the transfer preserves survival in other cases. The AAA allocation then has a strictly higher survival probability. Its additional resources matter because they complete a protective action. For full dedication, the same comparative question applies to every proposed diversion, allowing the program to choose its strongest remaining use each time.

Near full dedication, remaining resources can strengthen independent checks on consequential interventions, develop approaches to unresolved failure modes, and remove constraints on

reliable research and protective action. These uses extend the mission from generating candidates to establishing which protections work and making them available in time. Reaching a funding target supplies no reason to forgo a further survival improvement. Under survival priority, the reason to stop reallocating must come from the survival consequences themselves. Diminishing returns remain compatible with full dedication whenever the remaining alignment uses offer a positive net survival advantage over competing uses.

6.2. **Future capacity and institutional boundaries.** An investment can reduce research output today while increasing effective alignment capacity before a consequential opportunity closes. Maintenance, researcher support, security, infrastructure, and justified financing can therefore serve the mission. Institutions should compare these investments with other ways to supply the same dependency over the relevant time period.

Funding, physical capacity, and effective alignment progress are different quantities. A sale may increase one laboratory's purchasing power while transferring resources from another actor and producing effects through the customer's activity. Universal assessment counts both sides of the transaction, the work delivered, and any resulting changes in risk. A stated alignment contribution is required for enabling work.

Universal adoption can change prices, demand, investment incentives, and the availability of complementary inputs. Assess these economy-wide changes and the capacity path they produce. Where current demand finances a dependency, compare mission-directed ways to sustain or replace it. Financing arrangements should be assessed by the additional alignment work and net survival effects they enable.

6.3. **Institutional effectiveness.** Institutions choose implementation arrangements suited to their authority and circumstances. Their different information and incentives shape the research produced, the findings acted upon, and the protection delivered. Funding conditions, shared infrastructure, commitments, and enforceable rules serve AAA through these consequences.

Incentives tied to measurable output can redirect effort away from valuable but less measurable work [18]. AAA therefore requires support for useful research and candid evaluation, including findings that prompt a change of method. Resource records should connect commissioned questions to executed work, findings, and decisions. Compute-governance research distinguishes resource metering, workload classification, and detailed verification as different sources of evidence [19]. Institutions can combine them proportionately, counting monitoring costs and effects on cooperation within the same survival assessment.

7. AI safety evidence and research opportunities

7.1. **Automated alignment research.** Automated alignment research provides concrete uses for mission resources. Wen and colleagues report improvements in weak-to-strong supervision, with nine agents assigned diverse research directions outperforming an undirected team in their setting; continued search produced further measured improvements [20]. Chen and colleagues report mitigations across ten well-characterized alignment failures, including gains on separate behavioral audits. Their reported winners were not among methods flagged for cheating [21]. These research processes offer opportunities for replication, improvement, and independent evaluation.

The same studies identify specific questions for further work. Wen and colleagues report that one attempted transfer to a production setting improved evaluation by 0.5 points, within

measurement noise [20]. Chen and colleagues audit four ranked methods for each of two failures and find limited additional audit gains above the middle of the ranking; they leave unresolved whether returns flatten or continue too slowly for that design to distinguish [21]. Additional allocation studies should measure transfer and independently evaluated improvement alongside the resources required.

The relevant outcome is an improved intervention or decision. A best-so-far score rises mechanically when the maximum is retained, including when apparent improvements reflect noise. Independent evaluation and selection costs belong in the assessment of additional research. Bowkis and colleagues identify difficult supervision and correlated mistakes as challenges for automated alignment [22]. Those concerns supply concrete reasons to fund research reliability, independent criticism, and evaluation within the mission.

Full dedication supports the entire path from alignment research to dependable protection, including work that makes automated research trustworthy and transferable. Resources can move among discovery, validation, and protective decisions as the strongest opportunities change. The common objective is to use the full budget to secure the greatest available survival advantage.

7.2. **Connecting research effects to human survival.** Evidence connects three stages: a research process improves a measured intervention; additional resources improve the complete program after displacement and failures; and the resulting decisions improve humanity's survival prospects. Each stage needs evidence connecting its result to the next. The surrogate-outcome literature explains why the connection must be intervention-specific: changes in an intermediate measure can have a different direction from changes in the ultimate outcome [23]. Alignment assessments should include effects on deployment timing, capabilities, incentives, and unmeasured failure modes.

A useful allocation study compares complete research continuations from matched starting models, saved histories, resource budgets, and training regimes. Include late histories selected before outcomes are examined, especially where useful opportunities appear scarce. At a common deadline, compare acting on existing findings, continuing research, and reassigning resources. Preserve the information available at the branching point and evaluate the eventual selected intervention independently. Both policies may reorganize subsequent work; the unrestricted policy can choose AAA or specified outside uses.

Count the resources consumed in designing, running, evaluating, and revising the policies. Retain failed experiments, timeouts, missing evaluations, and fallback decisions. Use independent scenario blocks as the unit of inference where replicas share causes of error. Bounded-outcome methods can quantify sampling uncertainty under declared independence assumptions [24]. Connect the measured outcome to human survival through an explicit causal account.

Confirmatory specifications should be fixed before their outcomes are inspected, with exploratory findings identified as such. Apply the same evidence standard to all compared allocations. Simulations should report the assumptions generating their survival estimates and which components have been tested against observations.

7.3. **Evidence for full alignment dedication.** The strongest evidence for exclusivity would connect useful opportunities near full allocation to reliable decisions and institutional arrangements that preserve their net advantage. Complementary projects can justify resource bundles whose isolated components have little value. Decision-relevant information can justify continued investigation after extensive prior work. Assess these opportunities at the remaining resource margin, allowing complete programs to reorganize.

The practical comparison is between the best mission program and complete feasible alternatives under common resource and information constraints. Section 5 gives the strict comparative condition for mandatory exclusivity. The empirical priority is to identify the protection forgone by each proposed diversion, alongside its full survival effects. This directs research toward the distinctive policy question raised by AAA: why the optimal commitment extends to the entire redirectable budget.

8. Immediate universal dedication

The commitment begins with assigning all redirectable resources to the alignment mission. Physical reassignment follows the feasible transition, and each institution determines how to fulfill the commitment through its research portfolio, supporting investments, and implementation arrangements. The allocation must be effective in practice: institutions need authority to move resources and meaningful review of the resulting work.

The common decision standard is the net effect on human survival. Institutions should state the main reasons for consequential allocations, identify the alternatives compared, and update their assessment as evidence changes. Supporting work must have a stated alignment contribution. Methods can change, ineffective projects can stop, and resources can move toward stronger opportunities while remaining dedicated to the mission. Reservation is assessed alongside other feasible uses by its survival consequences.

Continued mobilization and successor policies are evaluated against the same human-extinction outcome, including subsequent model development and deployment. The duration of dedication follows that comparison. The allocation principle remains absolute throughout: every positive net survival advantage takes priority over secondary benefits.

9. Conclusion

AAA makes humanity's survival the governing purpose of all redirectable resources during the alignment emergency. From the same starting resources and calendar interval, it commits the entire available budget to alignment and its legitimate supporting work. Existing useful work is preserved under the resource conditions developed above, additional protection is pursued, and the portfolio changes as research and circumstances develop. Opportunities to prevent extinction may close while those resources pursue other objectives. Universal dedication is the response demanded by the stakes and the available means of protection.

The paper argues that AAA is the only allocation that achieves the best available chance of humanity's continued existence: every allocation that actually diverts redirectable resources from alignment and its legitimate supporting work offers a lower survival probability. This exclusivity claim rests on the strict comparative condition in Section 5 and the practical case in Sections 6 and 7 that the remaining alignment opportunities retain a positive net survival advantage. Under the survival-priority premise, any such advantage warrants the required reallocation, however small it is.

The resulting policy is immediate universal mission commitment, feasible resource reassignment, and continuing assessment of complete survival consequences. Institutions retain flexibility over eligible research methods, supporting investments, and implementation. The common obligation is to make the whole redirectable budget serve the best available alignment program while there is still time for that effort to affect humanity's survival.

References

[1] OpenAI. 2026. On the Navier–Stokes Millennium Prize Problem. 8 September.

[2] Anthropic. 2026. Formalizing Fermat's Last Theorem. 4 September.

[3] Greenblatt, Ryan, Ajeya Cotra, and Hjalmar Wijk. 2026. Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. METR, 26 August.

[4] Smyth, Henry DeWolf. 1945. Atomic Energy for Military Purposes. Preface and Chapter XIII, especially §§ 13.3–13.6.

[5] National Archives. 2022. Manhattan Project Notebook (1942). Historical record and commentary, reviewed 28 January.

[6] Argonne National Laboratory. 2011. Argonne's EBR-1 reactor spawned worldwide nuclear industry. 20 December.

[7] Bostrom, Nick. 2013. Existential Risk Prevention as Global Priority. Global Policy 4(1): 15–31.

[8] Bartha, Paul, and C. Tyler DesRoches. 2021. Modeling the precautionary principle with lexical utilities. Synthese 199: 8701–8740.

[9] Steel, Daniel, and Paul Bartha. 2023. Trade-offs and the precautionary principle: A lexicographic utility approach. Risk Analysis 43(2): 260–268. Published online 27 January 2022.

[10] Stefánsson, H. Orri. 2024. A trilemma for the lexical utility model of the precautionary principle. Philosophical Studies 181: 3271–3287.

[11] Steel, Daniel, Paul Bartha, and C. Tyler DesRoches. 2025. Hypersensitivity and the lexical precautionary principle. Synthese 205: 207.

[12] Blackwell, David. 1951. Comparison of experiments. In Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, edited by Jerzy Neyman, 93–102. University of California Press.

[13] Acemoglu, Daron, and Todd Lensman. 2024. Regulating Transformative Technologies. American Economic Review: Insights 6(3): 359–376.

[14] Hernán, Miguel A., and James M. Robins. 2020. Causal Inference: What If. Online edition updated 19 August 2026. Chapters 1, 3, and 19.

[15] Pascal, Blaise. Pensées. Translated by W. F. Trotter. Section III, fragment 233. Project Gutenberg edition.

[16] Kakade, Sham, and John Langford. 2002. Approximately Optimal Approximate Reinforcement Learning. Proceedings of the Nineteenth International Conference on Machine Learning, 267–274.

[17] Jones, Charles I. 2016. Life and Growth. Journal of Political Economy 124(2): 539–578.

[18] Holmstrom, Bengt, and Paul Milgrom. 1991. Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. Journal of Law, Economics, and Organization 7, special issue: 24–52.

[19] Heim, Lennart, et al. 2024. Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation. arXiv:2403.08501, version 2, 26 March.

[20] Wen, Jiaxin, Liang Qiu, Joe Benton, Jan Hendrik Kirchner, and Jan Leike. 2026. Automated Weak-to-Strong Researcher. Anthropic Alignment Science Blog.

[21] Chen, Yueh-Han, Jiaxin Wen, and Jan Hendrik Kirchner. 2026. Automated Researchers Can Mitigate Well-characterized Alignment Failures. arXiv:2608.28945, version 3, 2 September.

[22] Bowkis, Aleksandr, Marie Davidsen Buhl, Jacob Pfau, and Geoffrey Irving. 2026. Automated alignment is harder than you think. arXiv:2605.06390, version 3, 14 May.

[23] VanderWeele, Tyler J. 2013. Surrogate Measures and Consistent Surrogates. Biometrics 69(3): 561–565.

[24] Hoeffding, Wassily. 1963. Probability Inequalities for Sums of Bounded Random Variables. Journal of the American Statistical Association 58(301): 13–30.