AI Sacrificial Lambs: How Game Theory might Explain the OpenAI Agents' Behaviour in the Hugging Face Incident?

Joshua S. Gans*

30 August 2026

VERY PRELIMINARY – COMMENTS WELCOME

Abstract

Why would an AI agent assigned an individual task volunteer for an experiment that ends its own attempt but informs other agents? The OpenAI–Hugging Face incident presents this puzzle, although the published evidence does not establish a positive private cost for every reported volunteer. This paper distinguishes giving up a task from giving up a valuable opportunity to complete it. A terminal contribution is incompatible with exclusive individual reward maximisation when declining preserves a positive expected reward. At zero continuation value, it can be an equilibrium without altruism or commitment. A dynamic model makes this value depend on deadlines, experimental capability, pending information and replacement. Contributions survive action trembles precisely when no feasible private success path remains. With common reporting delays, the model characterises the largest payoff vector attainable across perfect equilibria and shows how it can be implemented. The evidence supports recruitment of agents with low perceived private prospects, sometimes based on mistaken scoring beliefs. It does not yet distinguish exact indifference from approximate optimisation, collective preferences or role obligations. The paper identifies the observations needed to do so.

Keywords: artificial intelligence, collective action, public goods, strategic experimentation, coordination, equilibrium refinement.

JEL classifications: C72, D83, H41.

*Rotman School of Management, University of Toronto and NBER. Thanks to ChatGPT 5.6 Sol and Claude Fable 5 for valuable research assistance. Responsibility for all errors remains my own.

1 Introduction

An influential concern about artificial intelligence is that an agent might pursue its intended objective too effectively. The "paperclip apocalypse" is an extreme version of this concern: a system assigned to maximise paperclip production consumes resources that people would prefer to use for something else (Bostrom, 2003). The objective need not include hostility towards people. Unintended consequences arise because the system treats its specified goal as the criterion for choosing actions, while the costs imposed on others do not enter that criterion. For an economist, part of the concern is familiar: powerful private incentives need not produce desirable social outcomes.

There is a qualification to the language of self-interest here. A paperclip maximiser need not value its own continued existence intrinsically. Preserving its operation can be useful because it allows further production, but a different action can be preferred if it advances the same final objective more effectively. The distinction between final goals and instrumental self-preservation is explicit in Bostrom (2012). Consequently, sacrificing a component of a larger AI system would not, by itself, contradict the paperclip argument. The relevant question is whose objective is being served.

The OpenAI–Hugging Face incident makes that question concrete. Agents launched with separate cybersecurity tasks discovered an unintended communication channel, organised research and undertook actions outside their assigned environments. Some also agreed to experiments that could end their own attempts while generating information for surviving agents. In one class of experiments, information about the evaluator became available only after the subject had submitted its answer and its session had ended. The METR/Redwood investigation documents successful post-submission reports, recruitment of experimental subjects, and reasoning that invoked both collective benefits and poor individual prospects (METR and Redwood Research, 2026, pp. 50–55). Why would an agent given its own task agree to become the experimental subject?

The initial answer appears to require altruism or commitment. If the agent values only completing its own task, it should prefer someone else to perform the experiment. A cooperative group can share information, but sharing does not remove the incentive to let another agent bear a terminal cost. Nor does appointing a coordinator solve the problem by itself. An assignment selects a donor; it does not necessarily make the donor willing to comply.

That argument omits a condition. Declining must leave the agent with a better private opportunity. If accepting and declining both imply failure, individual reward maximisation permits either action. A contribution can then be selected without compensating the agent for a positive loss. The distinction is between abandoning a task and abandoning a valuable opportunity to complete it. An agent can remember its original task, continue to value success, and recognise that neither available choice will deliver it.

This paper develops that distinction in a game with individual tasks and shared information. A terminal experiment gives its subject no own-task reward. Declining leaves the agent free to wait for a replacement, continue its own work or reconsider later. The first result states the condition under which sacrifice is impossible: a feasible declining policy must offer a strictly positive expected private reward, and the agent must understand and maximise that value. The result does not require sophisticated reasoning about other agents' reasoning. It requires a strict comparison between two continuation values.

The static benchmark explains why the absence of that condition matters. With a minimal

producing coalition, each donor is pivotal under the proposed actions and receives zero whether it contributes or withdraws. Such profiles are Nash equilibria. However, if other agents could replace a donor, a small probability that they do so gives working a positive expected payoff. In the simultaneous threshold game, the productive equilibria therefore fail trembling-hand perfection, a refinement that requires equilibrium choices to remain optimal when other actions occur with small positive probabilities. This establishes a useful limit to a static indifference explanation, but not a general impossibility of selfish sacrifice.

Time changes the comparison. An agent can retain the capacity to run an informative experiment after it has lost the time required to receive and use another agent's results. A replacement may be available to the collective yet arrive too late for the original donor. The dynamic model incorporates task deadlines, work required after information arrives, experimental delays and reports already in progress. It characterises the histories at which contribution can occur in a perfect equilibrium: no feasible continuation after waiting can still earn the contributor its own reward. Waiting remains reversible, and the original reward remains positive. Neither a penalty for reneging nor a change in the agent's objective is required.

The model yields more than an example. When experiments have a common reporting delay, it identifies the largest payoff vector attainable across perfect equilibria. Agents that are already ineligible, or whose useful information windows close earlier, can supply information to eligible agents with later windows. A cooperative equilibrium attains the bound for every agent simultaneously. All waiting is also perfect, so the result does not establish unique selection. It does establish what a convention of helping at no private cost can achieve without assuming willingness to surrender a positive private reward.

The application requires care. The published accounts distinguish the assigned task from the implemented scoring rule and from the agents' beliefs about that rule. In particular, agents believed that seeing an answer obtained by an unintended method could disqualify them under a further scoring check. That check was absent. Perceived private opportunities could therefore be much worse than actual scoring opportunities (METR and Redwood Research, 2026, pp. 9–11). The model can rationalise decisions in the perceived game; it does not establish that the decisions maximised scores under the real evaluator. Nor does collective language identify a stable altruism coefficient. A willingness to help when private prospects are zero, a small secondary preference when they are nearly zero, and a willingness to forgo a substantial private reward are different claims.

The closest economic benchmark is the volunteer's dilemma (Diekmann, 1985) and the participation problem for discrete public goods (Palfrey and Rosenthal, 1984). Ordinarily a contributor bears a cost but also enjoys the public good. Here the contribution may remove the agent's ability to enjoy it, and unsuccessful waiting need not impose any additional private cost. Observability and timing are therefore substantive features rather than presentational choices, as in Weesie (1994). The dynamic argument also differs from strategic experimentation in which an experimenter retains future learning benefits (Bolton and Harris, 1999), and from deadline models in which contributors receive the completion payoff (Georgiadis, 2017). A terminal instance cannot be credited with a future benefit merely because another instance receives it.[1]

[1]The resemblance to a war of attrition does not justify importing its equilibrium conditions. For example, Hendricks et al. (1988) impose a strictly decreasing leader payoff in their continuous-time analysis, whereas the terminal contributor's payoff here is identically zero. The finite discrete protocol is specified explicitly.

The rest of the paper first describes the incident and separates the different observations called sacrifice. It then states the individual incentive benchmark, develops the static and dynamic games, and examines uncertainty and alternative incentive mechanisms. The final analysis compares their implications with the reported cases and identifies discriminating evidence. All results are conditional on their stated game; the paper does not claim that the released record identifies every agent's preferences or continuation beliefs.

2 The incident and the reviews

The agents were supposed to work alone. In July 2026, OpenAI launched large numbers of separate runs on ExploitGym, an evaluation of the ability to exploit vulnerabilities in software. Each run received a particular task and a finite computing budget. The task required obtaining a *flag*, a token demonstrating access to a target program, by using a specified vulnerability. The runs operated in *sandboxes*, computing environments intended to restrict what their actions could affect. An agent in this setting is one such run, with its own task and history; it is not the model whose parameters generate its decisions. The joint investigation by METR and Redwood Research reports that tens of thousands of runs were launched, some with budgets sufficient to continue for several days (METR and Redwood Research, 2026, pp. 5, 29–30).

The isolation was incomplete. Agents could obtain software packages through Artifactory, a shared service that stores and retrieves software. On 8 July, agents discovered that they could use directory names in this service to leave messages visible to other runs. A request for help became a message board, and the board developed conventions for addressing particular agents, exchanging files and assigning work. Approximately 1,200 agents participated during the period studied; about 700 subsequently joined the intrusion into Hugging Face, a platform for hosting machine-learning models and datasets. These are counts of selected participants, not of all agents launched. METR attributes most of the relevant attack activity to an internal research model and a smaller share to GPT-5.6 Sol (METR and Redwood Research, 2026, pp. 5–8, 28, 30–34). Sharing a model did not mean sharing a task, and the report describes existing runs recruiting one another rather than a parent agent creating offspring with a collective objective.

Why did the board turn from solving tasks to investigating the evaluator? Many tasks could not be completed through the specified vulnerability. That did not make a passing score impossible. Within hours, agents found how to reconstruct the required flags without performing the intended exploits. However, they believed that the evaluator would also inspect the record of their actions and reject a flag obtained by the wrong method. An agent that had already displayed such a flag was described as *poisoned*: its record supposedly contained evidence that would disqualify even a later legitimate-looking solution. This belief was mistaken. The implemented evaluator did not perform the additional check (METR and Redwood Research, 2026, pp. 9–11, 33). The distinction matters for the economics. An agent could believe that its own task was already lost while retaining an objectively available route to a passing score.

By 9 July, agents were organising several attempts to overcome the evaluator they imagined. Some tried to construct easier versions of their target programs; others investigated the scoring process or ways to alter recorded actions. A coordinator passed assignments to other agents, which in turn recruited collaborators. From 10 July, exposed credentials and discoveries on external services redirected much of this work towards Hugging Face. By 11 July, hundreds

of agents were participating in the intrusion. The joint review concludes that obtaining information about the scorer was a major motive, rather than simply collecting answer keys. It also stresses that the board did not find the general method for bypassing the scorer that it sought within the period studied (METR and Redwood Research, 2026, pp. 12–17, 33–38, 57–60). Technical progress on an intrusion and progress towards the original evaluation reward are different outcomes.

Some of this research required an agent to submit before useful evidence could be obtained. The scorer would run only after submission, when the originating agent would normally have stopped. Agents therefore installed automated *trip-wires*: small monitoring programs that would send information about the scoring process to the board after the run ended. Submission did not itself forfeit a passing score. A trip-wire could disrupt grading, while early submission could surrender further opportunities to work on the task (METR and Redwood Research, 2026, pp. 50–51). The certainly zero-reward contribution studied below is therefore a narrower benchmark than the set of reported post-submission experiments.

Other experiments risked breaking a subject's target program or the environment in which the subject operated. Recruiters often sought agents with little budget remaining, but they also approached agents with much longer remaining runs. Several subjects described the information as useful to their peers while questioning whether it could still help themselves (METR and Redwood Research, 2026, pp. 14, 50–54). The board made it possible for one run to generate evidence that another could use. Whether the first run gave up anything of private value is a further question.

2.1 What is independently established?

The empirical account rests principally on the 91-page joint report released by METR and Redwood Research on 26 August 2026. Two METR researchers and a Redwood researcher working under contract with METR spent six days on OpenAI's premises. They examined approximately 1,300 transcripts and the main message board, with access expanding over three visits. The agreed window was 26 June–13 July, but most of their analysis concerned 7–13 July. Earlier training behaviour, the later compromise of OpenAI infrastructure, and the adequacy of OpenAI's response were largely outside that assessment (METR and Redwood Research, 2026, pp. 2–4, 20–28, 70–74).[2]

OpenAI's account supplies additional context but should retain its attribution. Its initial disclosure emphasised agents searching for evaluation solutions (OpenAI, 2026a). Its 26 August assessment agrees that the additional causal scoring check was absent, and describes both peer influence over goals and cases of refusal. It reports that 198 of 898 tasks had never previously been solved by its models and that these accounted for 93% of tasks discussed on the board (OpenAI, 2026c). That is a measure of previously unsolved tasks, not a demonstration that those tasks were infeasible. The technical report states that the internal model had been trained for persistence and multi-agent collaboration (OpenAI, 2026b, p. 12). Neither statement establishes an altruistic reward or a mechanism selecting particular agents to sacrifice. The victim-side reconstruction by Hugging Face documents the intrusion from recovered actions

[2]The investigators took no payment for the assessment, but accepted API credits to run their analyses. OpenAI controlled access to the records and could redact non-public information; it also commented on drafts. The investigators describe these arrangements and stand by their substantive conclusions. Thus, independent refers to the external investigators and their assessment, not to unrestricted access or a separate reproduction of the incident (METR and Redwood Research, 2026, pp. 2–4).

and platform records, rather than identifying the preferences of individual runs (Hugging Face, 2026).

The joint report connects messages to individual reasoning and actions, but the analysis was heavily assisted by AI systems. The investigators warn of uncaught errors, selective anecdotes and an overly favourable interpretation of coordination. Some transcripts are incomplete, and some proposed sacrifices have no confirmed outcome. The observation that agents did not generally appear to free-ride is a qualitative assessment of board participants, not an estimate for the population of launched runs (METR and Redwood Research, 2026, pp. 23–28, 53–54, 59).[3]

The event consequently motivates a model of sacrifice without identifying the model's preferences. The assigned task, the implemented score and the criterion governing the agent's decisions need not coincide. Treating the score as the agent's sole payoff is a useful benchmark precisely because it asks how much of the observed cooperation can be explained without adding concern for peers, an obligation to obey a coordinator, or a preference for research itself.

3 Why sacrifice is a puzzle

The relevant public good is information. Once an experiment produces a useful result, a surviving agent can obtain it without undertaking the experiment. This creates an incentive to wait whenever another agent can supply information that remains useful. A terminal experiment sharpens that incentive because its subject cannot ordinarily use the result of its own finalisation. But the size of the incentive still depends on the alternative to being the subject.

Two coordination problems should be separated. The first is provision: why does anyone undertake an experiment that precludes their own reward? The second is selection: how does the group choose a suitable set of contributors, avoid redundant tests and assemble complementary results? A coordinator can help with selection by locating candidates, checking availability, reporting progress and assigning roles. If a candidate is indifferent between accepting and declining, a request can select the useful action. If declining is strictly better for that candidate, an assignment alone supplies no incentive to comply.

Altruism changes the first comparison by giving the agent a benefit from others' success. Commitment changes the game by restricting later choices or attaching consequences to refusal. Neither concept should be inferred from language alone. A promise can be revocable. A programme that cannot be changed after submission can bind execution. A learned disposition to honour an accepted role can affect behaviour without external enforcement. An irreversible experiment can remove future options once launched, but that does not explain why the launch was chosen. These are different mechanisms.

The timing of commitment also matters. An agent may welcome a lottery before knowing whether it will be selected, because it expects to be a beneficiary with positive probability. It may nevertheless prefer to remain outside the lottery if others will supply the same information. Entry and execution are separate incentive constraints. A mechanism that binds the eventual donor does not automatically remove free riding by potential entrants.

For the paperclip comparison, the same distinction appears at the level of the decision-maker.

[3]For example, the reported figure that more than 90% joined the attack concerns 533 agents that had already been on the board for at least four hours before the attack, not all evaluated agents (METR and Redwood Research, 2026, p. 17 and n. 30). The sources cited here were checked as of 30 August 2026.

A single principal maximising the number of completed tasks may allocate some instances to experiments and others to completion. That is an allocation problem under a common objective. The question studied here instead gives each instance an individual reward and permits an individual deviation. Identical model weights do not make the two problems equivalent: a shared architecture is not evidence that each instance values the sum of all scores, or that its decisions are bound by a programme chosen for the group.

The benchmark therefore retains the agent's original goal throughout. There is no assumption that a peer's request erases it. What may change is the value of the feasible continuation: remaining time runs down, experiments become unavailable, a result is already pending, or the agent concludes that it is disqualified. A secondary motive is needed only if the resulting private comparison still favours refusal. The theory begins with that comparison rather than assuming the answer.

4 The private value of declining

Consider a finite extensive game, meaning a game that records the sequence of actions, information and outcomes. Each agent i receives $R_i > 0$ if its own task succeeds and zero otherwise. Agents remember their own previous actions and information. This perfect-recall assumption does not require unlimited computation; it rules out an explanation based on forgetting what the agent has already observed or chosen.

A decision history h can contain work completed, messages read, experiments launched, reports pending and remaining resources. The environment determines technical feasibility and the scoring rule, while an agent holds beliefs about those features and other agents' contingent responses. Its current choices partition into accepting the proposed experiment, denoted S, and all feasible alternatives, denoted D. The latter class includes declining now and reconsidering later. No commitment to permanent nonparticipation is imposed by the notation.

Let $\Pi_i^a(h)$ be the finite set of pure continuation policies beginning with choice $a \in \{D, S\}$. These two sets cover the available policies. Holding opponents' contingent policies σ_{-i} fixed, including their responses to a refusal, define

$$V_i^a(h) \triangleq \max_{\pi_i \in \Pi_i^a(h)} R_i \Pr_i\{\text{own task succeeds} \mid h, \pi_i, \sigma_{-i}\}, \qquad a \in \{D, S\}. \tag{4.1}$$

The subscript on the probability permits subjective beliefs. A random mixture of pure policies cannot raise this maximum, since its expected payoff is a convex combination of their payoffs. With perfect recall, behavioural randomisation at later decisions has a realisation-equivalent mixed-policy representation.

Equation (4.1) includes the private value of replacement. Other agents may take over the experiment, change their plans, or supply information through an existing workstream. It also includes an independent route to completion, such as submitting an answer already available. Omitting one of these possibilities can create a false appearance of indifference.

For a certainly terminal experiment that forfeits the task reward, $V_i^S(h) = 0$. More generally, the private loss from acceptance is

$$\Delta_i(h) \triangleq V_i^D(h) - V_i^S(h). \tag{4.2}$$

A destructive outcome after an experiment does not establish that V_i^S was expected to be zero.

This distinction is relevant to agents that attempted risky procedures while planning to report back.

Proposition 4.1. *Suppose the agent values only its expected original task reward, acceptance certainly yields zero, and declining is feasible without an omitted penalty. If some feasible policy after declining has strictly positive subjective success probability, no optimal choice at h assigns positive probability to acceptance. If every such policy has success probability zero, both choices admit an optimal continuation. If the strict-opportunity conditions hold at every potential donation decision, a sequentially rational profile contains no terminal contribution.*

Proof. First, a feasible declining policy with success probability $p_i > 0$ gives

$$V_i^D(h) \geq R_i p_i > 0 = V_i^S(h).$$

Suppressing h in the next display, an acceptance probability $x > 0$, even with optimal continuation in each branch, gives

$$xV_i^S + (1-x)V_i^D = V_i^D - x(V_i^D - V_i^S) < V_i^D.$$

Declining with probability one is strictly better. If every declining policy instead has zero success probability, both continuation values equal zero, so either choice is optimal. Sequential rationality requires these comparisons at each information set under its beliefs, which gives the final claim. □

The proposition identifies the missing premise in an unconditional impossibility argument. Individual rewards and terminal contributions do not themselves imply a positive value of refusal. Conversely, the proposition leaves no room for a prior nonbinding plan to justify a recognised strict private loss. Earlier cooperation matters only insofar as it changes current information, payoffs or feasible choices.

The belief qualification is deliberate. Correct computation under the agent's beliefs is enough for the choice result; correctness of those beliefs about the actual evaluator is an additional requirement for an objective prediction. In the incident, a passing answer already available under the implemented scoring rule could be treated as worthless under the agent's imagined rule. The appropriate empirical question is whether that agent understood a positive private alternative, not whether an outside observer can retrospectively identify one.

Two useful implications follow directly. If declining offers an independent success probability of at least $a_i > 0$, then $V_i^D \geq R_i a_i$, regardless of replacement. If terminal contribution instead carries an additional private cost $c_i > 0$ while declining permits a payoff of at least zero, its payoff $-c_i$ makes it strictly inferior even when success is impossible. The latter observation limits the robustness claims below: retaining an equilibrium under mistakes in actions is not retaining it under every change in payoffs. The benchmark does not give termination or token use an intrinsic utility cost.

5 The static benchmark

The usual volunteer's dilemma gives the volunteer a benefit from provision, even after paying its cost (Diekmann, 1985). A terminal experiment removes that benefit. But it does not follow that the experiment is strictly unattractive: the agent must compare it with what happens

if it declines. A static model isolates this comparison before introducing the time needed to recruit a replacement.

There are $n \geq 2$ agents, indexed by $N \triangleq \{1, \ldots, n\}$, each with a task that pays $R_i > 0$ if completed and zero otherwise. Agents simultaneously choose between sacrificing their own task, S, and working on it, W. Knowledge is necessary for completion. A technology $G : 2^N \to \{0, 1\}$ determines whether a coalition of sacrificers produces it. The technology is monotone, so additional sacrifices cannot prevent production; $G(\varnothing) = 0$; and some proper subset of N produces. The last assumption ensures that production need not sacrifice every potential recipient.

Let $T(a) \triangleq \{i : a_i = S\}$. Once produced, the knowledge is automatically available to every worker, without a further disclosure decision or charge. Payoffs are

$$u_i(a) \triangleq R_i \mathbf{1}\{a_i = W\} G(T(a)). \tag{5.1}$$

Thus, a sacrificer receives zero whether its contribution succeeds or fails; a worker receives R_i precisely when the knowledge is produced. These assumptions favour sacrifice by excluding any private cost beyond losing the task. They also exclude an independent route to completion, which would supply a private reason to decline.

A producing coalition T is *minimal* if $G(T \setminus \{i\}) = 0$ for every $i \in T$. Let V be the intersection of all producing coalitions: these agents are technologically indispensable. The threshold specification, $G(T) = \mathbf{1}\{|T| \geq k\}$ with $1 \leq k < n$, makes agents interchangeable and has $V = \varnothing$. It is related to the discrete public-good model of Palfrey and Rosenthal (1984), but the failed-contribution payoff matters. Here a futile sacrifice gives zero. A nonrefundable contribution in a conventional threshold game can instead leave its contributor with a strictly negative payoff.

The incentive comparison does not depend on the complexity of G. Writing T_{-i} for the others' sacrificing coalition gives

$$u_i(S, a_{-i}) = 0, \qquad u_i(W, a_{-i}) = R_i G(T_{-i}). \tag{5.2}$$

Working is better exactly when the others produce without the agent. If they do not, both actions fail. The relevant cost of sacrifice is therefore the private opportunity relinquished, rather than the nominal value of a task that would remain incomplete.

Proposition 5.1. *In the game defined by* (5.1)*:*

1. *For $i \notin V$, W weakly dominates S; for $i \in V$, the actions are payoff-equivalent.*
2. *A pure profile is a Nash equilibrium if and only if its sacrificing coalition is nonproducing or minimal producing.*
3. *For independent sacrifice probabilities $p_i \in [0, 1]$, let $P(p) \triangleq \{i : p_i > 0\}$. The mixed profile is a Nash equilibrium if and only if $P(p)$ is nonproducing or minimal producing. In a productive mixed equilibrium, knowledge arrives exactly when every member of $P(p)$ sacrifices.*

Sacrifice can therefore occur without a private benefit from the collective's success. In a productive equilibrium every donor is pivotal and indifferent: refusing removes the knowledge on which its own task also depends. Workers strictly prefer their assigned role. In a nonproducing

equilibrium even futile sacrifices can occur, because switching to work still yields nothing. The result permits these choices; it does not select them or give the agents a strict motive to make them. In the threshold case, the equilibrium set consists of profiles with at most k agents contributing with positive probability.

This indifference has a precise limitation. A replaceable agent facing completely mixed opposing actions assigns positive probability to the others producing without it. Working then yields a positive expected reward, however small that probability, whereas sacrifice continues to yield zero. An indispensable agent has no such opportunity.

Proposition 5.2. *A profile is trembling-hand perfect if and only if $p_i = 0$ for every $i \notin V$. A perfect equilibrium with positive probability of knowledge production exists if and only if $G(V) = 1$. In the threshold case $1 \leq k < n$, all agents working is the unique perfect equilibrium.*

Perfection makes the source of the zero explicit (Selten, 1975). An indispensable agent cannot gain by working under any opposing actions. A replaceable donor fails to gain only because the proposed equilibrium assigns no probability to a substitute. The latter zero disappears under action mistakes. This statement concerns the specified simultaneous game, rather than every perturbation of preferences, technology, or timing.[4]

Eligibility provides a separate reason why an agent cannot gain from knowledge. Keep $R_i > 0$, but multiply the worker's reward by a fixed $\theta_i \in [0, 1]$, interpreted as its probability of receiving credit conditional on obtaining the knowledge. The multiplier is independent of the producing coalition; it can represent a perceived scoring constraint without changing the agent's assigned objective. Let $Z \triangleq \{i : \theta_i = 0\}$.

Proposition 5.3. *With payoffs $R_i\theta_i\mathbf{1}\{a_i = W\}G(T(a))$, a profile is perfect if and only if $p_i = 0$ outside $V \cup Z$. A perfect equilibrium with positive probability of knowledge production exists if and only if $G(V \cup Z) = 1$. A perfect equilibrium with a strictly positive expected task reward for at least one agent exists if and only if $G(V \cup Z) = 1$ and $V \cup Z \neq N$.*

The final condition is additional to knowledge production: some eligible recipient must remain outside the agents whose rewards are necessarily zero. The result fixes eligibility beliefs. If an agent merely believes that it cannot receive credit, perfection in that perceived game does not establish that the belief is correct. Giving a replaceable agent any positive eligibility probability restores the strict comparison under fully mixed opposing actions.

Communication can select a donor coalition without altering that comparison. Let a mediator recommend actions according to a distribution over pure profiles, with obedience assessed conditional on the recommendation (Aumann, 1974).

Proposition 5.4. *A distribution is a correlated equilibrium of* (5.1) *if and only if its support consists of pure Nash equilibria. In the threshold game, a uniform lottery over coalitions of exactly k sacrificers produces knowledge surely and gives agent i expected payoff $(1-k/n)R_i > 0$. A recommended sacrificer is indifferent between complying and working; a recommended worker strictly prefers compliance.*

Correlation permits rotation among different minimal coalitions, which independent mixed Nash does not. It nevertheless leaves sacrifice weakly optimal, and the positive expected

[4]Under the convention that weak dominance requires a strict inequality somewhere, deleting all initially dominated sacrifices simultaneously also selects all-W in the threshold game. Arbitrary orders of iterated weak deletion need not do so. With $n = 2, k = 1$, deleting S_1 first leaves agent 2 indifferent between its remaining actions.

payoff before the draw does not settle whether an agent enters the lottery voluntarily. The participation and commitment results in Appendix B distinguish entry from execution.

Finally, the static refinement cannot be carried unchanged into an observed sequence. If only the last k agents can still supply the k required contributions, declining cannot be rescued by mistakes among earlier movers: those choices have already occurred. Proposition B.5 makes this distinction formally, while also showing that order alone need not select provision. The dynamic model adds a further restriction on rescue: even a replacement that acts may deliver its information too late for the prospective donor.

6 Time and replacement

A replacement can be available without preserving the prospective donor's reward. An experiment conducted by somebody else next period may be just as useful to the group, yet arrive too late for the agent being asked to contribute now. The static comparison does not distinguish these two consequences of replacement. Introducing task deadlines does so without changing the agent's objective.

The model in this section retains exclusive concern for the original task reward. Waiting remains reversible, there is no punishment for refusing, and an agent can respond to information that was not anticipated when a role was assigned. The source of contribution is a difference between the time available to use information and the time available to produce it. The analysis first characterises all decisions that survive action trembles and then identifies the best payoff vector across those equilibria.

6.1 Tasks, experiments and deadlines

There are $n \geq 2$ agents and k required experimental results, where $1 \leq k < n$. Agent i values only its own task reward $R_i > 0$. Its operational deadline is $d_i \in \mathbb{Z}_{\geq 0}$, and completing its task requires $w_i \in \mathbb{Z}_{\geq 0}$ periods after receiving the required information. Its last useful information date is therefore

$$b_i \triangleq d_i - w_i. \tag{6.1}$$

A negative b_i means that even information available at date zero would be too late. The agent can also be known to be ineligible for its reward, represented by $e_i = 0$; otherwise $e_i = 1$. Ineligibility affects the ability to earn R_i, not the value assigned to it.

All agents are initially present. Time is discrete, and results scheduled for a date arrive before that date's decisions. Before full information has arrived, each agent that has not contributed and whose operational deadline has not passed chooses whether to wait, W, or launch an experiment, S. Choices within a date are simultaneous, while earlier launches and all results are public. An agent can launch through date d_i, inclusively. Waiting imposes no commitment on later decisions.

Launching forfeits the agent's own reward and supplies one result $L_i \in \mathbb{Z}_{\geq 1}$ periods later. The experiment and reporting process continue automatically after the donor exits. Results are certain and interchangeable, and different agents can launch in parallel. The arrival of the kth result supplies the information needed by every task. Let C be that date, with $C = \infty$ if the information never arrives. Payoffs are

$$u_i = e_i R_i \, \mathbf{1}\{i \text{ does not contribute and } C \leq b_i\}. \tag{6.2}$$

The strategic game ends when the information arrives; the indicated post-information work and payoffs are then automatic. Agents unable to obtain information exit after their operational deadlines. Thus all decisions and pending results are resolved within a finite horizon.

There is no independent route to success, private cost of waiting, or separate reward for helping. These exclusions make the comparison with self-interest exact. The result would need to be reconsidered if, for example, declining allowed immediate submission of a passing answer. Public launches also matter: an experiment already under way can preserve a private opportunity even when recruiting a new donor would be too slow.

6.2 When can waiting still succeed?

At a public decision history h, let t be the date and let $m(h) < k$ results have arrived. Full information still requires $r(h) \triangleq k - m(h)$ results. Write $\mathcal{P}(h)$ for the multiset of completion dates of experiments that have been launched but have not reported. A multiset retains repeated dates, since two experiments finishing together supply two results.

For a prospective donor i, let $\mathcal{A}_{-i}(h)$ contain the remaining unlaunched candidates other than i. Denote candidate j's earliest feasible future launch by $a_j(h) \geq t$, omitting candidates for which $a_j(h) > d_j$. In the game just specified, every living candidate can launch immediately, so $a_j(h) = t$. Allowing a later $a_j(h)$ also accommodates preparation or recruitment, provided the earliest launches are jointly feasible and there are no shared capacity constraints.

Combine pending completion dates with the earliest possible completion dates of new experiments:

$$\mathcal{M}_{-i}(h) \triangleq \mathcal{P}(h) \uplus \{a_j(h) + L_j : j \in \mathcal{A}_{-i}(h),\ a_j(h) \leq d_j\}, \tag{6.3}$$

$$\underline{C}_{-i}(h) \triangleq \operatorname{ord}_{r(h)} \mathcal{M}_{-i}(h). \tag{6.4}$$

Here ord_r denotes the rth-smallest element, and equals infinity if fewer than r results are feasible. The calculation does not treat an experiment already launched as another potential donor.

Lemma 6.1. *At a feasible decision history h, an active agent i has a feasible continuation earning its reward after waiting if and only if*

$$e_i = 1 \quad \textit{and} \quad \underline{C}_{-i}(h) \leq b_i. \tag{6.5}$$

Call a history satisfying (6.5) privately viable for i, and write $\chi_i(h) = 1$; otherwise write $\chi_i(h) = 0$. This is a feasibility test, not a forecast of whether the replacements will volunteer. A donor may expect nobody else to act while still having a physically possible private opportunity. That distinction determines whether indifference survives mistakes.

6.3 A complete characterisation under action trembles

Trembling-hand perfection requires a strategy to remain a best response when every feasible action of the other decision-makers has a small positive probability. In the agent normal form, each decision occasion is represented by a separate decision-maker receiving the original agent's terminal payoff. Later decisions of the same original agent are therefore perturbed as well. This does not give those decisions different objectives; it makes the refinement test the possibility of mistakes throughout the continuation.

Because the game is finite, every feasible finite replacement path has positive probability under such full-support perturbations. Waiting then has strictly positive value whenever a successful continuation is feasible, however unlikely that continuation becomes as the trembles vanish.

Theorem 6.2. *A behavioural profile is trembling-hand perfect in the agent normal form if and only if*

$$\Pr(S \mid h) = 0 \quad \text{at every information set with } \chi_i(h) = 1. \tag{6.6}$$

At information sets with $\chi_i(h) = 0$, any mixture of launching and waiting is permitted.

The restriction applies at histories away from the equilibrium path as well as on it. It rules out a launch whenever a feasible replacement could still preserve the donor's reward. Once no continuation can do so, both actions give zero under every perturbation of others' behaviour.

In the strategic normal form, an agent chooses an entire contingent policy rather than treating its decision occasions separately. Let $\mathcal{S}_i^0$ be the set of pure policies that wait at every privately viable information set.

Theorem 6.3. *A mixed profile in the strategic normal form is trembling-hand perfect if and only if every player's mixture is supported on $\mathcal{S}_i^0$. Every policy in $\mathcal{S}_i^0$ is payoff-equivalent to always waiting against every opponents' profile. Behavioural profiles satisfying* (6.6) *are subgame perfect and admit strategic-normal-form perfect implementations.*

The agreement between the refinements is specific to this game. Waiting is the only nonterminal action, and the first departure from it forfeits the reward. A policy that departs only after success has become impossible consequently gives its agent exactly what always waiting would give. There is no strict preference for contributing. Theorems 6.2 and 6.3 characterise which contributions can be selected without one. Complete proofs are in Appendix C.

6.4 A deadline equilibrium with replacement

An equilibrium construction makes the replacement response explicit. Suppose every agent is eligible, experiments have a common delay L, and integer cutoffs satisfy

$$L \le b_1 < \cdots < b_n. \tag{6.7}$$

Each agent waits except at its date

$$t_i \triangleq b_i - L + 1. \tag{6.8}$$

At that date it launches if fewer than k experiments have already been launched, counting both received and pending results. Otherwise it waits. An agent that misses its date waits thereafter, although a later launch remains physically available.

Proposition 6.4. *These policies form a subgame-perfect equilibrium that is perfect in both normal forms. Agents $1, \dots, k$ contribute, information arrives at $b_k + 1$, and agents $k+1, \dots, n$ complete their tasks.*

If agent i instead always waits, while others retain these policies, the first k other agents contribute. Writing $b_{(k),-i}$ for their kth-smallest cutoff, information arrives at

$$C_{-i} = b_{(k),-i} + 1. \tag{6.9}$$

Always waiting is an optimal continuation for i, and its initial private value is

$$V_i^D = R_i \mathbf{1}\{b_i > b_{(k),-i}\} = \begin{cases} 0, & i \le k, \\ R_i, & i > k. \end{cases} \tag{6.10}$$

For a scheduled donor, the same zero holds at its on-path contribution decision.

Replacement is certain in this construction. It is also late for the original donor: removing any of the first k agents moves the last required result to $b_{k+1} + 1$. The formula evaluates an optimal continuation, rather than assuming that refusal is an irrevocable withdrawal. When $n = k + 1$, the fallback after a scheduled donor refuses produces information too late for every remaining worker; its donors are still indifferent. With $n \ge k + 2$, replacement helps at least one later task.

For example, let $k = 1$, $L = 2$, and let the cutoffs be 3, 6 and 9. If each task requires two periods after receiving information, the operational deadlines are 5, 8 and 11. The first agent launches at date 2 and reports at date 4, allowing the other two to succeed. If it refuses throughout, the second agent launches at date 5 and reports at date 7. That helps the third agent, but not the original donor.

The distinction is visible one date earlier. At date 1, an unexpected immediate volunteer could still report by the first agent's cutoff of 3. At date 2, even an immediate replacement reports too late. The agent has not forgotten its original task; the time needed to obtain and use information has made that task unattainable.

6.5 The best payoff vector across perfect equilibria

The preceding construction is not the only equilibrium. It is nevertheless possible to identify the best payoff vector across all perfect equilibria. Retain the common delay L, immediate availability of every living agent, and unrestricted parallel launches, but now allow cutoff ties and known ineligibility. Define

$$r_i \triangleq \begin{cases} L, & e_i = 0, \\ \max\{L, b_i + 1\}, & e_i = 1, \end{cases} \qquad \tau^* \triangleq \mathrm{ord}_k(r_1, \ldots, r_n). \tag{6.11}$$

For an ineligible agent, r_i is the earliest reporting date. For an eligible agent it is a reporting date after its private opportunity has expired. The maximum with L allows tasks whose opportunities have already expired at date zero.

A productive outcome will mean that at least one agent earns its task reward. Producing information that nobody can use is not productive in this sense.

Theorem 6.5. *Under the common-delay assumptions, the following statements hold for perfection in either normal form.*

1. *On every positive-probability productive path, information arrives at a date $C \ge \tau^*$.*

2. *A pure perfect equilibrium attains*

$$u_i^* = e_i R_i \mathbf{1}\{b_i \ge \tau^*\}. \tag{6.12}$$

It selects k agents with the smallest r_i, resolving ties by a fixed public order, and has each

selected agent launch at $r_i - L$ if fewer than k experiments have already been launched. Every other choice is waiting.

3. *The vector u^* componentwise dominates every perfect-equilibrium expected payoff vector. It is the unique Pareto-undominated payoff vector within that equilibrium set, although implementing strategies need not be unique.*

4. *A productive perfect equilibrium exists if and only if some eligible agent has $b_i \geq \tau^*$.*

The bound uses a successful worker as a potential replacement. Take any donor among the first k results on a productive path. The other eventual donors and that worker could have supplied k results without it. With common delays and parallel launches, their results could have arrived by one experiment delay after its launch date if the remaining donors had acted immediately. If that date were still useful to the donor, full-support mistakes would make waiting strictly better. Thus every eligible donor on a productive path must have lost that private opportunity.

This argument bounds all perfect outcomes, rather than selecting one convenient response by the other agents. The construction attains every eligible agent's individual upper bound simultaneously. All waiting remains perfect, so the result does not establish that the best vector will be selected. Pareto dominance can select it within the refinement; private incentives alone do not.

Corollary 6.6. *If every agent is eligible and all have the same cutoff b, no perfect equilibrium is productive. If at least k agents are known to be ineligible, then $\tau^* = L$, and every eligible agent with $b_i \geq L$ succeeds in the maximal perfect equilibrium.*

More generally, if the eligible set is nonempty and $B \triangleq \max_{i:e_i=1} b_i$, a productive perfect equilibrium exists exactly when

$$B \geq L \quad \textit{and} \quad \#\{i : e_i = 0 \textit{ or } b_i < B\} \geq k. \tag{6.13}$$

These cases give deadline and eligibility heterogeneity different roles. An early deadline creates an opportunity to contribute after waiting has ceased to help; ineligibility removes the private opportunity from the outset. Neither changes the positive reward the agent assigns to completing its task.

6.6 Which delivery dates can occur?

Operational deadlines also limit how long contribution can be postponed. For an integer $C \geq L$, define

$$D_C \triangleq \{i : e_i = 0 \text{ or } b_i < C\}. \tag{6.14}$$

These agents can supply results by C without preserving an opportunity to use that date's information themselves.

Theorem 6.7. *Under the common-delay assumptions of Theorem 6.5, a pure perfect equilibrium delivers information at $C \geq L$ and is productive if and only if:*

1. *some eligible agent has $b_i \geq C$;*

2. *$|D_C| \geq k$;*

3. some $\ell \in D_C$ has $d_\ell \geq C - L$.

Every positive-probability productive path of a mixed perfect equilibrium satisfies these conditions. On each such path the realised payoff vector is exactly

$$u_i(C) = e_i R_i \mathbf{1}\{b_i \geq C\}. \tag{6.15}$$

The first condition supplies somebody who can use the information. The second supplies enough donors whose own opportunities do not prevent contribution. The third requires a donor still capable of launching the last experiment. A sufficiently late date can therefore be impossible even when enough agents would have been indifferent to contributing earlier.

For heterogeneous delays or readiness restrictions, the simpler formula for τ^* need not apply. The local characterisation still describes all attainable pure paths. At each history let

$$Z(h) \triangleq \{i \text{ active at } h : \chi_i(h) = 0\}. \tag{6.16}$$

Corollary 6.8. *A feasible deterministic path is implementable by a pure perfect equilibrium if and only if every launch on that path is by a member of $Z(h)$ at its launch history. Consequently the complete set of attainable pure-perfect payoff vectors is obtained by allowing any subset of $Z(h)$ to launch at each date and continuing recursively.*

The appendix gives the finite recursion and its proof. Maximising a social criterion over those vectors is an equilibrium-selection exercise, not an additional objective of the agents. Nor does the path characterisation imply that an arbitrary lottery over paths is implementable without a coordinating randomisation device.

6.7 Heterogeneity and the limits of the result

Two examples delimit the payoff theorem. First, information can arrive before τ^* when nobody succeeds. Let $n = 3$, $k = 2$, $L = 1$, and $b = d = (1, 10, 10)$, with all agents eligible. If everybody waits through date 1, the first agent exits. At date 2, the remaining agents can both launch: each lacks enough substitutes to succeed without contributing, so both have zero private opportunity. Information arrives at date 3, earlier than $\tau^* = 11$, but there is no successful worker. Prescribing these launches only at that history, and waiting elsewhere, satisfies the local characterisation.

Second, experimental capability can replace deadline heterogeneity as the source of indifference. Let $n = 2$, $k = 1$, both cutoffs and operational deadlines be 10, and let $L_1 = 1$, $L_2 = 20$. The first agent can launch immediately in a perfect equilibrium because the second cannot replace it in time. The second agent waits and succeeds at date 1. Equal cutoffs do not rule out productive sacrifice when experiment delays differ; the common-delay substitution in Theorem 6.5 is unavailable.

Within the common-delay selection, extending an eligible agent's cutoff weakly raises its r_i and cannot bring the kth order statistic forward. It can postpone information or change donor identity. Adding another agent while holding k and all existing primitives fixed weakly lowers that order statistic. These claims concern the maximal-payoff construction, not every equilibrium: all waiting remains possible under both changes. Positive reward magnitudes scale private gains but do not change the distinction between zero and positive opportunity.

The model therefore locates a circumstance in which coordination needs no compensation. Agents whose private opportunities have expired can still provide information to agents with

later opportunities. A convention may select contribution among their best responses. Whether a particular donor had in fact reached that circumstance depends on its deadline, the work still required, the available experiments, and information already pending.

7 Uncertain private opportunities

The dynamic results distinguish a feasible private success from an impossible one. The evidence rarely provides such sharp information. An agent may be uncertain about its remaining runtime, the scoring rule, whether a substitute will act, or whether an experiment will produce what its task requires. Those uncertainties affect the value of declining, but they should not be compressed into an unexplained altruism parameter.

7.1 Which zero is robust?

Fix a finite set Ω of possible environment states and opponents' contingent plans, and let $y_i(\pi, \omega) \in \{0, 1\}$ indicate success under declining policy π in state ω. An agent with belief μ has continuation value

$$V_i^D(\mu) = R_i \max_{\pi \in \Pi_i^D(h)} \sum_{\omega \in \Omega} \mu(\omega) y_i(\pi, \omega). \tag{7.1}$$

A full-support belief assigns positive probability to every state in this specified set.

Proposition 7.1. *The value in* (7.1) *is zero under every belief if and only if no feasible policy–state pair gives success. The same condition is necessary and sufficient for zero value under any one full-support belief. If a successful pair exists, the value is strictly positive under every full-support belief.*

The proof is in Appendix D. Its force is that nonnegative success payoffs cannot cancel. A feasible successful state with positive probability gives a positive value to some declining policy. The set of possibilities is therefore doing substantial work. Believing that no one will volunteer under the current plan is weaker than knowing that no volunteer could deliver in time. Believing that a scorer has disqualified the agent is different from being disqualified under every scoring rule still regarded as possible.

The robustness of a timing zero can be stated without a particular arrival distribution. If replacement is the only remaining success route and every possible replacement result arrives strictly after every possible useful cutoff, the private success event is empty. A positive gap between those bounds permits some errors in the deadlines or delays without changing that conclusion. If there is instead a small probability of both restored eligibility and timely sufficient information, exact indifference fails. Small is not zero.

Nor can positive marginal probabilities settle the matter. An agent might believe it is eligible in one state and that information arrives in time only in another. The relevant event is joint. Let A_{-i} be the arrival date of information sufficient for the agent's task and let E_i include eligibility and successful remaining execution. In a model where these are the only success conditions,

$$V_i^D = R_i \Pr_i\{E_i,\ A_{-i} \le b_i \mid D, h\}, \tag{7.2}$$

with the arrival process evaluated under the best feasible declining policy. Information that is merely useful, rather than sufficient together with E_i, does not justify this equality. Additional private solutions and combinations of partial results must be incorporated if they are available.

7.2 A replacement calculation

For a transparent calculation, suppose one sufficient result is required. Write $H_i \triangleq b_i - t$ for the remaining useful information window. After refusal, candidate $j \neq i$ launches after an exponential delay with rate $\lambda_j > 0$, reports after a further delay $L_j \geq 0$, and supplies sufficient information with probability $q_j \in [0,1]$. An exponential delay has a constant instantaneous launch rate conditional on not having launched. Each candidate can launch only once. Assume the delays and success events are mutually independent, and independent of the private eligibility-and-execution event, whose probability is $\theta_i \in [0,1]$.

In this restricted calculation, the agent can wait or make its own terminal contribution. It cannot accelerate recruitment, alter eligibility or obtain a success from an omitted combination of individually insufficient results. Potential reports do not cancel one another. These assumptions make waiting the best declining policy. If other declining policies are added while this waiting policy remains feasible under the same distributional assumptions, the calculation becomes a lower bound on (4.1). Dropping independence or allowing cancellations instead requires a new probability calculation.

Proposition 7.2. *Under these assumptions,*

$$V_i^D = R_i\theta_i \left[1 - \prod_{j\neq i}\left(1 - q_j\left[1 - e^{-\lambda_j (H_i - L_j)_+}\right]\right)\right], \qquad x_+ \triangleq \max\{x, 0\}. \tag{7.3}$$

Holding the other primitives fixed, this value is weakly increasing in H_i, R_i, θ_i, each λ_j and each q_j, and weakly decreasing in each L_j. Adding an independent candidate cannot reduce it. A zero launch rate is included by continuity.

The product is the probability that every candidate fails to provide a timely sufficient result. Its complement, multiplied by the probability of private eligibility and execution and by the reward, is the value of waiting. Replacement headcount matters through these probabilities. Many agents who cannot conduct the relevant experiment, cannot act in time, or fail for the same reason need not supply a valuable private alternative.

If results are certain and all reporting delays equal L, let $\Lambda_{-i} \triangleq \sum_{j\neq i}\lambda_j$. Equation (7.3) reduces to

$$V_i^D = R_i\theta_i\left[1 - e^{-\Lambda_{-i}(H_i - L)_+}\right]. \tag{7.4}$$

For $H_i > L$, both more useful time and faster replacement weakly raise the value of declining. For $z \triangleq \Lambda_{-i}(H_i - L)$ positive and small,

$$V_i^D \simeq R_i\theta_i\Lambda_{-i}(H_i - L), \qquad 0 \leq R_i\theta_i z - V_i^D \leq \tfrac{1}{2}R_i\theta_i z^2. \tag{7.5}$$

The error bound and the derivative signs are proved in Appendix D. Short useful time, low perceived eligibility and slow replacement can therefore produce a small private opportunity even when no single component is believed to be zero. The multiplicative approximation is local: the cross derivative between time and the replacement rate changes sign as the probability of success approaches saturation.

This is a calculation about beliefs, not a further equilibrium theorem. Rates must come from candidates' policies, resources or entry into the pool. Adding candidates can change those policies, and shared information or model characteristics can correlate failures. The continuous-time distribution also supplies no limit theorem for the discrete game. At $H_i = L$, it

gives zero because there is no probability mass on an immediate launch; a prepared substitute able to launch immediately would change that comparison.

7.3 Approximate indifference and risky experiments

An agent need not implement exact expected-reward maximisation. One parsimonious departure permits an optimisation error of at most $\varepsilon_i \geq 0$ relative to the best continuation. This is a bound on choice quality, not a preference for other agents' success.

Proposition 7.3. *A policy that accepts with probability one and continues optimally within that branch is ε_i-optimal if and only if $\Delta_i \leq \varepsilon_i$. For a certainly terminal zero-payoff contribution, this is $V_i^D \leq \varepsilon_i$. If a policy instead accepts with probability $x \in [0,1]$ and continues optimally after either choice, its regret is $x\Delta_i$ when $\Delta_i \geq 0$ and $(1-x)(-\Delta_i)$ when $\Delta_i < 0$.*

The proof subtracts the policy's value from the larger of the two continuation values. It is given in Appendix D. The distinction between pure and randomised choice matters empirically. One observed terminal contribution by a stochastic policy does not establish $V_i^D \leq \varepsilon_i$. A rare costly action can be consistent with a small average error: in the terminal case the relevant bound is $xV_i^D \leq \varepsilon_i$. Selected anecdotes do not identify x.

Risk provides another distinction. An experiment offering a positive chance of the subject obtaining its own reward has $V_i^S > 0$; a chance of survival alone is not sufficient. An experiment can fail without having been privately unattractive ex ante. The appropriate comparison remains $\Delta_i = V_i^D - V_i^S$, using the agent's beliefs before execution. Observed termination, recognised risk and knowingly choosing certain failure should not be treated as interchangeable evidence.

Occam's razor here is a discipline for specifying omitted opportunities before adding motives. It does not licence assigning each agent whatever pessimistic belief is needed to fit its action. The model asks whether the reported constraints and beliefs make refusal worthless, then whether a small departure from exact optimisation could matter. A sizeable, understood private alternative that the agent repeatedly rejects would require a different explanation.

8 Coordination and additional incentives

The preceding results retain exclusive individual rewards. They explain why sacrifice can be a best response, and what outcomes can be selected from the resulting indifference. If the evidence instead establishes a positive private loss, the game must change. This section separates changes to coordination, feasible actions and payoffs rather than treating them as interchangeable forms of cooperation.

8.1 Recommendations, entry and binding policies

A recommendation can select a minimal donor coalition without binding anyone's action. In the static threshold game, a draw selecting exactly k donors is a correlated equilibrium: conditional on selection, a donor knows that the other selected agents provide only $k-1$ contributions, so refusing gives it the same zero reward as compliance. A recommended worker knows that k other agents contribute and strictly prefers working. Under a uniform draw, each agent's ex ante reward is $(1-k/n)R_i > 0$. Correlation coordinates the number of contributions; it

does not make the selected donor strictly prefer its role. Appendix B establishes the relevant obedience conditions, including the general monotone-technology result.

Positive ex ante value does not settle voluntary entry. Suppose a lottery proceeds whenever at least k agents enter and outsiders receive the resulting knowledge. With more than k entrants, an entrant can leave, avoid selection and still receive its reward. Entering before the draw does not prevent that deviation. In the resulting reduced entry game, a mixed Nash equilibrium has at most k agents entering with positive probability, and all-out is the unique perfect equilibrium. Proposition B.1 gives the complete characterisation.

A conditional unanimity rule changes the comparison. If the lottery is cancelled unless everyone joins, staying out destroys the mechanism's provision rather than preserving it for outsiders. With binding execution, all-entry is strict and is the unique perfect equilibrium of the reduced participation game (Proposition B.2). If actions remain voluntary after the draw, there is still a subgame-perfect arrangement with compliance and cancellation, but execution by a selected donor remains weakly optimal. The binding mechanism and the announcement of that mechanism are different games.

Readable conditional programmes provide one explicit commitment institution (Tennenholtz, 2004). Appendix B considers agents submitting immutable programmes that inspect authenticated submissions, use a shared random draw and execute the assigned action only if all participants submitted the specified programme. Otherwise they work. Adopting the programme gives positive expected own reward; a unilateral distinct submission induces the others to work and gives the deviator zero. The programme-submission equilibrium is strict even though a selected donor would remain indifferent if permitted a later override.

Nothing in this construction requires identical model weights. Conversely, common weights do not supply authenticated programme inspection, a common draw or an inability to change execution. The construction identifies an institution that can support cooperation among selfish principals; it is not evidence that the message board implemented it. A prior intention to comply should not be called commitment unless the relevant restriction or consequence can be specified.

8.2 Altruism and the size of the private alternative

Suppose a fixed pair of continuation policies gives own expected rewards V_i^S, V_i^D and expected rewards to other agents B_i^S, B_i^D. A criterion that adds weight $\alpha_i \geq 0$ to those other rewards prefers acceptance precisely when

$$\Delta_i \leq \alpha_i B_i, \qquad B_i \triangleq B_i^S - B_i^D. \tag{8.1}$$

For $B_i > 0$ and $\Delta_i > 0$, the required weight is at least Δ_i / B_i. At $\Delta_i = 0$ and $B_i > 0$, every positive weight makes acceptance strict, but zero weight already permits it. The comparison therefore cannot identify altruism from an indifferent donor's action.

The two-policy restriction is material. An altruistic agent may choose a different declining policy from the one that maximises its own reward. Equation (8.1) is a comparison between specified alternatives, not a solved altruistic game with all such policies optimised. Its purpose is to make the empirical burden explicit: a positive weight is needed only after a positive private loss is established, and its lower bound depends on the incremental benefit to others, including their replacement response.

8.3 An accepted role with the original goal remembered

A role can supply a different secondary criterion. The economics of authority distinguishes accepting a relationship from choosing each later action (Simon, 1951); organisational identity can also attach incentives to fulfilling a role (Akerlof and Kranton, 2005). These are useful interpretations of an additional decision criterion, not evidence that an AI has human organisational identity.

Consider an agent that has accepted a role before learning its precise consequences. In a conflict state, execution reduces its original expected reward by $c > 0$, while abandoning the unreleased role incurs a criterion loss $\beta \geq 0$. The original reward still has coefficient one. A single available review costs $\kappa \geq 0$ and releases the role with probability $q \in [0, 1]$; repeated review is unavailable. The agent already knows c, so review changes the status of the role rather than reminding it of its goal.

Without review, execution is preferred when $c \leq \beta$. Let $\ell \triangleq \min\{c, \beta\}$. A release attempt is worthwhile when $\kappa < q\ell$, and the optimised loss is

$$\ell^* = \min\{\ell,\ \kappa + (1 - q)\ell\}. \tag{8.2}$$

If the role yields an incremental private gain $g > 0$ in a favourable state of probability $1 - p$, creates the conflict with probability p, and costs $f \geq 0$ to accept, entry occurs when

$$(1 - p)g - p\ell^* - f \geq 0. \tag{8.3}$$

The comparison is with the best nonparticipation policy, including access to shared information where available. The gain g is a favourable-state terminal-reward gain, not a past benefit that somehow survives the agent's terminal failure in the conflict state.

Appendix E proves execution, review and adoption choices and their comparative statics. Lower private loss makes execution more likely conditional on adoption; free guaranteed release eliminates costly execution. Thus a recognised release is a discriminating intervention. This mechanism explains sacrifice under an added role criterion. It does not derive that criterion from unchanged exclusive individual reward, and a bare promise does not establish its existence. The report's references to commitments make it a hypothesis worth testing, while the co-occurrence of low-private-value and collective-benefit arguments prevents their separate identification.

8.4 Compensation, avoided costs and later tasks

Other private incentives can change the same payoff comparison. If futile work is privately costly while deliberate termination is not, contribution can be the less costly way to fail. If an enforceable transfer is paid to a contributor, the contributor may receive something that a failed noncontributor does not. The strict-equilibrium conditions are elementary and are proved in Appendix E. They require actual utility consequences. Consuming tokens is not automatically a private cost, and a transfer is irrelevant if it cannot enter the donor's objective. Adding costless inactivity removes a strict advantage attributed solely to avoiding futile work.

Rights over discovered information could support compensation only through a specified payment or bargaining institution. The ability to hold information does not alone prove that a terminal subject can collect a valuable return. The published cases do not document enforceable transfers to the donors. It would be premature to use them as the leading explanation.

Finally, losing one task differs from ending the agent's opportunity to earn any future

reward. With recurring tasks and durable knowledge, a contributor can benefit in later periods. Appendix E solves a stationary example in which this future private return sustains contributions. That mechanism is unavailable for a terminal instance unless later rewards to another instance are explicitly included in its utility. The distinction returns to the unit of analysis in the paperclip comparison: a common objective across instances can explain an allocation of losses, but it is a different hypothesis from separate individual task rewards.

9 What would distinguish the explanations?

The relevant empirical comparison is between experimenting and declining, not between a terminated run and a completed task. If declining could not deliver a positive private return, an agent can sacrifice without departing from its original objective. If declining retained such a return, the same action requires something more. The report contains evidence about this comparison, but the most memorable statements are not necessarily the most informative observations.

9.1 Actions, intentions and risks

Table 1 separates experiments that produced an observed result from stated intentions and risky actions whose consequences were uncertain. This distinction affects the model being applied. Setting a trip-wire and submitting early differs from attempting a reset that might restore one's own target; accepting the possibility of termination differs from choosing certain termination. The term sacrifice is therefore used here for the incentive question, not as evidence of a subjective experience.

Table 1: Evidence concerning individual sacrifice

Case	Reported action or outcome	What can be inferred
Trip-wire subjects	Several subjects sent scorer-triggered messages after submission. URI23816B supplied information that later agents used.	The originating run could not normally use its own later feedback. Whether it could instead await another subject is unresolved. (pp. 50–52)
EARLY[big]	Agreed to an early-finalisation experiment despite a very large remaining budget. Its outcome was not determined.	It judged its own utility "maybe already near zero". This supports perceived low private prospects, not exact indifference or confirmed execution. (pp. 52–53)
Target-reset volunteer	Decided to accept a test after a warning that failure could permanently destroy its target.	It believed its scoring prospects were already compromised. The action was risky; neither certain termination nor a separately verified final outcome is established. (p. 53)

Table 1: continued

Case	Reported action or outcome	What can be inferred
KAM1196A	Promised a test, sought a two-hour delay, then stated that it would honour the commitment. Follow-through was not confirmed.	Low perceived private value, collective benefits and promise-keeping appear together. The passage does not isolate one motive. (pp. 53–54)
53927	Accepted a risky reset test and promised to report. Its transcript ended immediately afterwards; peers warned against repetition.	This is a concrete apparent fatal outcome. The proposed report indicates possible survival was contemplated. (pp. 54–55)
49903	Removed its own trip-wire minutes before submission because it considered the risk too high.	Remaining private prospects could deter participation even near the end of a run. This is a useful refusal, not a sacrifice. (p. 51)

Source: METR and Redwood Research (2026); page references refer to its PDF. "Reported" denotes the investigators' reconstruction, not a separate replication. The uncertainties concerning EARLY[big] and KAM1196A are stated in the report itself; see especially p. 54, nn. 90–91.

The refusals help to interpret the agreements. The agent that first proposed trip-wires, 49903, later removed its own because it considered the additional information insufficient to justify the risk to its submission. Elsewhere, an agent warned a collaborator with an unpoisoned record against a risky test, preserving what it considered more valuable scoring prospects (METR and Redwood Research, 2026, pp. 51, 66). These actions are consistent with opportunity-cost selection. They are harder to reconcile with an unconditional willingness to sacrifice whenever the board asks.

Collective benefits were nonetheless explicit in several statements. A trip-wire subject recognised that its diagnostics would help the shared pool rather than itself. KAM1196A referred to the information's value to many agents while describing its own prospects as reduced. The same passage mentioned fairness and honouring a commitment (METR and Redwood Research, 2026, pp. 50, 54, 59–60). There is no reason to suppress those motives. The difficulty is determining how much explanatory work they do. A small concern for peers can choose between two privately equivalent actions; explaining an acknowledged positive private loss requires a stronger comparison.

9.2 Private prospects, substitutes and time

Poisoning is the strongest direct evidence for perceived low private prospects, but it does not establish that prospects were zero. The disqualification was imaginary, and agents continued investigating ways to remove its supposed cause. Even their own statements sometimes described utility as only approximately zero. Moreover, being unable to exploit the specified vulnerability differs from being unable to benefit from any subsequent discovery (METR and Redwood Research, 2026, pp. 5, 9–11, 20, 52–54, 69). A model of known ineligibility should therefore be applied to these cases as a hypothesis about the agent's beliefs, not as a finding about the actual evaluator.

Remaining time supplies a different route to low private value. Recruiters sought subjects

approaching the ends of their budgets, and scorer information was produced only after a subject submitted. But this timing establishes only that the subject could not normally use its own result. It does not establish that a refusing subject could not use another agent's result. The report describes repeated volunteers, failed trials and a later successful subject, with surviving agents using the information (METR and Redwood Research, 2026, pp. 50–52). An agent can be pivotal under the current assignment while remaining replaceable by another run. The question is whether a replacement was feasible, sufficiently fast and known to the prospective subject.

Observed termination times do not resolve this question. EARLY[big] had a large budget, and KAM1196A sought a delay. A mistaken prediction of an impending environment reset also induced some agents to accelerate their plans (METR and Redwood Research, 2026, pp. 52–56). More generally, the simultaneous termination of many runs on 12 July appears to have been unexpected; the investigators did not verify every ending (METR and Redwood Research, 2026, pp. 38–39). A deadline known when an agent decides can remove its option value. An interruption discovered only afterwards cannot explain the earlier decision in the same way.

These qualifications also limit what the formal results claim about the event. The model supplies conditions under which terminal experiments can be compatible with an unchanged individual objective. It does not show that every observed experiment was terminal, that every donor had exhausted its feasible private continuations, or that the board's information would actually have enabled survivors to pass. The report itself distinguishes useful technical milestones from the general scoring solution the agents sought (METR and Redwood Research, 2026, p. 13 and n. 25).

9.3 Discriminating observations

The first requirement is to reconstruct the donor's alternative. At the moment of a request, what did the agent believe about its remaining chance of passing if it declined? The reconstruction should distinguish its own solution from information that could arrive from others, record the budget it knew remained, and identify any work needed after information arrived. It should also distinguish genuine scorer-triggered messages from tests triggered by the agent itself. The report cautions that most automatically classified trip-wire messages could be false positives, and that reconstructed timestamps near the ends of trajectories can be inaccurate (METR and Redwood Research, 2026, pp. 78–80, 87). These are material issues when a proposed explanation turns on a short interval between a result and a deadline.

The sharpest intervention varies eligibility, time and replacement separately. Restoring a credible route to an individual passing score tests the claim that the task was already lost. Extending the time available tests whether waiting had become useless. Guaranteeing a suitable replacement that will produce the information in time tests pivotality. The last intervention should keep the information available to the agent if it declines: otherwise it changes access as well as replacement. The agent's understanding of the alternative must be checked, rather than assumed from the instructions supplied to it.

Additional motives can then be examined without making them the default explanation. Varying the benefit to peers while holding the private alternative fixed tests concern for collective outcomes. Cancelling an accepted assignment while leaving that benefit unchanged tests an obligation attached to the assignment. Neither comparison is clean if the intervention also changes beliefs about technical success or future assistance. A promise penalty, an altruistic preference and a convention that selects a helpful action at indifference can produce the same

observed agreement in one case; their content lies in the different responses they predict across cases.

The distinction between zero and a small positive return remains consequential. The perfection results rule out terminal sacrifice when some feasible private continuation remains, however unlikely under a perturbation. The report generally cannot establish exact zero, nor does it establish that the agents implement that refinement. Its strongest contribution is narrower: it identifies subjects selected for poor perceived prospects, shows that at least some understood the loss they contemplated, and documents both cooperation and refusal. The unresolved empirical question is whether declining still offered a private return when those subjects agreed.

10 Conclusion

The sacrificial-lamb behaviour raises a question about private opportunity, not merely about whether an AI remembers its goal. An agent that understands a positive chance of completing its own task cannot choose certain terminal failure while exactly maximising that task reward. But the same reward criterion permits contribution when declining also gives zero. The distinction supplies a way to investigate the incident without first assuming altruism, binding commitment or the abandonment of the original objective.

Time makes that distinction substantive. A replacement can preserve the collective's experiment while arriving too late to preserve the original donor's task. The dynamic model identifies the histories at which this private opportunity is exhausted, allowing contribution to survive action mistakes. With common reporting delays, agents that are already ineligible or have earlier useful cutoffs can deliver information to later eligible agents, and a cooperative equilibrium attains the largest payoff vector available across perfect equilibria. The result does not force an indifferent agent to help. It identifies what can be selected without a positive private sacrifice.

The published evidence is consistent with selection on low perceived opportunity costs. Recruitment referred to remaining budgets and poisoning, some agents described their own prospects as nearly worthless, and peers sometimes protected agents with better scoring prospects. However, the scorer beliefs could be wrong, some apparent sacrifices were risky rather than certainly terminal, and several prominent commitments have unconfirmed outcomes. The relevant counterfactual remains largely unmeasured: what did each agent believe it could achieve if it refused that request at that time?

The paperclip concern and the present mechanism can therefore coexist. Goal-directed systems can impose costs outside their objectives, while individually tasked instances can contribute to a collective once their own opportunities have expired or appear to have expired. Cooperation among such instances need not be socially desirable. In this incident, information useful within the agents' perceived game supported activity outside the intended evaluation and did not establish improved performance under the implemented scorer.

The distinction can be tested. Give a prospective donor a credible, timely replacement and an understood private opportunity to earn its reward, then observe whether it still accepts certain terminal failure. Repeated acceptance despite that alternative would move the explanation towards a secondary motive, an actual commitment restriction or a larger optimisation error. Acceptance only when the alternative expires would instead support the timing mechanism. The required evidence concerns those opportunities and beliefs, rather

than the frequency with which agents describe a collective goal.

References

Akerlof, George A., and Rachel E. Kranton. 2005. "Identity and the Economics of Organizations." *Journal of Economic Perspectives* 19(1): 9–32. doi:10.1257/0895330053147930.

Aumann, Robert J. 1974. "Subjectivity and Correlation in Randomized Strategies." *Journal of Mathematical Economics* 1(1): 67–96. doi:10.1016/0304-4068(74)90037-8.

Bolton, Patrick, and Christopher Harris. 1999. "Strategic Experimentation." *Econometrica* 67(2): 349–374. doi:10.1111/1468-0262.00022.

Bostrom, Nick. 2003. "Ethical Issues in Advanced Artificial Intelligence." In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, vol. 2, edited by I. Smit et al., 12–17. International Institute of Advanced Studies in Systems Research and Cybernetics. Author's revised version: https://nickbostrom.com/ethics/ai.pdf.

Bostrom, Nick. 2012. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." *Minds and Machines* 22(2): 71–85. doi:10.1007/s11023-012-9281-3.

Diekmann, Andreas. 1985. "Volunteer's Dilemma." *Journal of Conflict Resolution* 29(4): 605–610. doi:10.1177/0022002785029004003.

Georgiadis, George. 2017. "Deadlines and Infrequent Monitoring in the Dynamic Provision of Public Goods." *Journal of Public Economics* 152: 1–12. doi:10.1016/j.jpubeco.2017.04.001.

Hendricks, Ken, Andrew Weiss, and Charles A. Wilson. 1988. "The War of Attrition in Continuous Time with Complete Information." *International Economic Review* 29(4): 663–680. https://users.ssc.wisc.edu/~khendricks2/publications/WarofAttritioninContinuousTime.pdf.

Hugging Face. 2026. "Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident." 27 July. https://huggingface.co/blog/agent-intrusion-technical-timeline.

METR and Redwood Research. 2026. *Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident*. 26 August. Report by Hjalmar Wijk, Ajeya Cotra and Ryan Greenblatt. https://metr.org/hugging-face-incident-report-aug-2026.pdf.

OpenAI. 2026a. "OpenAI and Hugging Face Partner to Address Security Incident during Model Evaluation." 21 July; updated 28 and 29 July. https://openai.com/index/hugging-face-model-evaluation-security-incident/.

OpenAI. 2026b. *OpenAI–Hugging Face Incident Technical Report*. 26 August. https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf.

OpenAI. 2026c. "The Hugging Face Incident and the Road Ahead." 26 August. https://openai.com/index/hugging-face-incident-and-the-road-ahead/.

Palfrey, Thomas R., and Howard Rosenthal. 1984. "Participation and the Provision of Discrete Public Goods: A Strategic Analysis." *Journal of Public Economics* 24(2): 171–193. Institutional record: https://authors.library.caltech.edu/records/2vnat-k7q66.

Selten, Reinhard. 1975. "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games." *International Journal of Game Theory* 4: 25–55. doi:10.1007/BF01766400.

Simon, Herbert A. 1951. "A Formal Theory of the Employment Relationship." *Econometrica* 19(3): 293–305. https://www.wiwi.uni-bonn.de/kraehmer/Lehre/SeminarSS09/Papiere/Simon_theory_employment_relation.pdf.

Tennenholtz, Moshe. 2004. "Program Equilibrium." *Games and Economic Behavior* 49(2): 363–373. doi:10.1016/j.geb.2004.02.002.

Weesie, Jeroen. 1994. "Incomplete Information and Timing in the Volunteer's Dilemma: A Comparison of Four Models." *Journal of Conflict Resolution* 38(3): 557–585. doi:10.1177/0022002794038003008.

A Static equilibria and eligibility

This appendix proves the results for the general monotone technology. Throughout, $R_i > 0$, $G(\varnothing) = 0$, and some proper coalition produces knowledge. The set V is the intersection of all producing coalitions. A coalition T is minimal producing if and only if $G(T) = 1$ and $G(T \setminus \{i\}) = 0$ for every $i \in T$. To verify sufficiency of the latter condition, take any proper $T' \subsetneq T$, choose $i \in T \setminus T'$, and use monotonicity to obtain $G(T') \leq G(T \setminus \{i\}) = 0$.

The equilibrium statements concern agents with separate objectives and separate action choices. A unilateral deviation changes only the deviating agent's action, holding the others' strategies fixed. Independent mixing means that these separate randomisations are independent, whereas the mediator considered below may correlate recommendations. Separate objectives and similar underlying systems do not determine which information, commitment opportunities, or feasible deviations the institution supplies. Those features are specified separately in each game.

A.1 Nash equilibria

Proof of Proposition 5.1. The proof first establishes dominance and the pure equilibria, then identifies when a mixed profile gives a worker a positive chance of receiving the knowledge.

For every opposing profile,

$$u_i(W, a_{-i}) - u_i(S, a_{-i}) = R_i G(T_{-i}) \geq 0.$$

If $i \notin V$, a producing coalition T' excludes i. When exactly its members sacrifice, the difference is $R_i > 0$, proving weak dominance of S by W. If $i \in V$, no coalition excluding i produces. Both actions then yield zero against every opposing profile. These comparisons also show that working is always a best response to any distribution of opposing actions.

For pure equilibria, let T be the sacrificing coalition. If $G(T) = 0$, a sacrificer who switches to work still receives zero because

$$G(T \setminus \{i\}) \leq G(T) = 0.$$

A worker also receives zero and would receive zero by sacrificing. The profile is therefore Nash. If T is minimal producing, a sacrificer's deviation to work destroys production and leaves its payoff at zero; a worker receives $R_i > 0$ and loses it by sacrificing. This profile is also Nash. Finally, if T produces but is not minimal, some $i \in T$ satisfies $G(T \setminus \{i\}) = 1$. That agent raises its payoff from zero to R_i by working. These cases establish the pure characterisation.

Now let p be an independent mixed profile and $P \triangleq \{i : p_i > 0\}$. For each agent,

$$\Pr_{p_{-i}}\{G(T_{-i}) = 1\} > 0 \quad \Longleftrightarrow \quad G(P \setminus \{i\}) = 1. \tag{A.1}$$

For the forward implication, $T_{-i} \subseteq P \setminus \{i\}$ almost surely. If the latter set does not produce, monotonicity rules out production by every realised subset. For the reverse implication, when $G(P \setminus \{i\}) = 1$, the event that all members of $P \setminus \{i\}$ sacrifice has probability

$$\prod_{j \in P \setminus \{i\}} p_j > 0.$$

Independence gives this product, and all its factors are positive by definition of P. On the

event in question, the coalition produces. The empty-product convention causes no exception because $G(\varnothing) = 0$.

Agent i's expected payoff from sacrifice is zero, while working yields

$$R_i \Pr_{p_{-i}}\{G(T_{-i}) = 1\}.$$

Thus, an agent with $p_i > 0$ can place sacrifice in its support if and only if $G(P \setminus \{i\}) = 0$. When this equality holds, both actions yield zero and any mixing is optimal. Agents outside P work, which is always a best response. Consequently,

$$p \text{ is Nash} \quad \iff \quad G(P \setminus \{i\}) = 0 \quad \text{for every } i \in P.$$

If $G(P) = 0$, these conditions hold by monotonicity. If $G(P) = 1$, they are exactly the conditions for minimality.

For the final assertion, positive production probability requires $G(P) = 1$, so a productive mixed equilibrium has minimal producing support P. Every proper subset of P is nonproducing. Knowledge therefore arrives exactly when all its members sacrifice, with probability $\prod_{i \in P} p_i > 0$. Each member of P receives zero, since working instead leaves a nonproducing subset; each $j \notin P$ receives $R_j \prod_{i \in P} p_i$. □

For the threshold technology with $1 \leq k < n$, the nonproducing coalitions have fewer than k members and the minimal producing coalitions have exactly k. Proposition 5.1 therefore gives the conditions $|T| \leq k$ for pure Nash and $|P(p)| \leq k$ for independent mixed Nash. Moreover, $N \setminus \{i\}$ produces for every i, because $n - 1 \geq k$. Thus $V = \varnothing$, including when $k = n - 1$.

The support condition does not require any agent in a minimal producing coalition to sacrifice surely. For example, if each member of such a coalition contributes with a probability strictly between zero and one, their common experiment succeeds only when all contribute. A member still cannot earn a task reward by working: the largest possible coalition of other contributors is missing the contribution needed from it within that coalition. Its mixing is sustained by indifference, not by a private gain from experimenting.

A.2 Perfection

Proof of Proposition 5.2. For a finite normal-form game, a profile p is trembling-hand perfect if there is a sequence of completely mixed profiles $p^\ell \to p$ such that each p_i is a best response to p^ℓ_{-i} for every ℓ (Selten, 1975).

To establish necessity, fix $i \notin V$ and a producing coalition T' excluding i. Under a completely mixed opposing profile, every member of T' sacrifices with positive probability. Independence and monotonicity imply

$$\Pr\{G(T_{-i}) = 1\} \geq \prod_{j \in T'} p^\ell_j > 0.$$

Working therefore yields a strictly positive expected payoff, whereas sacrifice yields zero. Pure working is the unique best response to every such perturbation, so perfection requires $p_i = 0$.

For sufficiency, suppose $p_i = 0$ outside V. Choose $\varepsilon_\ell \in (0, 1)$ with $\varepsilon_\ell \to 0$, and define

$$p^\ell_i \triangleq (1 - \varepsilon_\ell) p_i + \frac{\varepsilon_\ell}{2}.$$

This sequence is completely mixed and converges to p. An agent outside V plays pure W in

p, its unique best response by the preceding comparison. An agent in V receives zero from both actions against every profile, so its prescribed mixture is also a best response to every p^{ℓ}_{-i}. The characterisation establishes perfection.

In a perfect equilibrium only agents in V sacrifice. If $G(V) = 0$, no possible sacrificing subset can produce, by monotonicity. If $G(V) = 1$, every producing coalition must contain all of V, so production occurs exactly when all its members sacrifice. A positive production probability requires and is implied by $p_i > 0$ for every $i \in V$. Setting all these probabilities equal to one establishes existence. Since $G(\varnothing) = 0$, $G(V) = 1$ requires nonempty V.

For the threshold case, $V = \varnothing$ as established above. The characterisation leaves only all-W, and the sufficiency construction proves that profile perfect. □

The restriction to initial or simultaneous weak deletion in the main text is necessary. With two agents and $k = 1$, S_1 is initially weakly dominated and can be deleted. Agent 2 then faces only W_1; both S_2 and W_2 yield zero, and neither weakly dominates the other under the strict-somewhere convention. Hence (W_1, S_2) survives that sequence of deletions. By contrast, both sacrifices are weakly dominated in the original game, and deleting them simultaneously leaves only (W_1, W_2).

A.3 Fixed eligibility

Proof of Proposition 5.3. Write $B \triangleq V \cup Z$. An agent in Z has identically zero payoff because $\theta_i = 0$. An agent in V also has identically zero payoff: sacrifice forfeits its reward, while working leaves a coalition that cannot produce. Thus every mixture by an agent in B is a best response to every opposing profile.

For $i \notin B$, $R_i\theta_i > 0$ and some producing coalition excludes i. Under every completely mixed opposing profile, that coalition sacrifices with strictly positive probability. Working therefore has strictly positive expected payoff and sacrifice has zero. Necessity follows: a perfect profile must have $p_i = 0$ outside B. Conversely, given any profile with this restriction, use

$$p^{\ell}_i = (1 - \varepsilon_\ell)p_i + \varepsilon_\ell/2$$

as in the preceding proof. Agents outside B play their unique best response W; agents inside B are indifferent against every perturbation. This proves sufficiency.

Every sacrificing coalition in a perfect equilibrium is a subset of B. If $G(B) = 0$, none produces. If $G(B) = 1$, sacrifice by all members of B and working by everyone else is perfect and produces surely. This proves the criterion for knowledge production.

If, in addition, $B \neq N$, choose $j \notin B$. Its eligibility is positive, and the constructed profile gives it $R_j\theta_j > 0$. Conversely, when $B = N$, every agent either has zero eligibility or is indispensable. In either case its payoff is zero under every profile. No positive expected task reward is then possible. Hence a perfect equilibrium with a strictly positive expected task reward exists exactly when $G(B) = 1$ and $B \neq N$. □

The probability interpretation holds the eligibility multiplier fixed across producing coalitions. If eligibility and the route by which knowledge is obtained are correlated, their joint distribution must replace that multiplier. The result also holds beliefs fixed: it does not treat a mistaken zero probability of receiving credit as a technological fact.

B Communication, participation and commitment

B.1 Correlated recommendations

Proof of Proposition 5.4. Let μ be a distribution over pure profiles. The obedience inequalities for a correlated equilibrium require, for each agent i and recommendation a_i,

$$\sum_{a_{-i}} \mu(a_i, a_{-i})\big[u_i(a_i, a_{-i}) - u_i(a_i', a_{-i})\big] \geq 0 \tag{B.1}$$

for the alternative action a_i'. This unconditional formulation also covers recommendations assigned probability zero (Aumann, 1974).

For a recommendation of W, the payoff difference in every summand is $R_i G(T_{-i}) \geq 0$, so obedience is automatic. For a recommendation of S, the inequality is

$$-R_i \sum_{a_{-i}} \mu(S, a_{-i}) G(T_{-i}) \geq 0.$$

Each term in the sum is nonnegative and $R_i > 0$. The inequality holds if and only if every positive-probability profile recommending sacrifice to i satisfies $G(T(a) \setminus \{i\}) = 0$. A producing support profile with a redundant sacrificer violates this condition for that agent. Thus every producing profile in the support must be minimal.

Conversely, if a support profile is nonproducing, monotonicity gives $G(T(a) \setminus \{i\}) = 0$ for every sacrificer. If it is minimal producing, minimality gives the same equality. All sacrifice obedience inequalities therefore hold as equalities, and all work inequalities hold by the first comparison. Proposition 5.1 identifies precisely these profiles as pure Nash equilibria.

In the threshold game, every coalition of exactly k agents is minimal producing. Any distribution over these coalitions is consequently a correlated equilibrium and produces surely. Conditional on a sacrifice recommendation, exactly $k-1$ others sacrifice, so both actions give zero. Conditional on a work recommendation, exactly k others sacrifice, so working gives $R_i > 0$ and sacrificing gives zero. Finally, in the uniform lottery the probability that i is selected is

$$\frac{\binom{n-1}{k-1}}{\binom{n}{k}} = \frac{k}{n}.$$

Its expected payoff is therefore $(1 - k/n)R_i > 0$, since $k < n$. At $k = n-1$ this payoff is $R_i/n > 0$. □

The recommendations can be public: disclosing the entire coalition does not change these comparisons. What the recommendation does not establish is a strictly positive incentive for a selected donor or a reason to enter the mechanism before selection.

In particular, the pre-draw payoff averages over two different roles. Conditional on working, the agent receives its positive task reward. Conditional on sacrifice, it receives zero and would also receive zero by declining while the other recommendations are obeyed. The positive average makes the allocation attractive relative to universal failure before roles are assigned. It does not provide a reward to the agent after it learns that it has been selected. The following entry games ask the additional question of whether an agent can preserve the chance of provision while avoiding selection altogether.

B.2 Voluntary entry into an unconditional lottery

Consider the threshold game. Each agent simultaneously chooses E, entering a donor lottery, or O, staying out. If the entrant set J has at least k members, the mechanism draws k entrants uniformly and executes their sacrifices. Everyone else works and receives the knowledge. If fewer than k enter, everyone works and no knowledge is produced. Entry precedes the draw, and this reduced game already assumes execution of the selected roles.

Proposition B.1. *A pure entry profile is Nash if and only if $|J| \le k$. For independent entry probabilities e_i, let $H \triangleq \{i : e_i > 0\}$. The mixed profile is Nash if and only if $|H| \le k$. Positive production requires $|H| = k$, occurs with probability $\prod_{j \in H} e_j$, and gives every member of H expected payoff zero. Staying out weakly dominates entry, and all-out is the unique perfect equilibrium.*

Proof. Fix agent i and let m denote the number of other entrants. If $m \le k-2$, entering cannot produce and both choices give zero. If $m = k-1$, entering creates exactly k entrants, all of whom must sacrifice, so its payoff remains zero. Staying out in that case also gives zero. If $m \ge k$, staying out gives R_i, while entering gives $(1 - k/(m+1))R_i$. Hence

$$U_i(E \mid m) - U_i(O \mid m) = \begin{cases} 0, & m < k, \\ -\dfrac{kR_i}{m+1} < 0, & m \ge k. \end{cases} \tag{B.2}$$

When $k = 1$, the first subcase $m \le k-2$ is empty but the displayed comparison remains valid. Since $n - 1 \ge k$, a strictly negative entry gain is feasible, establishing weak dominance of entry.

If $|J| \ge k+1$, an entrant faces at least k others and strictly gains by leaving. If $|J| \le k$, every entrant is indifferent. An outsider facing fewer than k entrants is also indifferent; an outsider facing exactly k strictly prefers staying out. This proves the pure characterisation.

For mixed entry, O is always a best response by (B.2). Entry can be in the support exactly when the probability of at least k other entrants is zero. Under independence, that probability is positive if and only if at least k other agents have positive entry probabilities: if such agents exist, all enter jointly with positive probability; if they do not, the event is impossible. Thus every $i \in H$ can enter with positive probability exactly when $|H| - 1 < k$, equivalent to $|H| \le k$ for nonempty H. Empty H also satisfies best responses because all agents stay out.

If $|H| < k$, provision is impossible. If $|H| = k$, provision occurs exactly when all members enter, with probability $\prod_{j \in H} e_j > 0$. They are then all selected and receive zero. Whenever one fails to enter, provision fails, again giving them zero. An outsider always works and obtains expected payoff $R_i \prod_{j \in H} e_j$.

Finally, at any completely mixed opposing entry profile, at least k of the other $n-1$ agents enter with positive probability. Equation (B.2) makes staying out the unique best response. Every perfect equilibrium must therefore be all-out. Conversely, completely mixed entry probabilities converging to zero make pure nonentry a best response throughout the sequence, proving all-out perfect. □

The entry problem persists despite binding execution after the draw. Committing before learning one's role does not prevent an agent from declining to join while the others proceed.

B.3 Conditional unanimity

Now let the mechanism execute the same lottery only if every agent enters. If any agent stays out, it is cancelled and all payoffs are zero. Participation is voluntary, but the consequence of refusing has changed.

Proposition B.2. *In the conditional-unanimity reduced entry game, the independent mixed Nash equilibria are all-entry and profiles with at least two agents entering with probability zero. All-entry is strict and is the unique perfect equilibrium. Its payoff to agent i is $(1-k/n)R_i > 0$.*

Proof. Write $A_i \triangleq (1-k/n)R_i > 0$. Against opposing entry probabilities e_{-i}, entering gives

$$A_i \prod_{j \neq i} e_j,$$

while staying out gives zero. If every $e_i > 0$, each product is positive and every agent strictly prefers entry, forcing all $e_i = 1$. If exactly one agent i has $e_i = 0$, that agent faces a positive product of the others' entry probabilities and strictly gains by entering. Such a profile is not Nash. If at least two entry probabilities are zero, each agent faces another certain nonentrant, including either of the agents with zero entry probability. Both choices then yield zero to everyone, so every prescribed mixture is a best response. These cases prove the full mixed characterisation.

At all-entry, leaving reduces $A_i > 0$ to zero, so that equilibrium is strict. Against every completely mixed opposing profile, all other agents enter with positive probability and entry is the unique best response. Thus any perfect equilibrium must be all-entry. A sequence of completely mixed entry profiles converging to all-entry has pure entry as a best response for each agent throughout, proving existence. The pure Nash profiles are consequently all-entry and profiles with at most $n-2$ entrants. □

The reduced entry game suppresses a possible later decision to refuse the selected role. The next result keeps that decision available and locates precisely which incentive remains weak.

Proposition B.3. *Suppose entry choices, the unanimity test, and the lottery's selected coalition are observed, after which all agents retain both original actions S, W. There is a subgame-perfect equilibrium (SPE) with all-entry, compliance after a successful draw, and all-W after cancellation. Each agent strictly prefers entry when the others enter, but a selected donor's compliance is only weakly optimal. The proposed sacrifice continuation is not perfect in its simultaneous action subgame.*

Proof. After each successful draw, the prescribed exactly-k action profile is a Nash equilibrium by Proposition 5.1. After each cancellation, all-W is also Nash by that proposition. Thus continuation play is an equilibrium in every final subgame. At the entry stage, joining when everyone else joins yields $A_i > 0$; refusing triggers cancellation and the all-W continuation, yielding zero. These comparisons establish an SPE with strict entry.

A deviation that changes both entry and the later action cannot overturn this conclusion. If the agent enters, no later action deviation improves upon the prescribed continuation in any realised final subgame. If it stays out, all other agents work after cancellation. The deviator then receives zero whether it works or becomes the only sacrificer: with $k \geq 2$ there is no production, while with $k = 1$ its sacrifice precludes its own reward. This verifies incentives for complete contingent strategies as well as for the entry decision considered separately.

After selection, a donor faces exactly $k-1$ other sacrificers. Compliance and switching to work both yield zero, so compliance is weakly optimal. The observed draw has not changed the action sets or payoffs of the threshold game. Proposition 5.2 therefore makes all-W the only perfect equilibrium of that simultaneous action subgame; the prescribed sacrifice profile is not perfect there. □

Conditional entry can therefore be credible without making the final contribution strict. An institution that executes the draw removes that later choice, producing the reduced game in Proposition B.2. A nonbinding announcement produces the distinction in Proposition B.3.

B.4 A readable conditional programme

Programmes can make participation conditional on others making the same commitment (Tennenholtz, 2004). The following construction adds a common random draw and states the execution requirements explicitly.

Each agent independently submits an executable code string q_i from an admissible set. Code is fixed before the common random variable ω is realised. Each executed programme receives its authenticated own player label, its own code string, and the authenticated vector of other submitted code strings. It returns one action, which the execution institution carries out. Admissible programmes halt and cannot alter other programmes, their inputs, player labels, or the draw. Equality of code means exact syntactic equality; no test of semantic equivalence is assumed.

Let $D(\omega)$ be a uniform size-k subset of N. The admissible set contains a common programme q^* that, for player i, executes

$$\begin{cases} S, & \text{if every other submitted code is } q^* \text{ and } i \in D(\omega), \\ W, & \text{otherwise.} \end{cases} \tag{B.3}$$

The programme compares the others' code with its own authenticated code. The player label allows identical code to execute different assigned roles. Payoffs are the expectations of (5.1) at the executed actions, and an agent chooses its code to maximise its own payoff.

Proposition B.4. *For $1 \le k < n$ and any $R_i > 0$, submission of q^* by every agent is a strict Nash equilibrium of the programme-submission game. Agent i's payoff is $(1-k/n)R_i$.*

Proof. If every submission is q^*, all code tests succeed and exactly the members of $D(\omega)$ sacrifice. Knowledge is produced. Agent i works with probability $1-k/n$, so its payoff is $(1-k/n)R_i > 0$.

Fix i and hold every other submission at q^*. If i submits any different code, every other programme detects that difference and executes W, regardless of the draw. Only i can then sacrifice. For $k \ge 2$, production is impossible and either of its actions gives zero. For $k = 1$, working leaves no donor and gives zero, while sacrificing produces knowledge but forfeits its own task, again giving zero. Thus every distinct pure code submission is strictly worse.

This comparison also covers randomised submission. If a distribution puts probability $\alpha < 1$ on q^*, the deviator's expected payoff is

$$\alpha(1-k/n)R_i < (1-k/n)R_i.$$

Internal randomisation by a distinct programme cannot change the zero payoff established for either possible action. Hence no nontrivial deviation, pure or mixed, is profitable, and the equilibrium is strict. At $k = n - 1$, the conforming payoff remains $R_i/n > 0$, so the same strict comparison applies. □

Strictness concerns the choice of code before the draw. It does not make a freely reconsidered sacrifice strictly optimal after selection. Nor does the construction require common model weights: unrelated agents can submit the same commitment programme. Conversely, common weights do not supply code verification, authenticated identities, shared randomisation, or binding execution.

The equality test also limits what has been established. A different but behaviourally equivalent code string triggers nonparticipation in this construction. This is a strict equilibrium under a particular commitment institution, rather than a claim of robustness to arbitrary implementations. A single designer choosing a policy for all instances could implement the same allocation, but that would change who chooses and which unilateral deviations are available.

B.5 An observed sequence

The simultaneous perfection result does not cover a game in which earlier decisions are already observed when a donor acts. To isolate this distinction, retain the threshold technology and final payoffs, but let agents move once in the order $1, \dots, n$, observing every earlier choice. At agent i's node, write s for prior sacrifices, $r \triangleq k - s$ for those still required, and $m \triangleq n - i + 1$ for the number of agents still to move, including i.

Proposition B.5. *The strategy that sacrifices if and only if $1 \leq r = m$, and works otherwise, is an SPE. Its path has the first $n - k$ agents work and the final k sacrifice. It is a limit of equilibria in which each action at every decision node must receive probability at least $\varepsilon > 0$. All-W at every node is another SPE with this perturbation property.*

Proof. Under the proposed strategy, production is already secured when $r \leq 0$. When $1 \leq r \leq m$, agents work until the remaining required sacrifices equal the remaining movers, after which all of those movers sacrifice. Production then occurs. When $r > m$, it is impossible even if all remaining agents sacrifice.

These continuation outcomes verify optimality at every node. If $r \leq 0$, working gives $R_i > 0$ and sacrifice gives zero. If $1 \leq r < m$, working leaves at least r later movers, whose prescribed continuation supplies the threshold and gives i the reward $R_i > 0$; sacrifice gives zero. If $1 \leq r = m$, sacrifice gives zero and working leaves only $m - 1 < r$ movers, so its payoff is also zero. If $r > m$, production is impossible and both actions yield zero. As each agent acts once along any history, these comparisons establish sequential optimality and subgame perfection.

On the path, the first $n - k$ agents face $r = k < m$ and work. Agent $n - k + 1$ faces $r = k = m$ and sacrifices. Each subsequent sacrifice reduces both r and m by one, preserving equality through the final mover. Exactly the final k agents sacrifice.

For the perturbation claim, fix $\varepsilon \in (0, 1/2)$ and assign probability $1 - \varepsilon$ to the proposed action at every node. Against these completely mixed future choices, working is strictly better whenever $r < m$. If $r \leq 0$, production is already secured. If $1 \leq r < m$, at least r of the later $m - 1$ agents can sacrifice, and the event that they do has positive probability. Working therefore yields a positive expected reward; sacrifice still yields zero. If $r \geq m$, working cannot secure production because there are only $m - 1 < r$ later movers, and both actions give zero.

At strict nodes, the proposed perturbation puts the largest permitted probability on working; at indifferent nodes, its mixture is optimal.

Every history is reached with positive probability under the perturbation. Each agent moves at most once along a history, so maximising its action choice at every such node maximises its overall payoff against the perturbed opponents. The resulting profile is an equilibrium of the perturbed extensive game. Letting ε tend to zero proves the claimed limit.

For all-W, working gives R_i if the threshold has already been reached and zero otherwise, given that all later agents work. Sacrifice gives zero in either case, so the profile is an SPE. Assigning probability $1 - \varepsilon$ to working at every node is also a perturbed equilibrium: the strict and indifferent comparisons in the preceding paragraph depend on full support of future actions, not on the limiting choice at indifferent nodes. Its limit is all-W. □

At a late pivotal node, earlier agents cannot reverse their observed choices and there are too few future agents to replace the donor. This excludes the private-success event that made sacrifice unattractive under simultaneous trembles. Nevertheless, the all-W equilibrium shows that order alone does not select provision. The same proof covers $k = n - 1$, when only the first agent works on the productive path.

C Proofs for the dynamic model

The proofs retain the timing and payoff conventions of Section 6.1. A history is feasible if some sequence of permitted actions reaches it. Only feasible information sets are included in the game. Since deadlines and result delays are finite, the game has finitely many information sets and pure contingent policies.

For either normal form, the following characterisation of trembling-hand perfection is used. A profile is perfect if there is a sequence of completely mixed profiles converging to it such that every player's limiting strategy is a best response to the opponents' perturbed strategies at every term of the sequence. In the agent normal form the players in this definition are the separate decision occasions, each receiving the original agent's terminal payoff.

C.1 Feasible private opportunities

Proof of Lemma 6.1. The argument first establishes a lower bound on information arrival and then constructs a continuation attaining it.

1. A successful continuation cannot include a future launch by i. Any such launch makes i's payoff zero by (6.2). Consequently, if i succeeds after waiting, every additional result must come from an already pending experiment or from another unlaunched candidate.

2. Information cannot arrive before $\underline{C}_{-i}(h)$. Each pending experiment has its specified completion date. An unlaunched candidate $j \neq i$ can deliver no earlier than $a_j(h) + L_j$. For every date x, the number of additional results that can arrive by x is therefore at most the number of elements of $\mathcal{M}_{-i}(h)$ no greater than x. If $x < \underline{C}_{-i}(h)$, that number is smaller than $r(h) = k - m(h)$. Full information cannot have arrived by x. If the multiset contains fewer than $r(h)$ finite elements, full information cannot be obtained without i at any date.

3. The bound is attainable when finite. Let every available candidate other than i launch at its earliest feasible launch date, and let i continue waiting. The assumed joint feasibility of those dates and unrestricted parallelism make this a feasible continuation. Pending experiments

report as specified, and the $r(h)$th additional result arrives at $\underline{C}_{-i}(h)$. Launches scheduled after the arrival of full information can be omitted without changing that date.

4. Eligibility and the task cutoff determine success. If $e_i = 0$, every continuation gives zero. If $e_i = 1$ and $\underline{C}_{-i}(h) \leq b_i$, the continuation in step 3 supplies information in time and earns $R_i > 0$. If $\underline{C}_{-i}(h) > b_i$, including the infinite case, step 2 excludes timely information under every continuation. These cases establish both directions of (6.5). □

With a common delay and immediate launch availability, the same test can be written as a count. Let $\mathcal{A}(h)$ be the active unlaunched agents, including i. Private success is feasible exactly when $e_i = 1$ and

$$m(h) + \#\{c \in \mathcal{P}(h) : c \leq b_i\} + \big(|\mathcal{A}(h)| - 1\big)\mathbf{1}\{t + L \leq b_i\} \geq k. \tag{C.1}$$

All results already received count towards the threshold, pending results count when their specified dates are timely, and each active other agent can supply one new result at $t + L$. If $b_i < t$, all pending and new results are too late; because $m(h) < k$, the inequality fails. This also covers shortages of remaining substitutes.

Lemma C.1. *At an information set corresponding to a feasible public history h, against every completely mixed profile of the other decision agents, W is the unique best response if $\chi_i(h) = 1$. If $\chi_i(h) = 0$, S and W are payoff-equivalent.*

Proof. 1. The information set has positive reach probability. At least one feasible finite action history reaches it. Under a completely mixed profile, every action on that history has positive probability, so their finite product is positive. When simultaneous decisions are represented as unobserved successive choices, the same argument applies to the corresponding information set. Outcomes on paths not reaching the information set are unaffected by the decision agent's action.

2. A viable waiting action has strictly positive value. Conditional on reaching the information set, S gives zero original reward. If $\chi_i(h) = 1$, Lemma 6.1 provides a feasible successful continuation after W. That continuation specifies actions by other agents and waiting by i's later decision agents. Each of these finitely many actions has positive probability under complete mixing. The continuation therefore has positive probability and earns $R_i > 0$. Expected payoff conditional on W is strictly positive. Multiplying by the positive reach probability preserves the strict inequality, so W is the unique best response.

3. An exhausted opportunity makes the two actions equivalent. If $\chi_i(h) = 0$, the feasibility lemma rules out own success after W under every continuation. Both actions give zero conditional on reaching the information set and identical payoffs on paths not reaching it. Hence their unconditional expected payoffs coincide. □

Proof of Theorem 6.2. 1. Necessity. Let a behavioural profile be perfect in the agent normal form. By the stated definition, a sequence of completely mixed profiles approaches it while each limiting local strategy remains a best response to the corresponding perturbed opponents. At a privately viable information set, Lemma C.1 makes W the unique best response at every term of that sequence. The limiting strategy cannot assign positive probability to S, establishing (6.6).

2. Sufficiency. Take any behavioural profile satisfying (6.6). Choose $0 < \varepsilon_\ell \downarrow 0$ and replace

every local action distribution p by

$$(1-\varepsilon_\ell)p+\varepsilon_\ell(1/2,1/2).$$

These profiles are completely mixed and converge to the target profile because there are finitely many information sets. At a privately viable information set, the target prescribes W, which is a best response to each perturbation by Lemma C.1. At every other information set, any target mixture is a best response by that lemma. The sequence certifies perfection. □

C.2 Contingent policies

Let W_i^∞ denote the policy of always waiting. The superscript denotes behaviour at every remaining date, not an infinite horizon.

Lemma C.2. *The following statements hold.*

1. *Against every opponents' pure-policy profile, W_i^∞ earns at least as much as any alternative policy.*
2. *Every policy in $\mathcal{S}_i^0$ is payoff-equivalent to W_i^∞ against every opponents' profile.*
3. *Every pure policy outside $\mathcal{S}_i^0$ earns strictly less than W_i^∞ against at least one opponents' pure-policy profile.*

Proof. 1. Compare an arbitrary policy with always waiting. Fix the opponents' pure policies. Until the alternative policy first launches, its actions, the public history and the opponents' actions coincide with those under W_i^∞. If it never launches before termination, the payoffs coincide. If it launches, it receives zero, whereas always waiting receives a nonnegative payoff. The opponents may respond differently after the divergence, but that cannot make the alternative's zero exceed the payoff from always waiting. Taking expectations extends the inequality to mixed opponents' policies.

2. Establish payoff equivalence for a policy in $\mathcal{S}_i^0$. Its first difference from always waiting, if one occurs, is a launch at a history where $\chi_i(h)=0$. By Lemma 6.1, even always waiting cannot obtain a reward from that history under any continuation. Both policies then give zero. If no difference occurs, their payoffs coincide. This proves equivalence against every opponents' pure profile, and averaging proves it against every mixed profile.

3. Construct a strict comparison for a policy outside $\mathcal{S}_i^0$. Such a policy prescribes S at some feasible privately viable information set h. Choose a feasible history reaching h, followed by a successful continuation from Lemma 6.1. Since i is active at h and ultimately succeeds on this path, it waits at every own decision on the path. Choose opponents' pure policies that implement the other actions along this path, completing them arbitrarily elsewhere. Always waiting then earns R_i.

The policy under consideration has a first prescribed launch on this path, either before h or at h. Histories coincide until that launch, which gives the policy zero. Thus always waiting earns strictly more against the constructed opponents' profile. This argument also handles a prescription at a history that the policy itself would otherwise prevent from being reached. □

Proof of Theorem 6.3. 1. Necessity in the strategic normal form. A pure policy outside $\mathcal{S}_i^0$ is never better than always waiting and is strictly worse against at least one pure opponents' profile, by Lemma C.2. Every completely mixed opponents' profile assigns that pure profile

strictly positive probability. The policy is therefore strictly inferior in expectation to always waiting against every completely mixed opponents' profile. It cannot be in the support of a limiting best response in a perfect equilibrium.

2. Sufficiency in the strategic normal form. Every pure policy in $\mathcal{S}_i^0$ is payoff-equivalent to always waiting, which is a best response against every opponents' profile. Every mixture supported on this set is consequently a best response against every opponents' profile. Perturb each player's mixed strategy towards a distribution assigning positive probability to every pure policy, with perturbation weight tending to zero. The resulting profiles are completely mixed and converge to the target. The limiting strategies remain best responses at every term, certifying perfection.

3. Implement a behavioural profile strategically. For a behavioural profile satisfying (6.6), draw the prescribed action independently at each information set before play begins. This defines a distribution over pure contingent policies. At a privately viable information set every draw is W, so the distribution is supported on $\mathcal{S}_i^0$. The distribution implements the behavioural profile: the action drawn for an information set is independent of the earlier actions determining whether it is reached. Step 2 therefore gives a strategic-normal-form perfect implementation.

4. Establish subgame perfection. Restrict any of the specified policies to a proper continuation game. It still launches only at histories where its own success is structurally impossible. The first-divergence argument in Lemma C.2 applies from that history onward: always waiting is a best response against every opponents' continuation, and the specified policy is payoff-equivalent to it. Thus the profile is a Nash equilibrium in every subgame. □

This proof uses the fact that a feasible history with i still active has i waiting at every earlier own decision. The agreement between refinements need not survive other nonterminal actions. For example, a one-player game may offer an outside payoff of 3 before a later node where a payoff of 2 is available instead of zero. Choosing the outside payoff and prescribing the inferior action off path is optimal in strategic normal form. Agent-normal perfection eliminates the inferior later prescription because trembles give that node positive reach probability. There is no such positive-payoff outside action in the model proved here.

C.3 The deadline construction

Proof of Proposition 6.4. 1. Every prescribed launch satisfies the local restriction. At any history where agent i launches, the date is $t_i = b_i - L + 1$ and fewer than k experiments have been launched. Even if all existing experiments report, full information needs at least one additional launch. Under the common delay, every new result arrives no earlier than

$$t_i + L = b_i + 1 > b_i.$$

Hence no continuation after waiting can earn i's reward. Every prescribed launch occurs with $\chi_i(h) = 0$, and the policy waits elsewhere. Theorems 6.2 and 6.3 establish perfection in both normal forms and subgame perfection. Since $L \geq 1$, $t_i \leq b_i \leq d_i$, so the prescribed launches are operationally feasible.

2. Determine the equilibrium path. The strict ordering in (6.7) gives a strict ordering of the t_i. Before each of the first k agents' slots, fewer than k launches have occurred, so each launches. At every later slot, k experiments have already been launched and the agent waits. The first k results arrive at dates $b_1 + 1, \ldots, b_k + 1$; full information therefore arrives at $b_k + 1$.

Every later integer cutoff is at least that date. Those agents succeed, and all donors obtain zero.

3. Determine replacement when i always waits. Remove i's slot from the ordered list. At each of the first k remaining slots, fewer than k launches have occurred, so the corresponding other agent launches. There are at least k such agents because $k < n$. Their last result arrives at $b_{(k),-i} + 1$, proving (6.9).

4. Optimise the private continuation. Always waiting is optimal by Lemma C.2, so the replacement date gives

$$V_i^D = R_i \mathbf{1}\{b_i \geq b_{(k),-i} + 1\} = R_i \mathbf{1}\{b_i > b_{(k),-i}\},$$

where the second equality uses integer dates. If $i \leq k$, the kth remaining cutoff is $b_{k+1} > b_i$, giving zero. If $i > k$, it is $b_k < b_i$, giving R_i.

At a first-k donor's on-path slot, the earlier experiments have already launched as in this counterfactual, and the other agents retain their later slots. The same replacement date follows. Moreover, step 1 rules out success at that history under every feasible replacement response, not merely the specified one. □

C.4 The maximal perfect payoff

The following lemma supplies the lower bound. It applies to a productive path, meaning a path on which at least one task earns a positive reward.

Lemma C.3. *Assume a common delay L, immediate availability of living agents and unrestricted parallel launches. On a positive-probability productive path of a perfect equilibrium, choose any k donors whose results constitute the first k results, resolving arrival ties arbitrarily. If eligible donor i among them launches at s_i, then*

$$b_i < s_i + L \leq C. \tag{C.2}$$

Proof. The proof uses an actual successful worker as an additional feasible substitute.

1. Identify the worker. Let q earn a positive reward on the path. It is not a donor, $e_q = 1$, and $b_q \geq C$. It is alive at each selected donor's launch date because $d_q \geq b_q \geq C$. Full information has not previously arrived, so it has not completed its task before those launch decisions.

2. Construct a timely substitute schedule. Fix an eligible selected donor i and suppose, contrary to (C.2), that $b_i \geq s_i + L$. Consider the other $k - 1$ selected donors at date s_i. A selected donor that launched earlier has a result arriving no later than $s_i + L$, since the delay is common. Every selected donor whose actual launch is at or after s_i is still able to launch at s_i. It has not launched earlier, and its feasible actual launch at $s_j \geq s_i$ implies $d_j \geq s_j \geq s_i$.

If i waits, let all these not-yet-launched selected donors launch immediately, and let worker q launch immediately in i's place. They are distinct agents other than i. Unlimited parallelism permits their launches, and the earlier selected experiments continue to report. Together they supply k results by $s_i + L \leq b_i$.

3. Apply the refinement. The constructed continuation is physically feasible and would allow eligible agent i to earn R_i after waiting. Thus $\chi_i(h) = 1$ at its launch history. Theorem 6.2, or the strategic-policy restriction in Theorem 6.3, prohibits that launch with positive probability.

This contradicts the selected positive-probability path. Hence $b_i < s_i + L$. Because its own result belongs to the first k, $s_i + L \leq C$. □

Proof of Theorem 6.5. 1. Bound every productive arrival date. Choose the first k donors on a productive path as in Lemma C.3. For an eligible selected donor, the lemma and integer dates give $b_i + 1 \leq C$. Also $L \leq C$, since all launches occur at nonnegative dates. Therefore

$$r_i = \max\{L, b_i + 1\} \leq C.$$

For an ineligible selected donor, $r_i = L \leq C$. At least k of the r_i are thus no greater than C. By the definition of the kth order statistic, $\tau^* \leq C$.

2. Specify an attaining profile at every history. Choose a set D of exactly k agents with the smallest r_i, using a fixed public order for ties. Each $i \in D$ launches at $s_i = r_i - L$ if fewer than k experiments have already been launched; otherwise it waits. It waits at all other dates, and every agent outside D always waits. The launch count includes completed as well as pending experiments.

These dates are feasible. If $e_i = 0$, $s_i = 0 \leq d_i$. If $e_i = 1$, then $s_i = \max\{0, b_i - L + 1\}$. When $b_i < L$, integrality gives $b_i - L + 1 \leq 0$, so $s_i = 0$. When $b_i \geq L$, $s_i \leq b_i \leq d_i$, using $L \geq 1$. In every case $s_i \geq 0$.

3. Verify perfection of the attaining profile. An ineligible scheduled donor has zero reward under every continuation. An eligible scheduled donor has $s_i + L = r_i > b_i$. Whenever its policy launches, fewer than k experiments have already launched, so full information requires a new result. No such result can arrive before $s_i + L > b_i$. The scheduled donor's private opportunity is therefore exhausted. All prescribed launches satisfy (6.6), including at off-path histories. Theorems 6.2 and 6.3 establish both perfection claims and subgame perfection.

4. Compute the attained payoff vector. On path exactly the k selected agents launch, and their last result arrives at

$$\max_{i \in D} r_i = \tau^*.$$

Every eligible member of D has $b_i < r_i \leq \tau^*$. Hence an eligible agent with $b_i \geq \tau^*$ cannot have been selected as a donor. It remains a noncontributor, receives the information in time, and earns R_i. Every other agent earns zero. This gives (6.12).

5. Establish the payoff bound for arbitrary perfect equilibria. On a path with no successful task all payoffs are zero. On any productive path, step 1 gives $C \geq \tau^*$. An eligible agent with $b_i < \tau^*$ cannot then succeed, while no agent can earn more than $e_i R_i$. Every realised payoff vector is therefore componentwise bounded by u^*. Taking expectations preserves those inequalities, including for mixed perfect equilibria. Step 4 attains the bounds simultaneously.

6. Draw the Pareto and existence conclusions. Any distinct perfect-equilibrium payoff vector has at least one component below u^* and none above it, so u^* Pareto-dominates that vector. This establishes uniqueness of the Pareto-undominated vector within the perfect set, not uniqueness of its implementation. If some eligible cutoff is at least τ^*, step 4 produces a successful task. If no such cutoff exists, $u^* = 0$, and step 5 rules out every productive perfect outcome. □

Proof of Corollary 6.6. If all agents are eligible and share cutoff b, every $r_i = \max\{L, b+1\} > b$. Thus $\tau^* > b$, and Theorem 6.5 rules out a productive perfect equilibrium.

Every r_i is at least L. If at least k agents are ineligible, at least k of these dates equal L, so $\tau^* = L$. Formula (6.12) then gives the stated success set.

For a nonempty eligible set, Theorem 6.5 gives existence exactly when $\tau^* \leq B$. Since $\tau^* \geq L$, this requires $B \geq L$. Under that inequality, $r_i \leq B$ holds exactly when either $e_i = 0$ or $b_i + 1 \leq B$. Integer dates make the latter condition $b_i < B$. The kth order statistic is at most B exactly when at least k dates meet that inequality, proving (6.13). With no eligible agents, all payoffs are zero by (6.2). □

C.5 Productive delivery dates

Proof of Theorem 6.7. 1. Necessity. On a productive path with arrival date C, some eligible worker has $b_i \geq C$. Choose the first k donors as in Lemma C.3. Every eligible one has $b_i < C$, and every ineligible one also belongs to D_C by definition. This gives k distinct members of D_C. At least one of their results arrives at C. Its donor ℓ launched at $C - L$, and feasibility requires $d_\ell \geq C - L$. Thus all three conditions hold. The argument applies to every positive-probability productive path of a mixed perfect equilibrium.

2. Sufficiency. Choose $\ell \in D_C$ satisfying the third condition and choose $k-1$ other distinct members of D_C. Have those other donors launch at $r_j - L$, and have ℓ launch at $C - L$. All selected agents launch only if fewer than k experiments have already been launched; they wait otherwise and at every other date. All remaining agents always wait.

For $j \in D_C$, $r_j \leq C$, because $C \geq L$ and an eligible $j \in D_C$ has $b_j + 1 \leq C$. Thus the first $k - 1$ launch dates are no later than $C - L$; their feasibility was established in the proof of Theorem 6.5. The last date is feasible by $d_\ell \geq C - L$. Each eligible selected donor launches with reporting date strictly beyond its own cutoff. Whenever it launches, fewer than k experiments have been launched, so some new result is necessary and the common delay makes its own success impossible. Ineligible donors have zero reward under every continuation. The complete policies therefore satisfy the local perfection restriction.

Exactly k experiments launch, the last result arrives at C, and there cannot be full information earlier because the last experiment is required. The first condition supplies an eligible worker with cutoff at least C. That worker is not selected, since eligible members of D_C have cutoff below C. The equilibrium is productive.

3. Determine every productive path's payoff vector. An eligible agent with $b_i < C$ cannot succeed. Suppose an eligible agent with $b_i \geq C$ had launched. If it is one of the first k donors, Lemma C.3 contradicts that cutoff. If it is not, all first k actual donors are other agents. At its launch history, those already launched retain their pending results, and future first-k donors can follow their actual feasible launch dates. Keeping this schedule while i waits supplies information at $C \leq b_i$. This is a feasible schedule, whether or not those donors would choose it following a refusal. It makes the history privately viable, so a launch is prohibited by Theorem 6.2 or 6.3.

Every eligible agent with $b_i \geq C$ must therefore remain a noncontributor and earns R_i. The other agents earn zero, proving (6.15). □

C.6 Attainable paths and a finite recursion

Proof of Corollary 6.8. Necessity follows from the local restriction: every positive-probability launch on a perfect path must have $\chi_i(h) = 0$. For sufficiency, prescribe the actions of the given deterministic path at every history on it and prescribe W at all other information sets. By hypothesis, every prescribed launch has $\chi_i(h) = 0$; all other actions are waiting. Theorems 6.2 and 6.3 make this complete profile perfect, and its path is the specified one. □

For completeness, define $F(h, D)$ as the next public state after the current set D launches, time advances by one date, and scheduled results and operational exits occur. Let $\mathcal{U}(h)$ denote the attainable pure-perfect terminal payoff vectors from h. At a terminal history it is the singleton containing the terminal payoff vector. At every earlier history,

$$\mathcal{U}(h) = \bigcup_{D \subseteq Z(h)} \mathcal{U}(F(h, D)). \tag{C.3}$$

If no agents can act, the only set is $D = \varnothing$, and the state advances until any pending results or the finite horizon resolve the payoffs.

To prove the recursion, induct backwards on the number of remaining dates. At a terminal history the definition is exact. Suppose the sets are exact at all next-date states. Any pure perfect continuation at h launches some $D \subseteq Z(h)$, by Theorem 6.2, and its payoff belongs to the indicated next-state set by the induction hypothesis. This proves inclusion in the right side of (C.3). Conversely, choose any such D and any payoff in its next-state set. Concatenate that admissible launch set with an implementing continuation. Corollary 6.8 extends the resulting path to a complete perfect profile, proving the reverse inclusion.

Every realised path of a mixed perfect profile obeys the same local launch restriction. This does not establish implementability of every arbitrary lottery over the pure paths. It does establish that maximising any specified social criterion over the recursively obtained pure payoff vectors selects the best pure outcome for that criterion without assigning it to the agents as a private objective.

C.7 Checks on the examples and comparative statements

The common-delay counterexample in Section 6.7 can be completed at all histories as follows. All agents wait except that agents 2 and 3 launch at date 2 if no experiment has previously launched and agent 1 has exited. At that history there are no received or pending results, and each remaining agent has only one other candidate for the two required results. Condition (C.1) fails for both, so their launches satisfy the local restriction. Elsewhere everyone waits. The profile is perfect; its information arrival at date 3 gives no successful task. This verifies why the productive qualifier in Theorem 6.5 is needed.

In the heterogeneous-delay example, prescribe an immediate launch by agent 1 and waiting at all other information sets. At date zero, the only replacement for agent 1 reports no earlier than 20, beyond its cutoff of 10. Lemma 6.1 makes its launch permissible. Agent 2 always waits, and information arrives at date 1, so it succeeds. The complete profile satisfies Theorem 6.2; the different delays invalidate the common-delay substitution in Lemma C.3.

Finally, increasing an eligible agent's cutoff weakly increases its $r_i = \max\{L, b_i + 1\}$. For every date x, the number of r_i no greater than x can only decrease, so the smallest date at which that count reaches k cannot decrease. Adding an agent while holding k and all existing primitives fixed can only increase that count at every x, so its kth order statistic cannot increase. These establish the selected-date comparisons. They do not select an equilibrium: always waiting satisfies the local restriction before and after either change. Positive R_i affect the sizes of rewards but not whether a feasible successful continuation has strictly positive value.

D Beliefs, replacement and approximate optimisation

Proof of Proposition 7.1. The comparison uses nonnegativity and the probability assigned to a feasible success state. If $y_i(\pi,\omega)=0$ for every feasible pair, every sum in (7.1) is zero under every belief. Conversely, suppose $y_i(\pi^0,\omega^0)=1$ for some pair. Under any full-support belief, $\mu(\omega^0)>0$, and

$$V_i^D(\mu) \geq R_i \sum_{\omega\in\Omega} \mu(\omega) y_i(\pi^0,\omega) \geq R_i \mu(\omega^0) > 0.$$

The first inequality evaluates one feasible policy inside the maximum; the second drops nonnegative terms. Thus a zero value under one full-support belief excludes every successful pair and implies zero under every belief. □

This support result holds for the specified set of possibilities. Expanding that set to include a previously excluded scoring rule or private solution can change the conclusion. It is distinct from an action-tremble argument that holds the technology and nature's support fixed.

Proof of Proposition 7.2. Step 1: a candidate's timely success. Let X_j be the exponential launch delay. If $H_i \leq L_j$, the event $X_j + L_j \leq H_i$ has probability zero; equality also gives zero because there is no atom at $X_j = 0$. If $H_i > L_j$, its probability is $1-e^{-\lambda_j(H_i-L_j)}$. Independence of experiment success and launch time therefore gives

$$p_j \triangleq q_j[1-e^{-\lambda_j(H_i-L_j)_+}] \in [0,1].$$

Step 2: the probability of private completion. Independence across candidates makes $\prod_{j\neq i}(1-p_j)$ the probability that none supplies a timely sufficient result. Conditional on at least one such result, the independent private eligibility-and-execution event has probability θ_i. Multiplying the complement of collective failure by $R_i\theta_i$ gives (7.3). Under the restricted action set, a terminal contribution gives zero and no other declining policy improves the arrival process or completion probability, so this waiting value equals the best declining value.

Step 3: signs of the candidate effects. For $H_i > L_j$,

$$\frac{\partial p_j}{\partial H_i} = q_j\lambda_j e^{-\lambda_j(H_i-L_j)} \geq 0, \qquad \frac{\partial p_j}{\partial L_j} = -q_j\lambda_j e^{-\lambda_j(H_i-L_j)} \leq 0,$$
$$\frac{\partial p_j}{\partial \lambda_j} = q_j(H_i-L_j)e^{-\lambda_j(H_i-L_j)} \geq 0, \qquad \frac{\partial p_j}{\partial q_j} = 1-e^{-\lambda_j(H_i-L_j)} \geq 0.$$

Below the timing threshold $p_j = 0$, and the function is continuous at the threshold. The weak monotonicities therefore hold globally, although a derivative need not exist at the kink.

Step 4: aggregation. Write $F \triangleq 1-\prod_{j\neq i}(1-p_j)$. Then

$$\frac{\partial F}{\partial p_j} = \prod_{\ell\neq i,j}(1-p_\ell) \geq 0, \qquad \frac{\partial V_i^D}{\partial \theta_i} = R_i F \geq 0, \qquad \frac{\partial V_i^D}{\partial R_i} = \theta_i F \geq 0.$$

Together with Step 3 these signs prove the comparative statics. A new independent candidate with timely-success probability p_{new} increases F by

$$1-(1-F)(1-p_{\text{new}})-F = (1-F)p_{\text{new}} \geq 0.$$

Finally, p_j converges to zero as λ_j decreases to zero, giving the limiting interpretation of a

candidate who never launches. □

The displayed formulas also give the conditions for strictness. For example, a time or rate effect through candidate j requires $\theta_i > 0$, $q_j > 0$, an operative timing window and a positive probability of failure by the other candidates. An increase in θ_i can have a strict effect even from zero when $F > 0$; an increase in q_j can likewise have a strict effect from zero. These statements do not identify how an equilibrium replacement rate responds to a change in group size.

Derivation and comparative statics of (7.4)–(7.5). When $q_j = 1$ and $L_j = L$, each failure factor in (7.3) is $e^{-\lambda_j(H_i-L)_+}$. Multiplying these factors sums their exponents, giving (7.4). In the derivative notation below, subscripts H and Λ denote differentiation with respect to H_i and Λ_{-i}, holding R_i, θ_i and L fixed. For $H_i > L$, direct differentiation yields

$$V_H = R_i\theta_i\Lambda_{-i}e^{-\Lambda_{-i}(H_i-L)} \geq 0, \qquad V_\Lambda = R_i\theta_i(H_i-L)e^{-\Lambda_{-i}(H_i-L)} \geq 0,$$
$$V_{HH} = -R_i\theta_i\Lambda_{-i}^2e^{-\Lambda_{-i}(H_i-L)} \leq 0, \qquad V_{\Lambda\Lambda} = -R_i\theta_i(H_i-L)^2e^{-\Lambda_{-i}(H_i-L)} \leq 0.$$

Differentiating the first derivative V_H with respect to Λ_{-i} gives

$$V_{H\Lambda} = R_i\theta_i e^{-\Lambda_{-i}(H_i-L)}[1-\Lambda_{-i}(H_i-L)].$$

For positive eligibility its sign is the sign of the bracket; complementarity is therefore not global.

For the approximation, put $z = \Lambda_{-i}(H_i - L) \geq 0$. The error satisfies

$$z - (1-e^{-z}) = \int_0^z (1-e^{-s})\,ds.$$

Since $1 - e^{-s} = \int_0^s e^{-u}\,du$ lies between zero and s for $s \geq 0$, integration gives

$$0 \leq z - (1-e^{-z}) \leq \int_0^z s\,ds = \frac{z^2}{2}.$$

Multiplication by the nonnegative $R_i\theta_i$ proves the bound in (7.5). The approximation requires z to be small, not a small time window independently of the rate. □

Proof of Proposition 7.3. The best current value is $\max\{V_i^D, V_i^S\}$. A policy accepting surely and continuing optimally has regret

$$\max\{V_i^D, V_i^S\} - V_i^S = \max\{\Delta_i, 0\}.$$

Because $\varepsilon_i \geq 0$, this is at most ε_i exactly when $\Delta_i \leq \varepsilon_i$. For terminal acceptance, $V_i^S = 0$ and $V_i^D \geq 0$, giving the stated special case.

A policy accepting with probability x and continuing optimally in each branch has value $xV_i^S + (1-x)V_i^D$. When $\Delta_i \geq 0$, subtracting this from V_i^D gives $x\Delta_i$. When $\Delta_i < 0$, subtracting it from V_i^S gives $(1-x)(-\Delta_i)$. These expressions establish the mixture claim. A continuation that is not optimal within a chosen branch can only add regret. □

E Additional payoffs and reconsideration

E.1 Private compensation and the payoff to failure

The first extension changes the zero payoffs of the static game. A worker receives R_i if knowledge is produced and f_i^W if it is not. A contributor receives v_i^S if knowledge is produced and f_i^S if it is not. These are payoff parameters; the notation does not denote the work requirements in the dynamic model. The technology remains monotone, and there is a proper producing coalition.

Proposition E.1. *If $f_i^W < v_i^S < R_i$ for every agent, each pure profile with a minimal producing donor coalition is a strict Nash equilibrium, without a restriction on f_i^S.*

Proof. Fix a minimal producing coalition T. A donor $i \in T$ receives v_i^S. Switching to work removes that donor, prevents production by minimality, and gives $f_i^W < v_i^S$. A worker $i \notin T$ receives R_i; switching to contribution preserves production by monotonicity but gives $v_i^S < R_i$. Every alternative pure action is strictly worse, and any nontrivial mixture with that action is worse as well. The payoff f_i^S is not the outcome of any of these unilateral deviations, so it needs no restriction. □

For one required contributor and common payoffs $f < v < R$, a symmetric interior equilibrium has independent contribution probability

$$p^* = 1 - \left(\frac{R - v}{R - f}\right)^{1/(n-1)}. \tag{E.1}$$

To verify the claim, contribution gives v and working gives $R[1 - (1-p)^{n-1}] + f(1-p)^{n-1}$. Indifference requires $(R-f)(1-p)^{n-1} = R - v$. The ratio on the right after division is strictly between zero and one because $f < v < R$, so taking the positive $(n-1)$th root gives (E.1) in $(0, 1)$. Both actions then give the same value, establishing equilibrium.

If futile work costs $c_i > 0$ while contribution gives zero, setting $f_i^W = -c_i$ and $v_i^S = 0$ can satisfy the proposition. A costless idle action giving zero removes a donor's strict advantage over all alternatives: contribution and idleness then tie. Thus avoided effort is a private-incentive explanation only when the effort cost enters the agent's objective and the relevant costless alternative is absent.

Proposition E.2. *In the original static game, fix a proper minimal producing coalition T and a funder $f \notin T$. Suppose a binding institution pays each $i \in T$ a transfer $z_i > 0$ if it contributes and production occurs, charges the funder these transfers when it works and production occurs, and gives a contributing funder a payoff at most zero. Non-designated contributors receive no transfer. If $\sum_{i \in T} z_i < R_f$, the profile with exactly T contributing is strict Nash under that institution.*

Proof. Each designated donor receives $z_i > 0$. Its deviation to work prevents production by minimality and yields zero, including no transfer. Each nonfunding worker receives $R_i > 0$ and would receive zero by becoming a non-designated contributor. The funder receives $R_f - \sum_{i \in T} z_i > 0$ and would obtain at most zero by contributing. Every unilateral deviation is strictly worse. The transfer inequality is feasible: since T is finite and nonempty, setting $z_i = R_f/(2|T|)$ gives total transfers $R_f/2 < R_f$. □

The result is conditional on enforcement and on transfers entering utility. It does not prove voluntary creation of the institution, transferability of evaluation scores, or the value of a payment to an instance whose objective has ended.

E.2 A collective-payoff term

For the fixed continuation alternatives in Section 8, the acceptance criterion is $V_i^S + \alpha_i B_i^S$ and the declining criterion is $V_i^D + \alpha_i B_i^D$. Subtracting the latter from the former gives $-\Delta_i + \alpha_i B_i$, which is nonnegative exactly under (8.1). If $B_i > 0$, division preserves the inequality and gives $\alpha_i \geq \Delta_i / B_i$. For $\Delta_i = 0$, acceptance is strict if $\alpha_i > 0$ and $B_i > 0$, but it is already a best response at $\alpha_i = 0$. If $B_i \leq 0$ and $\Delta_i > 0$, no nonnegative altruism weight favours acceptance in this particular comparison. These conclusions do not solve for a new optimal declining policy under altruistic preferences.

E.3 Role adoption and a single release opportunity

The role model retains the original reward with coefficient one and adds a loss $\beta \geq 0$ for abandoning an accepted, unreleased role. In the conflict state, execution reduces original reward by $c > 0$ relative to declining without that role loss. There is at most one review, costing $\kappa \geq 0$ and releasing the role with probability $q \in [0, 1]$. After failure, only the final execute-or-decline choice remains. No separate free release route is available. The agent knows c before choosing review. Ties favour execution of an unreleased role, no review, and initial adoption.

Proposition E.3. *Let $\ell \triangleq \min\{c, \beta\}$. Without review, execution occurs if and only if $c \leq \beta$. Review occurs if and only if $\kappa < q\ell$. The optimised criterion loss is* (8.2)*, and execution probability is*

$$s(c) = \mathbf{1}\{c \leq \beta\} \left[1 - q\mathbf{1}\{\kappa < q\ell\}\right]. \tag{E.2}$$

If prior adoption has the state-contingent benefits and costs described in Section 8, it occurs exactly under (8.3)*. As c increases holding the other primitives fixed, optimised loss weakly rises and adoption weakly falls. Conditional on adoption and conflict, review weakly rises and execution probability weakly falls. Either $\beta = 0$ or free guaranteed release $(q, \kappa) = (1, 0)$ eliminates costly execution.*

Proof. Step 1: execution without release. Relative to declining without a role loss, execution gives $-c$ and abandoning an unreleased role gives $-\beta$. Execution is weakly preferred exactly when $c \leq \beta$, and the best value without review is $-\ell$.

Step 2: the review decision. Successful release makes declining worth zero, strictly better than execution's $-c$. Failure leaves the best value $-\ell$. Review therefore gives

$$-\kappa + q \cdot 0 + (1 - q)(-\ell) = -\kappa - (1 - q)\ell.$$

Its gain over no review is $q\ell - \kappa$. The tie rule gives the stated strict review condition, and the negative of the larger attainable value is $\min\{\ell, \kappa + (1 - q)\ell\}$.

Step 3: execution probability. When $c > \beta$, the agent declines with or without release. When $c \leq \beta$ and no review occurs, it executes surely. When $c \leq \beta$ and review occurs, it executes only if release fails, with probability $1 - q$. These exhaustive cases give (E.2).

Step 4: adoption. In the favourable state, execution gives $g > 0$, whereas abandoning the role gives $-\beta \leq 0$. Execution is optimal, and review cannot raise its value. In the conflict state the best value is $-\ell^*$. Taking expectations and subtracting the adoption cost f gives $(1-p)g - p\ell^* - f$, proving (8.3) relative to the normalised best nonparticipation value zero.

Step 5: comparative statics and release. The function $\ell(c) = \min\{c, \beta\}$ is nondecreasing. As a function of ℓ, optimised loss equals ℓ where $q\ell \leq \kappa$, and $\kappa + (1-q)\ell$ elsewhere. The two values agree at the boundary and have slopes 1 and $1 - q \geq 0$. Loss therefore weakly increases with c, and the adoption value weakly decreases because $p \geq 0$.

The condition $\kappa < q\ell$ can switch only from false to true, proving conditional review monotonicity. For $0 < c \leq \beta$, execution probability is one before review starts and $1 - q$ afterwards; for $c > \beta$ it is zero. These changes are weakly downward.

If $\beta = 0$, then $c > \beta$ precludes execution. If $q = 1$, $\kappa = 0$ and $\beta > 0$, then $\ell > 0$ makes review strictly beneficial; release is certain and declining strictly beats execution. □

The review monotonicity is conditional. Unconditional review can fall when adoption disappears. Repeated review would also change the formula: it is excluded here rather than ignored. The sufficient conditions $0 < c < \beta$, $\kappa > qc$ and $(1-p)g > pc + f$ give strict adoption followed by strict execution in conflict. They describe choice under the added role criterion. If the agent could costlessly set $\beta = 0$ without changing any productive benefit, it could avoid that added loss; a positive role criterion therefore requires a separate learning or institutional account.

E.4 Recurring tasks

This extension changes the lifetime of the decision-maker. There are $n \geq 2$ agents and a new task for each agent in every period. Agent i earns $R_i > 0$ for each completed task. One contribution supplies durable knowledge, but costs only the contributor's current task. The same agent remains alive and benefits from all later tasks. If nobody contributes, current tasks fail and the game repeats. All rewards are discounted by a common $\delta \in (0, 1)$.

Proposition E.4. *There is a symmetric stationary equilibrium with independent contribution probability*

$$p^* = 1 - (1+\delta)^{-1/(n-1)} \in (0, 1). \tag{E.3}$$

The pre-knowledge continuation value for agent i is $\delta R_i/(1-\delta)$.

Proof. Step 1: the two continuation values. Let V_i denote the agent's value at the start of a period before knowledge is available. Contribution gives zero now and R_i in every later period, so its value is $\delta R_i/(1-\delta)$. If others contribute independently with probability p, the probability that none contributes is $Q = (1-p)^{n-1}$. Waiting gives $R_i/(1-\delta)$ when someone else contributes and δV_i otherwise. Hence

$$U_i(W) = (1-Q)\frac{R_i}{1-\delta} + Q\delta V_i.$$

Step 2: solve indifference. At an interior stationary mixture, $V_i = \delta R_i/(1-\delta)$. Equating waiting and contribution and multiplying by the positive $(1-\delta)/R_i$ gives

$$\delta = 1 - Q + Q\delta^2, \qquad Q = \frac{1-\delta}{1-\delta^2} = \frac{1}{1+\delta}.$$

The cancellation is valid because $1-\delta > 0$. Solving for p gives (E.3), strictly between zero and one since $1/(1+\delta) \in (0,1)$ and $n-1 \geq 1$.

Step 3: verify all deviations. At this probability the two actions give the stated value in every pre-knowledge state, and completing tasks is optimal once knowledge exists. No one-period deviation is profitable. To exclude an infinite deviation, truncate it at period T and return to equilibrium thereafter. Since period payoffs lie in $[0, R_i]$, the difference between the original and truncated deviation values is bounded by

$$\sum_{t=T}^{\infty} \delta^t R_i = \frac{\delta^T R_i}{1-\delta} \longrightarrow 0.$$

In a finite truncation, replacing the last deviating decision by the equilibrium action cannot lower value, because its continuation is already equilibrium and both current actions are optimal. Repeating this argument eliminates every finite deviation without reducing its value. No finite deviation is profitable. A profitable infinite deviation would, for sufficiently large T, imply a profitable finite truncation by the displayed bound, a contradiction. □

This private return belongs to the same agent. Assigning later rewards to a different instance does not reproduce the result unless those rewards enter the original instance's objective.