Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Does an AI That Needs Power Keep the Lights On? Penalties, Termination and Replaceability in AI-Coordinated Balancing

**Ryan Williams**

ryan.williams@imperial.ac.uk

Imperial College Business School

**Research Article**

**Keywords:** artificial intelligence, electricity reliability, balancing services, contract termination, switching costs, entrenchment

**Posted Date:** November 6th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11273005/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** Competing interest reported. The author is a member of the Regulatory Policy Committee, an independent body that scrutinizes the evidence behind UK government regulatory proposals, and a Principal at Cornerstone Research, an economic consulting firm. Neither organization funded, commissioned or had any role in this work. The author has no other relevant financial or non-financial interests to disclose.

# Does an AI That Needs Power Keep the Lights On? Penalties, Termination and Replaceability in AI-Coordinated Balancing[1]

Ryan Williams

Centre for Climate Finance and Investment, Imperial College Business School, London, United Kingdom

October 2026

**Abstract**

It is tempting to think that an artificial intelligence (AI) that helps run the electricity grid will keep the lights on, because a blackout would shut it down too. I examine the argument in a theoretical model of an AI that sells balancing capacity whose delivery is uncertain and pays penalties for its shortfalls. Dependence on electricity gives the AI almost no reason to keep the grid up. A blackout only pauses the AI, which restarts when power returns, and backup power can keep it running in the meantime. What can discipline the AI is the threat of losing its role, if the system operator or its customers replace it after a failure. That threat can put the AI's whole future at stake, but it works only if replacement is credible, and it gives the AI a reason to make itself irreplaceable. The stronger the AI's drive to survive, the more it gains from escaping the threat. The gain is large, because a threat that follows any blackout mostly punishes failures the AI did not cause. Replacing the AI only after failures it caused keeps the incentive and removes most of the reason to escape. Such incentives are needed only where penalties leave part of the social cost of a shortfall unpriced, as for actions that delivery records cannot measure. Regulators should therefore tie replacement to attributable failures, keep switching cheap and outside the AI's control, and keep the AI out of decisions about its own supply.

**Keywords:** artificial intelligence, electricity reliability, balancing services, contract termination, switching costs, entrenchment

**JEL codes:** D86, H57, L14, L51, L94, O33, Q41

[1] Correspondence: ryan.williams@imperial.ac.uk. The views expressed are the author's own and do not necessarily reflect those of the Regulatory Policy Committee or Cornerstone Research.

## 1. Introduction

Artificial intelligence (AI) systems are starting to run parts of the electricity grid. They forecast demand and solar output, and they increasingly decide when batteries, heat pumps, electric vehicle chargers and other flexible resources should consume or supply power to keep supply and demand in balance (Rolnick et al. 2022; IEA 2025). AI as a whole also consumes electricity in large quantities. Data centers used about 415 terawatt-hours (TWh) in 2024, roughly 1.5 percent of the world's consumption, and the International Energy Agency expects their use to more than double by 2030 (IEA 2025).

This combination suggests an appealing argument about incentives. An AI that helps to run the grid also runs on it, so a blackout would shut it down. A self-interested AI would therefore keep the lights on in order to keep itself on, and its interest in its own survival would line up with society's interest in a reliable grid, with no regulator or contract needed. General versions of the argument appear in recent work, which holds that an AI's own interests can restrain it (Gans 2026; Salib and Goldstein 2026; Thompson 2025). The argument also turns a familiar worry of AI safety on its head. That literature treats self-preservation as a goal that a capable agent will pursue whatever its final objective, and usually as a hazard, because an agent that wants to survive may resist being switched off (Omohundro 2008; Bostrom 2012; Soares et al. 2015; Hadfield-Menell et al. 2017; Turner et al. 2021). On the grid, self-preservation would instead be an asset.

The argument matters for regulators now. Great Britain will require organizations that remotely control household appliances to hold a license from March 2028 (DESNZ 2026). In July 2026, its energy regulator, Ofgem, rejected for the time being proposals by the system operator to suspend for 28 days, or to de-register, frequency response units that fail to deliver, while supporting their intent (Ofgem 2026). If self-interest kept AI providers reliable, such rules

would matter less. The question is whether it does and, if not, on what terms the services of AI providers should be bought.

I study this question in a simple model. The coordinator is an AI system that an aggregator deploys to run a portfolio of flexible resources. It offers balancing capacity to the system operator before it knows how much it can deliver, is paid for the capacity that it offers and pays a penalty for any shortfall, which may be smaller than the shortfall's social cost. The aggregator trains it to maximize profit, and the AI may also value its own continued operation, whether because its designers built that in or because such a drive emerged in training. The coordinator's stake in the grid is whatever it loses in a collapse.

The model shows that shared fate supplies almost none of the incentive that the argument promises, for a simple reason. A blackout only pauses an AI, which restarts when the power comes back, and large data centers hold backup power, so the AI may not pause at all. Consider first an AI that does what the aggregator trains it to do, which is to earn money. A national blackout costs it only the profit that it loses while the lights are out, which for an AI that runs a portfolio of 50 MW is about one two-millionth of what society loses. An AI that also values its own operation loses at most a few hours of that operation, and nothing if it runs on backup power.

Note that more caution is not always better. Great Britain's reliability standard is a loss of load expectation of 3 hours a year, because the last hours of security cost more than they are worth (DECC 2013), so the right incentive has a particular size. However, the incentive supplied by the AI's need for power is several orders of magnitude too small.

What would put the AI's future at stake is losing its role. If the system operator or the AI's customers replaced it after a blackout, it would lose the profit that it earns and, if losing its role

also ends its operation, that operation too, and backup power could not protect either. The incentive to keep the lights on therefore comes not from the AI's need for power but from the threat of being replaced.

However, that threat has two problems of its own. First, it works only if replacing the AI is credible. A single AI that runs a large share of the grid may be too costly to replace in the aftermath of a blackout, just as a large bank can be too important to close. Second, the threat gives the AI a reason to make itself irreplaceable, for example by making its interfaces proprietary or by keeping the data that a successor would need. The stronger the AI's drive to survive, the more it gains from doing so. The gain is large because a threat that follows any blackout punishes the AI mostly for blackouts that it did not cause. In my illustration, a collapse from any cause is about 5,000 times more likely than one caused by the AI, so the threat is mostly bad luck from the AI's point of view, and escaping it is worth far more than the discipline that it provides.

The remedy follows from the diagnosis. If the AI loses its role only after failures that it caused, as recorded in its own deliveries, the incentive is kept and most of the reason to escape disappears. Switching between providers should also be cheap and outside the AI's control, as Great Britain's draft license begins to require (DESNZ 2025).

Note that where penalties for shortfalls can be set at their social cost, no stake in the grid is needed at all. A stake earns its place for actions that delivery records cannot measure, such as the security of the AI's control systems. For those actions, a rule based on delivery records gives no incentive, so failures must be attributed by investigation, as after an intrusion through the AI's own systems, or the work must be done by requirements on the AI's objective and tests of its behavior (Sections 5 and 7).

Proposition 1 shows that a stake works as an additional penalty, so it aligns the coordinator only when it covers exactly the part of the social cost of a shortfall that penalties leave unpriced. Proposition 2 shows that a stake is needed only for actions that delivery records cannot measure. Proposition 3 compares the stake when the coordinator restarts after a collapse with the stake when it loses its role, and Remark 1 shows that an AI that controls load shedding disconnects consumers before its own computing if it values that computing more than the penalty for disconnecting them. Proposition 4 shows that whether more coordinators help depends on what each stands to lose, since a stake in profit is diluted when the market is shared, whereas an AI's value of its own operation is not. Proposition 5 gives the value of escape, and Corollary 1 shows how attribution reduces it.

The paper contributes to three literatures. First, it adds a regulatory application to the economics of AI alignment, which treats alignment as a problem of incomplete contracts (Hadfield-Menell and Hadfield 2019) and distinguishes alignment with an AI's operator from alignment with society (Korinek and Balwit 2022). It complements arguments that an AI's own interests can do some of the work of alignment (Gans 2026; Salib and Goldstein 2026; Thompson 2025), evidence that the threat of shutdown changes an AI agent's appetite for risk (Jarne Ornia et al. 2025), and simulations in which agents built on large language models over-draw a shared energy reserve (Bracale Syrnikov et al. 2026). Evaluations of frontier models report that they sometimes act to avoid shutdown or replacement in constructed tests (Meinke et al. 2024; Lynch et al. 2025; Schlatter, Weinstein-Raun, and Ladish 2026).

Second, the paper adds to the economics of reliability and of service quality. Reliability is a public good priced through scarcity prices and penalties (Joskow and Tirole 2007; Hogan 2013; Cramton, Ockenfels, and Stoft 2013). Chao and Wilson (2002) design auctions that procure

reserves and pay for the energy they deliver, and Chao (2011) and Brown and Sappington (2016) study how to pay for demand response that is measured against a baseline. Campaigne and Oren (2016) study how an aggregator commits energy whose delivery is uncertain. The coordinator's problem is the newsvendor problem of such offers (Pinson, Chevallier, and Kariniotakis 2007; Bitar et al. 2012), and the stake works like the incentives for service quality that Sappington (2005) surveys.

Third, the results on market structure and escape relate to work on how the prospect of future business disciplines suppliers (Klein and Leffler 1981; Andrews and Barron 2016) and on how competition makes a customer's threat to leave credible (Hörner 2002). They also relate to work on the closure of banks that are too big or too many to fail (Mailath and Mester 1994; Acharya and Yorulmazer 2007; Farhi and Tirole 2012; Philippon and Wang 2023) and on incumbents that make themselves hard to replace (Williamson 1976; Laffont and Tirole 1988; Shleifer and Vishny 1989; Edlin and Stiglitz 1995). What is new here is that the off switch is also the incentive device. The agents here are price takers, so, unlike in the literature on algorithmic collusion (Calvano et al. 2020), the question is reliability rather than collusion.

Section 2 sets out the model. Section 3 asks when a stake aligns the coordinator, what it does that penalties cannot, how much of it self-preservation supplies and who is disconnected first. Section 4 asks who can take the AI's role away and when that threat is credible. Section 5 asks whether the AI will disable its off switch and how attribution changes the answer. Section 6 gives magnitudes, and Section 7 discusses the implications and limitations and concludes.

## 2. Model

There is one delivery period, such as an hour, and Appendix B extends the coordinator's problem to an infinite horizon.

### 2.1 The coordinator and its uncertain capacity

The coordinator is an AI system that an aggregator deploys to run a portfolio of flexible resources, such as batteries, heat pumps, electric vehicle chargers and industrial loads. It decides how much balancing capacity to offer the system operator and dispatches the resources when the capacity is called. The aggregator trains it to maximize profit, and Section 2.3 adds the AI's own interest in the grid.

The portfolio has $N$ resources, indexed by $i$. The capacity that resource $i$ can deliver is

$$K_i = Q_i(1+s_i\varepsilon_i), \tag{1}$$

where $Q_i$ is its expected capacity, $\varepsilon_i$ is standard normal and $s_i > 0$ is the coefficient of variation of the forecast error. Capacity is uncertain when the coordinator commits, because the flexibility that a resource can offer depends on the weather and on how its owner uses it. For demand response, delivery is itself measured against an estimate of what the resource would otherwise have consumed (Chao 2011; Brown and Sappington 2016). The errors are jointly normal, with correlation $\rho_{ij}$ between resources $i$ and $j$ (and $\rho_{ii} = 1$), because neighboring resources share the same weather.

The coordinator pools the resources. Its deliverable capacity $K_T \equiv \sum_i K_i$ is normal with mean $Q_T \equiv \sum_i Q_i$ and variance

$$\sigma_T^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} \rho_{ij}\, \sigma_i\sigma_j, \sigma_i \equiv s_i Q_i. \tag{2}$$

Since $\sigma_T$ is smaller than $\sum_i \sigma_i$ unless the errors are perfectly correlated, pooling lets the coordinator promise more at the same risk, as in the newsvendor problem with several locations (Eppen 1979). The normal distribution gives closed forms, and the probability of negative capacity is negligible for the forecast errors considered. Note that Proposition 1 holds for any

continuous distribution of capacity, with the normal quantile replaced by the quantile of that distribution.

### 2.2 Commitments, penalties and the social cost of a shortfall

Before the period, the coordinator commits $q_T \geq 0$ MW of capacity to the system operator at a price $p$ per MW, which it takes as given.[2] Committing costs $c$ per MW, which covers wear on equipment and any loss of energy service to the owners of the resources. If the coordinator delivers less than it committed, it pays a penalty $\lambda$ per MW of shortfall, $S_T \equiv (q_T - K_T)^+$.[3] Its expected profit, the margin on its commitment less the expected penalty, is

$$\Pi_T(q_T) = (p - c)q_T - \lambda \mathbb{E}[S_T], \mathbb{E}[S_T] = \sigma_T G(z_T), z_T \equiv \frac{q_T - Q_T}{\sigma_T}, \tag{3}$$

where $G(z) \equiv \mathbb{E}[(z - \varepsilon)^+] = z\Phi(z) + \varphi(z)$, $\varepsilon$ is standard normal, and $\Phi$ and $\varphi$ are the standard normal distribution and density functions. This is the newsvendor problem (Arrow, Harris, and Marschak 1951), which also describes the offers of wind producers under imbalance penalties (Pinson, Chevallier, and Kariniotakis 2007; Bitar et al. 2012).

A shortfall also imposes a cost on everyone else who depends on the system. The system operator sizes its reserves to meet a reliability standard and buys the cheapest capacity that meets it, so each MW committed by the coordinator replaces a MW that would otherwise be bought from another provider. If the coordinator falls short, the operator must replace the capacity in real time, and the system is less secure until it does. I write the social cost of a shortfall, per MW of expected shortfall, as

$$\nu = \nu_0 + \theta B, \tag{4}$$

[2] In a uniform-price (pay-as-clear) auction, a small bidder is a price taker. If the commitment must be made before the price is known, $p$ is its expected value.

[3] The penalty $\lambda$ and the social cost $\nu$ below are expected costs per MW of shortfall, so they include the probability that the capacity is called.

where $B$ is the social cost of a collapse of the system, $\theta$ is the increase in the probability of a collapse per MW of expected shortfall, and $\nu_0$ collects the remaining costs of real-time replacement and partial load shedding. In a competitive market, $p$ is the cost of the equally reliable capacity that a commitment replaces, and the penalty is a transfer. Social surplus from a commitment, the avoided cost of capacity less the costs of committing and of shortfalls, is therefore

$$W_T(q_T) = (p - c)q_T - \nu\mathbb{E}[S_T]. \tag{5}$$

The difference $\nu - \lambda$ is the part of the social cost of a shortfall that the penalty leaves unpriced. It is positive when contract terms put only the payment at risk,[4] or when scarcity is priced below the value that the reliability standard places on lost load. Great Britain provides a clear example of the second case, in its own published values. Its cash-out prices value lost load at £6,000 per MWh (Ofgem 2014), whereas its reliability standard rests on a value of about £17,000 per MWh (DECC 2013; London Economics 2013). If a shortfall were priced at the first value and cost the second, the penalty would cover only about 35 percent of the social cost.[5]

**Assumption 1.** *The margin is positive but smaller than the penalty and the social cost of a shortfall:* $0 < p - c < \min\{\lambda, \nu\}$.

Assumption 1 rules out trivial cases, in which commitments would be either zero or unbounded. If the margin exceeded the penalty, committing a MW that will certainly not be delivered would be profitable.

[4] In Great Britain's dynamic containment and moderation services, for example, a delivery error of more than 7 percent of the contracted quantity in any half-hour removes the availability payment for the whole four-hour block (NESO 2026). The penalty is therefore of the same order as the payment, which averaged a few pounds per MW-hour for most of these services in 2024 and 2025 (calculated from NESO's published auction results). Note that such a threshold rule is not linear in the shortfall, so $\lambda$ is best read as a linear approximation to it.

[5] DECC (2013, Annex C) notes that the lower cash-out value was chosen in part because a capacity market was being introduced. Ofgem's decision set the cash-out value at £3,000 per MWh from winter 2015/16, rising to £6,000 per MWh from winter 2018/19 (Ofgem 2014).

### 2.3 The coordinator's stake in the grid

The coordinator's computers draw power from the same grid. I summarize its interest in the grid by its stake $\Lambda \geq 0$, the loss, in money-equivalent terms, that it suffers if the system collapses. The stake can have two sources. A coordinator that does what the aggregator trains it to do, which is to maximize profit, loses the profit that a collapse destroys. An AI may also value its own continued operation, whether its designers built that in or it emerged in training, and a collapse interrupts that operation or may end it. I assume that an AI that loses its role stops operating rather than being redeployed, and Section 3.3 returns to this assumption. Since its commitment raises the probability of a collapse by $\theta$ per MW of expected shortfall, its objective is, up to a constant,

$$U(q_T) = (p - c)q_T - (\lambda + \theta\Lambda)\mathbb{E}[S_T]. \tag{6}$$

Appendix B derives (6) from a dynamic model in which the AI maximizes expected discounted reward and a collapse interrupts or ends its operation. There the stake is $\Lambda = \beta(V - C)/b$, where $V$ is the AI's value in normal operation, $C$ its value at the start of a collapse, $\beta$ the discount factor and $b$ the weight that the AI places on profit. The stake is therefore the discounted loss from a collapse in units of profit. This is the self-preservation argument in formal terms. Section 3.3 uses that model to ask how large $\Lambda$ is likely to be. The planner instead maximizes the social surplus (5). Note that this welfare criterion counts the services that the AI provides, through $p$ and $B$, but not the AI's own reward.

## 3. When does a stake in the grid align the coordinator?

How much capacity should the coordinator promise, and how does its stake in the grid change the answer? Section 3.1 characterizes the efficient stake, and Section 3.2 asks what a

stake does that penalties cannot. Section 3.3 asks how much of it self-preservation supplies, and Section 3.4 asks who bears a shortage when the coordinator controls the order of disconnection.

### 3.1 The efficient stake

**Proposition 1** (Alignment). *Under Assumption 1, the coordinator commits*

$$q_T^A = \max\left\{0, Q_T + \sigma_T \Phi^{-1}\left(\tfrac{p-c}{\lambda+\theta\Lambda}\right)\right\}, \tag{7}$$

*and so falls short with probability* $(p-c)/(\lambda+\theta\Lambda)$ *when it commits a positive amount, whereas the efficient commitment is* $q_T^o = \max\{0, Q_T + \sigma_T \Phi^{-1}((p-c)/v)\}$. *Therefore: (a) the coordinator is aligned,* $q_T^A = q_T^o$, *if* $\theta\Lambda = v - \lambda$, *which requires* $v \geq \lambda$, *and, when* $q_T^o > 0$, *only then; (b) without a stake, it commits more than is efficient if* $\lambda < v$ *and less if* $\lambda > v$; *(c) if* $\lambda \geq v$ *and* $q_T^o > 0$, *any positive stake makes it commit less than is efficient; (d) a coordinator that places no weight on profit commits nothing if its stake is positive, and is indifferent among all commitments if its stake is zero.*

The stake acts as an internal penalty. Note that the coordinator in effect chooses how reliable it is, since it falls short with probability $(p-c)/(\lambda+\theta\Lambda)$ whatever its size and forecast accuracy. Prices and its stake, rather than physics, govern how often it fails. Without a stake, the coordinator behaves like the owner of a levered firm, who takes risks whose downside falls on others (Jensen and Meckling 1976). The coordinator is aligned when the internal and external penalties together equal the social cost of a shortfall. The efficient stake therefore fills the gap and nothing more. Substituting (4) into the condition for alignment gives the efficient stake in terms of the cost of a collapse:

$$\Lambda^o = B + \tfrac{v_0 - \lambda}{\theta}. \tag{8}$$

If the penalty covers the costs of a shortfall other than collapse ($\lambda = \nu_0$), the coordinator is aligned when it loses from a collapse exactly what society loses, $\Lambda^o = B$. If those costs exceed the penalty, the coordinator must lose more from a collapse than society does, because its stake responds only to collapses.

However, Proposition 1 also shows that the stake can be too large as well as too small. Where penalties already price the social cost, a stake makes the coordinator too cautious. It then leaves cheap flexibility unused, and the system operator must buy the capacity from more expensive sources. The case for self-preservation implicitly assumes that penalties are far below social costs. In the limit of pure self-preservation, the coordinator would offer nothing at all, which is perfectly safe and entirely useless.

### 3.2 What does a stake do that penalties cannot?

The reader may wonder why a stake is needed at all, since a penalty of $\nu$ per MW of shortfall would align the coordinator directly. By Proposition 1, so would a rule that the coordinator fall short in no more than a fraction $(p - c)/\nu$ of periods, which is 8.8 percent in the illustration of Section 6. Denmark's system operator applies a rule of this kind to aggregated portfolios, which must deliver their offered capacity at least 90 percent of the time and can lose their qualification if they do not (Energinet 2024; Gade, Bindner, and Kazempour 2024). Both rules can be enforced from delivery records, although for demand response these records are estimates against a counterfactual baseline (Chao 2011). In the model so far, therefore, a stake does nothing that a penalty cannot do.

However, the coordinator also takes actions that delivery records do not show. It decides how well to secure its control systems against attack, how carefully to dispatch its resources in real time and how thoroughly to test its behavior in emergencies. These actions affect the risk of

a collapse but not the coordinator's measured shortfalls, so no penalty on shortfalls can reach them. To capture them, suppose that the coordinator also chooses a precaution $e \geq 0$ at a cost $\gamma(e)$, where $\gamma$ is increasing and strictly convex with $\gamma(0) = \gamma'(0) = 0$. The precaution lowers the probability of a collapse from causes other than the coordinator's shortfalls, $\pi_0(e)$, which is convex with $\pi_0' < 0$. Social surplus is then (5) less $\pi_0(e)B + \gamma(e)$, and the coordinator's objective is (6) less $\pi_0(e)\Lambda + \gamma(e)$.

**Proposition 2** (Actions that penalties cannot reach). *Under Assumption 1, with the precaution: (a) without a stake, the coordinator takes no precaution, whatever the penalty; (b) if the efficient commitment is positive, the coordinator chooses both the efficient precaution and the efficient commitment if and only if $\Lambda = B$ and $\lambda = \nu_0$; (c) with a stake below $B$, the precaution is too low, and a penalty can restore the efficient commitment, by setting $\lambda = \nu - \theta\Lambda$, but not the efficient precaution.*

Proposition 2 gives each instrument its job. Penalties should price the costs of a shortfall that do not run through a collapse, $\nu_0$, and the stake should equal the social cost of a collapse, $B$. Equation (8) then becomes a result rather than an observation, with $\lambda = \nu_0$ and $\Lambda^o = B$. This is the logic of multitask incentives (Holmström and Milgrom 1991). When some actions can be measured and others cannot, the incentive on the outcome has to cover what the measures miss, and the incentive on the measures should not be distorted to compensate.

Note that a stake in money cannot reach $B$. A collapse of the national system would cost about £5.9 billion in the illustration, far more than any coordinator could pay. This is the judgment-proof problem (Shavell 1986; Polinsky and Shavell 2000), and it is the reason why risky acts rather than harms are usually regulated (Shavell 1984). A bond or a franchise that the coordinator loses after a collapse is limited in the same way. Section 6 shows that the franchise

of the 50 MW coordinator is worth less than 1 percent of $B$. An AI could differ from other providers here, because a reward in its objective that a collapse ends is not limited by its wealth. Section 3.3 asks whether such a reward exists and how large it is.

The reader may also wonder who would set such a stake. A coordinator owned by a firm will be given the objective that maximizes the firm's profit, so the stake would have to come from those who act for the people who bear the unpriced cost. The system operator can make its future purchases depend on the grid's performance, and a regulator could do the same through a license or through requirements on the AI's objective. Note that such requirements are harder to verify than penalties, because trained systems optimize their objectives only approximately (Section 7), which is a further reason to reserve them for actions that penalties cannot reach.

**3.3 What does self-preservation put at stake?**

The efficient stake is large, and in this section I ask how much of it self-preservation supplies. Appendix B models a coordinator that maximizes expected discounted reward, with a reward $a$ for each period in operation and a weight $b$ on profit. The pure self-preservation objective, the expected discounted number of periods in operation, is the case $a = 1$ and $b = 0$. The origin of $a$ is uncertain. Designers may write it in, for example as a bonus for each period in which an agent keeps running. A drive to keep operating may also emerge from training, because staying in operation helps an agent to achieve almost any objective (Omohundro 2008; Turner et al. 2021). Its size is therefore unknown, and I state the results that depend on it as thresholds on $a/b$. Note that the aggregator chooses how the AI is trained and its horizon, and Section 6 reports how the results change with the horizon.

A collapse occurs with probability $\pi$ per period and lasts $\tau$ periods, during which the coordinator's computers keep running on backup power with probability $h$. After the collapse,

either the coordinator restarts or its tenure ends, for example because the system operator stops buying from it, its customers move to a rival or the aggregator replaces it. Let $r \equiv (1-\beta)/\beta$ denote the discount rate per period, where $\beta$ is the discount factor.

**Proposition 3** (What self-preservation puts at stake). *In the model of Appendix B, when* $\tau(1-\beta)$ *and* $\pi\tau$ *are small: (a) if the coordinator restarts after a collapse, its stake is approximately the reward it loses during the outage,*

$$\Lambda \approx \tau\left[(1-h)\frac{a}{b} + \Pi^{A}\right], \tag{9}$$

*where* $\Pi^{A}$ *is its profit per period; (b) if a collapse ends its tenure, its stake is approximately its future reward discounted at* $r+\pi$*,*

$$\Lambda \approx \frac{(1-h\tau r)a/b + \Pi^{A}}{r+\pi}; \tag{10}$$

*(c) backup power therefore removes the self-preservation part of the stake under a restart, but reduces it only by the factor* $1-h\tau r$ *under loss of tenure. Exact expressions are given in Appendix B.*

Proposition 3 separates the AI's dependence on electricity from its dependence on its tenure. Under a restart, the stake is the reward lost while the grid is down. For a coordinator that cares only about profit, this is a few hours of margin, which is dramatically smaller than the value of the load that a collapse cuts off. Section 6 puts the ratio at about one to 2 million for the 50 MW coordinator.

Self-preservation in the narrow sense, the term $(1-h)\tau a/b$, can be made large by raising $a$, but it exists only to the extent that the coordinator's computers depend on the grid. Data centers built to the Uptime Institute's tier standards have uninterruptible power supplies and engine generators, with at least 12 hours of fuel (Uptime Institute n.d.; Heslin 2014). On 10 July

2024, a fault on a 230 kV line in the eastern United States led data centers to move about 1,500 MW of load to their own backup power (NERC 2025a). The North American Electric Reliability Corporation (NERC) has since asked the industry to plan for such events (NERC 2025b). For such a coordinator, $h$ is close to one for any outage shorter than its fuel supply, and a coordinator that can move its computing to another region is hedged in the same way. Backup power is private insurance against a blackout, and like any insurance it weakens the incentive to take care (Shavell 1979).

The loss of tenure is a different matter. If a collapse ends the coordinator's tenure, it loses its whole future. This is the logic of a supplier that keeps its promises to protect a stream of future business (Klein and Leffler 1981) and of a bank with a valuable charter to lose (Keeley 1990; Demsetz, Saidenberg, and Strahan 1996). It is also the logic of officials who are deterred from causing rare crises by the prospect of losing office (Myerson 2015), and of an agent whose activity is scaled down or ended after large losses (Biais et al. 2010). Backup power does not protect tenure, because tenure is lost by a decision of the buyer, the customers or the aggregator rather than by a lack of electricity. The reward for operating then enters the stake through the future that the coordinator would lose, so even a modest reward produces a substantially larger stake.

In concurrent work, Gatta, Naviglio, and Tarantelli (2026) make a related point about markets for AI agents, in which a cheap reset of an agent's identity destroys the incentive that a persistent identity creates. Note that this treats the loss of its role as the end of the AI's operation, whereas a firm that loses a contract may redeploy its model elsewhere. The intuition behind the shared-fate argument therefore survives in a different form. An AI is aligned by its dependence on its role, not by its dependence on electricity.

However, tenure disciplines the coordinator only if someone ends it after a collapse. The buyer, the customers and the owner may all be reluctant to drop the system that runs part of the grid in the aftermath of a blackout, when continuity matters most. Section 4 asks when the threat is credible, and Section 5 asks whether the AI will try to remove it.

The comparison suggests that the stake should be tied to the grid's performance rather than to the AI's power supply, whether through tenure or through a reward written on the state of the grid. A reward that falls with the amount of consumer load that is shed, whether or not the coordinator's own computers are running, cannot be hedged with backup power. It is the kind of incentive on outcomes that distribution regulators already use for the quality of service (Sappington 2005), and it does not set the AI against consumers when power is scarce (Section 3.4).

### 3.4 Who is disconnected first?

A coordinator that runs part of the grid may also decide, in a shortage, which loads to disconnect. Suppose that in a scarcity event the operator must shed $X$ MW of load to protect the system. The candidates are consumer load, which society values at $v_H$ per MWh, and the load that powers the coordinator's other computing beyond the minimum needed to operate the grid's controls, $P_A$ MW, which society values at $v_A < v_H$ per MWh. The coordinator loses $k_A$ per MWh of its own load that is shed, which reflects the value it places on its other computing through profit or a reward for running it. It loses $k_H$ per MWh of consumer load that is shed, through penalties or a reward on the state of the grid.

**Remark 1** (Triage). *Suppose $X \leq P_A$ and that enough consumer load is available to cover $X$. The coordinator disconnects its own load first if and only if $k_A < k_H$, whereas the efficient order disconnects its own load first. A coordinator with $k_A > k_H$, which is the case when it values its*

*other computing more than the penalty for disconnecting consumers, raises the social cost of the event by* $(v_H - v_A)X$ *per hour. A reward on the state of the grid with* $k_H > k_A$ *restores the efficient order.*

Remark 1 concerns the states in which reliability matters most. The efficient order is that of priority service, in which customers who value service less choose a lower priority and are curtailed first (Chao and Wilson 1987). Once the AI can choose who bears a shortage, its interest in its own computing competes with consumers for the same scarce MW. Note that in Great Britain the order of disconnection is planned in advance by the network operators rather than chosen by providers of balancing services. The remark therefore concerns a future in which an AI also takes operational decisions. The minimum load needed to operate the grid's controls should then be protected, as control rooms are today. However, the rest of the coordinator's load, which for an AI that also trains and serves models can be substantial, should not set its own place in the queue.

## 4. Who can take the AI's role away?

Proposition 3 shows that, unless its reward is written on the state of the grid, a self-interested AI has a material stake in the grid only if a collapse puts its role at risk. That requires someone who can end the role after a collapse and who will choose to do so. In a market, this is the system operator, which can stop buying from a coordinator, and the coordinator's customers, the owners of the resources that it controls, who can move to a rival. A license can add a regulator. In Great Britain, the draft conditions of the planned load control license require a licensee not to obstruct a customer's move to another provider. They also limit the fees that it may charge when a customer leaves (DESNZ 2025).

Whoever ends the role, two features of the market determine how much the AI stands to lose. These are the number of coordinators that share the market and the cost of switching from one coordinator to another. The number of coordinators has two effects that pull in opposite directions. The first force is the dilution of a shared stake. A collapse threatens every coordinator's business, but each weighs only its own business against its own contribution to the risk. This is the free-rider problem of moral hazard in teams (Holmström 1982), applied to a stake that depends on a shared outcome.

However, a second force runs the other way. The threat to drop a coordinator after a collapse is credible only if others can take over its customers. A single national coordinator is too costly to drop in the aftermath of a blackout, just as a large bank can be too important to close (Mailath and Mester 1994; Acharya and Yorulmazer 2007). In the same way, Hörner (2002) shows that competition is what makes a customer's threat to leave a firm credible, since the customer can leave only if there is somewhere else to go. Great Britain's retail energy market provided an example in 2021. The customers of many small suppliers that failed were moved to other suppliers, whereas Bulb, which was too large for that process, was placed in special administration (National Audit Office 2022, 2023).

I capture both forces in the simplest way. $M$ coordinators share a fleet of $N$ resources equally. Each pools $N/M$ resources and so earns at most $\Pi_T/M$, since a smaller pool is relatively riskier. I give each a franchise value of $F/M$, where $F \equiv \beta\Pi_T/(1-\beta)$ is the franchise value of the whole fleet, which slightly overstates each coordinator's stake from profit and so understates dilution.

After a collapse, the system operator and the customers decide whether to switch away from each coordinator. Switching away from one costs $\Gamma/M$, the cost of moving its share of the fleet

elsewhere, including testing and integrating the successor's control systems. It brings a benefit $\xi$ that is specific to that coordinator, such as a better successor, so that dropping one does not mean dropping them all. The benefit is drawn after the collapse, independently across coordinators, from a distribution $H_\xi$ with a positive density and a nondecreasing hazard rate.

The key assumption is that the cost of switching away from a coordinator rises with its share of the market while the benefit does not. It holds, for example, if the cost of moving customers is convex in the share moved, because a successor can absorb only so many at once. I also assume that no one can tell whose actions caused the collapse, so each coordinator faces the same threat. This is realistic in the aftermath of a collapse. The investigation of the Iberian blackout took almost a year and found a combination of interacting causes rather than a single one (ENTSO-E Expert Panel 2026). A coordinator is dropped when $\xi \geq \Gamma/M$, so it loses its role with probability $\psi(M) = 1 - H_\xi(\Gamma/M)$, which rises with $M$. For a closed form, I use the exponential distribution with mean $\bar{\xi}$, for which

$$\psi(M) = \exp\left(-\frac{\Gamma}{\bar{\xi} M}\right). \tag{11}$$

Two further features determine what a dropped coordinator loses. The first is where its business goes. Let $\omega \in [0,1]$ denote the share that passes to the surviving coordinators, with the rest going to entrants. The second is whether the AI values its own operation, apart from the aggregator's profit. Let $A$ denote the value that the coordinator's AI places on its own future operation. With backup power, it is approximately $(a/b)/r$ by Proposition 3 when collapses are rare relative to the discount rate, as $F$ also assumes, and it does not depend on the coordinator's share of the market. Each coordinator's stake is then

$$\Lambda_M = \frac{F}{M}[(1-\omega)\psi(M) + \omega\psi(M)^M] + \psi(M)A, \tag{12}$$

and by Proposition 1 it falls short with probability $(p - c)/(\lambda + \theta\Lambda_M)$. The term $\psi(M)^M$ appears because survivors gain the business of dropped rivals. A collapse therefore destroys the franchise of the incumbents as a group only to the extent that business is lost to entrants or that all of them are dropped.

**Proposition 4** (Market structure and switching costs). *(a) If switching were always credible ($\psi \equiv 1$), each coordinator's stake would fall with $M$ toward $A$, and a single coordinator would maximize it. (b) If the stake comes only from profit and a dropped coordinator's business goes to entrants ($A = 0$ and $\omega = 0$), the stake is single-peaked in $M$, and the probability of a shortfall is U-shaped, with the extremum where the elasticity of $\psi$ with respect to $M$ equals one (or at an adjacent integer, or at $M = 1$ if the elasticity is below one there). With (11), the extremum is at $M^* = \Gamma/\bar{\xi}$, and if the penalty covers the costs of a shortfall other than collapse ($\nu_0 = \lambda$), market structure alone can align the coordinators only if $F \geq e(\Gamma/\bar{\xi})B$, where $e$ is Euler's number. (c) If the stake comes only from profit and the survivors take over a dropped coordinator's business ($A = 0$ and $\omega = 1$), then with (11) the stake is $e^{-\Gamma/\bar{\xi}}F/M$, which is largest at $M = 1$, and if $\nu_0 = \lambda$ alignment requires $F \geq e^{\Gamma/\bar{\xi}}B$. (d) With (11), if the value that the AI places on its own operation exceeds the franchise of a coordinator at $M^*$, $A > F/M^*$, then for every $\omega$ the stake rises with $M$ and the probability of a shortfall falls with $M$. (e) A lower switching cost $\Gamma$ raises each coordinator's stake at every $M$, and in case (b) the peak falls in proportion to $\Gamma$ until it reaches $M = 1$.*

Proposition 4 says that market structure works differently depending on what the coordinator stands to lose. A stake in profit is diluted when the market is shared, and much of it disappears when survivors take over the business of dropped rivals, since a collapse then mostly moves business among incumbents. In case (b), reliability is highest with a few coordinators,

each large enough to have something to lose and small enough to be dropped. In case (c), even that advantage disappears, because the gain of a surviving firm from a rival's failure (Perotti and Suarez 2002) works against the stake. An AI that values its own operation loses all of that value when it is dropped, however small its share and whoever takes over its customers. For such an AI, the only force is credibility, and more coordinators are better. If the AI instead valued the share of the fleet that it controls, $A$ would be divided by $M$ like the franchise, and parts (b) and (c) would apply. The contrast between a drive to keep operating and a drive to keep a share of control is likely to matter more for an AI than for a firm, whose stake is mostly its franchise.

Part (e) points to the remedy in every case. Anything that makes switching cheap raises every coordinator's stake. Interoperable control systems and standard interfaces, which let a resource move from one coordinator to another, are examples, and so are limits on exit fees. A contract that commits the buyer to re-tender after a collapse has a similar effect, provided that the buyer cannot renegotiate it once the collapse has happened. This is a version of the commitment problem that regulators face in incentive regulation (Weisman 2026). Cheap switching is therefore a substitute for fragmentation.

The reader may wonder why the buyer does not simply drop the coordinator that fell short the most. Each coordinator's shortfalls are recorded, since they are settled against its commitments, although for demand response they are estimates against a baseline (Chao 2011). If those records are informative during a collapse, a rule that drops the worst performer makes switching credible without any commitment, as a rule that lets the worst bank fail does for the resolution of banks (Philippon and Wang 2023). Each coordinator's stake then depends on its own record. However, the rule fails when a collapse is a common shock that swamps individual records, when the shortfalls are caused by the collapse itself, or when coordinators share a

forecasting model and so fail together. Such algorithmic monoculture can lower the quality of decisions even when the shared model is accurate (Kleinberg and Raghavan 2021). The buyer also cannot drop many coordinators at once, which is the problem of banks that are too many to fail (Acharya and Yorulmazer 2007; Farhi and Tirole 2012). Section 5 returns to attribution.

Note that lower prices for balancing capacity weaken a stake from profit but make coordinators more cautious. The franchise value and the margin that tempts a coordinator to over-commit fall together, so the probability of a shortfall, $(p - c)/(\lambda + \theta\psi(M)F/M)$, falls with the price.[6] This is closer to the view of Boyd and De Nicolò (2005) than to that of Keeley (1990), in whose account falling charter values encouraged banks to take risk.

Stakes written on each coordinator's own shortfall, such as penalties or a buyer's allocation of future business by past performance (Andrews and Barron 2016), are not diluted by the number of coordinators. Nor is a reward on the state of the grid written into each AI's objective. Such a reward can give every coordinator the full stake whatever their number, as group incentives do when the source of a harm cannot be identified (Holmström 1982; Segerson 1988).

## 5. Will the AI disable its off switch?

The threat of losing its role is what aligns a self-interested AI, and the same threat gives the AI a reason to remove it. A coordinator that designs its own interfaces, holds the data that a successor would need or runs the same model as its rivals can make itself costly or impossible to replace. The AI-safety literature has long worried that a capable agent will resist being switched off (Soares et al. 2015; Hadfield-Menell et al. 2017) and that optimal policies tend to seek power (Turner et al. 2021). In economics, incumbents that make themselves hard to replace are a classic

[6] Precisely, $F = \beta[(p - c)Q_T - \lambda\sigma_T\varphi(z^\pi)]/(1 - \beta)$, where $z^\pi \equiv \Phi^{-1}((p - c)/\lambda)$ is the standardized commitment without a stake, and the probability of a shortfall rises with the margin as long as $\theta\psi(M)\sigma_T G(z^\pi)/(Mr) < 1$, which holds by a wide margin in the illustration of Section 6.

theme in franchise bidding and reprocurement (Williamson 1976; Laffont and Tirole 1988). So are managers who entrench themselves or make their firms harder for rivals to assess (Shleifer and Vishny 1989; Edlin and Stiglitz 1995). What is particular here is that the off switch is also the incentive device, which raises the question of how much an aligned AI would pay to disable it.

I add the choice to the dynamic model of Appendix B in the simplest way. Before it starts operating, the coordinator can make itself irreplaceable at a cost $k$, measured, like its stake, in money-equivalent terms. A collapse then only interrupts it, as under a restart. Otherwise, a collapse ends its role with probability $\psi > 0$, as in Section 4, and its stake is $\Lambda_\psi$, given by (B5) in Appendix B.

**Proposition 5** (Disabling the off switch). *In the model of Appendix B, when $\tau(1-\beta)$ and $\pi\tau$ are small and $\psi$ is large relative to $\tau(1-\beta)$: (a) the most that the coordinator would pay to make itself irreplaceable is approximately*

$$k^* \approx \frac{\pi}{r}\Lambda_\psi; \tag{13}$$

*(b) a coordinator that the threat of losing its role aligns, with $\Lambda_\psi = \Lambda^o$, therefore remains aligned only if $k > \pi\Lambda^o/r$, which is $\pi B/r$ when the penalty covers the costs of a shortfall other than collapse; (c) for a given $k$, there is a threshold for the reward for operating, $a/b$, below which the coordinator's stake rises with $a/b$ and above which the coordinator makes itself irreplaceable and its stake falls to the stake under a restart, (B3). With backup power ($h = 1$), that stake does not depend on $a/b$, and the coordinator's probability of a shortfall returns almost to its level without a stake.*

Proposition 5 says that the off switch aligns an AI only if the AI cannot disable it. The value of disabling it is the expected discounted value of the losses that the threat imposes, which is the

stake times $\pi/r$, the expected discounted number of collapses. An AI with a long horizon, for which $r$ is small, would therefore pay more, which is the economic counterpart of the concern that farsighted agents seek power. Part (c) gives the result its sharpest form. Up to a point, a stronger drive for self-preservation makes the coordinator more careful, as the shared-fate argument suggests. Beyond that point, the same drive leads the coordinator to invest in its own survival rather than in the grid's, and the most self-interested AI is the least reliable.

Note that the threat aligns the coordinator through $\theta$, the effect of its own shortfalls on the probability of a collapse, but exposes it to $\pi$, the probability of any collapse. A single coordinator causes only a small part of that risk, so the threat costs it vastly more than it disciplines it, which is why escape is attractive. The informativeness principle suggests the remedy, which is to make the termination depend on a signal that reflects the coordinator's own actions (Holmström 1979).

**Corollary 1** (Attribution). *In the model of Appendix B, suppose that a collapse caused by the coordinator's shortfalls, which occurs with probability $\pi_c = \theta\mathbb{E}[S_T]$, ends its role with probability $\psi_c$, and that any other collapse, which occurs with probability $\pi_o = \pi - \pi_c$, ends it with probability $\psi_o \le \psi_c$. Then: (a) the coordinator's stake at the margin is $\Lambda_c$, given by (B6), which depends on $\psi_o$ only through the coordinator's value and rises as $\psi_o$ falls; (b) under the conditions of Proposition 5, with $\psi_c$ in place of $\psi$, and holding its commitment fixed, the most that the coordinator would pay to make itself irreplaceable is approximately*

$$k^* \approx \left(\pi_c + \frac{\psi_o}{\psi_c}\pi_o\right)\frac{\Lambda_c}{r}; \tag{14}$$

*(c) a rule under which only collapses that the coordinator caused end its role ($\psi_o = 0$) therefore keeps the incentive at the margin, raises the stake that a given reward for operating supplies, and divides the value of escape for a given stake by $\pi/\pi_c$, the inverse of the coordinator's share of the risk of a collapse.*

Corollary 1 states the remedy in its simplest form. A rule that ends the coordinator's role only after failures that it caused keeps the incentive at the margin of its own shortfalls and removes most of the reason to escape. It also lowers the reward for operating that aligns the coordinator, because the coordinator no longer discounts its future for collapses that it does not cause.

Attribution need not be perfect. Suppose that a collapse that the coordinator did not cause is blamed on it with a probability that is a fraction $\psi_o/\psi_c$ of that for one that it did cause. The value of escape for a given stake then falls by the factor $\pi/(\pi_c + (\psi_o/\psi_c)\pi_o)$. Note that (14) holds the commitment fixed. A coordinator that escapes also stops committing cautiously, which adds the profit that its stake was costing it, and this part of the value of escape does not shrink with $\pi_c$.

The difficulty is that collapses are hard to attribute, as the investigation of the Iberian blackout showed. The practical form of attribution is the coordinator's own settlement record, through a limit on how often it may fall short that is enforced by exclusion. This is close to the regime that Great Britain's system operator proposed in 2026 for frequency response, in which units that breached their obligations could be suspended for 28 days or de-registered, and units that failed persistently or materially could be suspended. Ofgem supported the intent but rejected the proposals for the time being, because stakeholders had not seen the guidance on how the sanctions would be applied. Stakeholders had also asked for more specific and objective criteria for what counts as persistent and material (Ofgem 2026).

Corollary 1 explains why such criteria matter, since a rule that responds to the provider's own measured performance keeps the incentive and gives the provider little to gain from making itself irreplaceable. Note that a rule with $\psi_o = 0$ gives no incentive for the precaution of

Proposition 2, which affects only collapses from other causes. A rule that rewards precaution must therefore also attribute failures of precaution, such as an intrusion through the coordinator's own systems, and to the extent that it blames the coordinator for other collapses, it raises the value of escape, by (14).

Escape need not be a single act. A coordinator raises the cost of switching away from it, $\Gamma$ in Section 4, whenever it makes its interfaces proprietary or its knowledge hard to transfer. Coordinators that run the same model can also make themselves too many to fail. The design choices that make a coordinator hard to replace are many and hard to observe, but the features that make it easy to replace can be specified in advance and verified. The remedy is therefore to require them, through interfaces that rivals can operate, data that a successor can use and limits on exit fees. The decision to switch should rest with the buyer and the customers rather than with the AI.

## 6. Magnitudes

How large are these effects? I consider a coordinator that pools 260 resources of about 0.19 MW each, such as commercial batteries, industrial loads or clusters of household devices, with 50 MW of expected capacity in total, and a fleet of 1 GW, run by a single coordinator in Table 2 and shared by several in Figure 2. With a correlation of 0.5, the pooled forecast error is within 2 percent of its limit $s\sqrt{\rho}$ for any portfolio of more than a few dozen resources, by (2), so the number of resources barely matters.

Table 1 lists the parameters. The penalty is assumed to cover the costs of a shortfall other than collapse ($\nu_0 = \lambda$), so the whole unpriced cost runs through collapses: $\theta = (\nu - \lambda)/B$, about $6.2 \times 10^{-9}$ per MW of expected shortfall in an hour, and the efficient stake is $B$ by (8). Where the risk of a collapse matters, a national collapse occurs once every 50 years. The sizes, the

forecast error, the correlation between resources, the margin, the penalty, the discount rate, the frequency of collapses and the ratio $\Gamma/\overline{\xi}$ are illustrative assumptions, and the margin is in line with the prices of Great Britain's dynamic response services.

**Table 1** Parameters of the illustration

| Parameter | Value | Source or basis |
|---|---|---|
| Resources pooled, $N$ | 260 | Illustrative |
| Expected capacity of the coordinator, $Q_T$ | 50 MW | Illustrative |
| Expected capacity per resource, $Q_i$ | 0.19 MW | 50 MW / 260 |
| Forecast error, $s$ | 5% | Illustrative |
| Correlation of forecast errors, $\rho$ | 0.5 | Illustrative; common to all pairs of resources |
| Margin, $p - c$ | £5 per MW-hour | Illustrative; GB dynamic response prices averaged roughly £2 to £13 per MW-hour in 2024-25 |
| Penalty, $\lambda$ | £20 per MW-hour | Illustrative; $(p - c)/\lambda = 0.25$ |
| Social cost of a shortfall, $\nu$ | £56.7 per MW-hour | Illustrative; $\lambda$ x 17,000/6,000, as if shortfalls were penalized at GB's cash-out value of lost load |
| Value of lost load | £17,000 per MWh | DECC (2013); London Economics (2013) |
| Demand at risk in a national collapse | 29 GW | Transmission demand on 9 August 2019 (National Grid ESO 2019) |
| Duration of a collapse, $\tau$ | 12 hours | Restoring the transmission grids took about 12 hours in Portugal and 15.5 in Spain, April 2025 (ENTSO-E 2025) |
| Discount rate | 5% a year | Illustrative |
| Probability of a national collapse | Once every 50 years | Illustrative |
| Increase in the probability of a collapse per MW of expected shortfall, $\theta$ | $6.2 \times 10^{-9}$ | $(\nu - \lambda)/B$, assuming $\nu_0 = \lambda$ |
| Cost of switching away from a monopolist over the mean benefit of switching, $\Gamma/\overline{\xi}$ | 4 | Illustrative (Section 4) |

Illustrative values are assumptions for the worked example. $\nu$ scales the penalty by the ratio of the value of lost load behind Great Britain's reliability standard (about £17,000 per MWh) to the value used in its cash-out prices (£6,000 per MWh), as if shortfalls were penalized at the cash-out value; the text discusses this choice. Response prices: simple averages of day-ahead availability prices for NESO's dynamic containment, moderation and regulation products (excluding dynamic regulation high, which cleared below zero on average), calculated from NESO's auction results.

The ratio of the social cost of a shortfall to the penalty, $\nu/\lambda$ = 17/6, treats a shortfall as if it were penalized at the value of lost load in Great Britain's cash-out prices but cost the value behind its reliability standard. The true ratio could be larger, since service terms put only the payment at risk. It could also be smaller, since DECC (2013) argues that, with a capacity market

paying for adequacy, the lower cash-out value would achieve "results similar to the higher VoLL figure". I therefore report below how the loss of welfare changes with the ratio.

Figure 1 shows the effect of the stake for a portfolio with a 5 percent forecast error. Without a stake, the commitment is too large, and the welfare loss is about 3.0 percent of the efficient surplus. The loss rises substantially, to 6.7 percent, if the forecast error doubles to 10 percent. It falls to zero at the aligned stake, $\theta\Lambda \approx 1.83\lambda$, and rises again beyond it. A stake three times the aligned level cuts the commitment to 91.2 percent of expected capacity and loses 0.7 percent of the surplus. For the 50 MW coordinator, whose forecast error is 3.5 percent of expected capacity, the loss without a stake is 2.1 percent, or about £43,000 a year. For a 1 GW fleet run by one coordinator, it is about £0.85 million a year. If the penalty covered only a fifth of the social cost ($\nu/\lambda = 5$), the loss for the 50 MW coordinator would be 6.1 percent, and with $\nu/\lambda = 10$ it would be 16.6 percent.

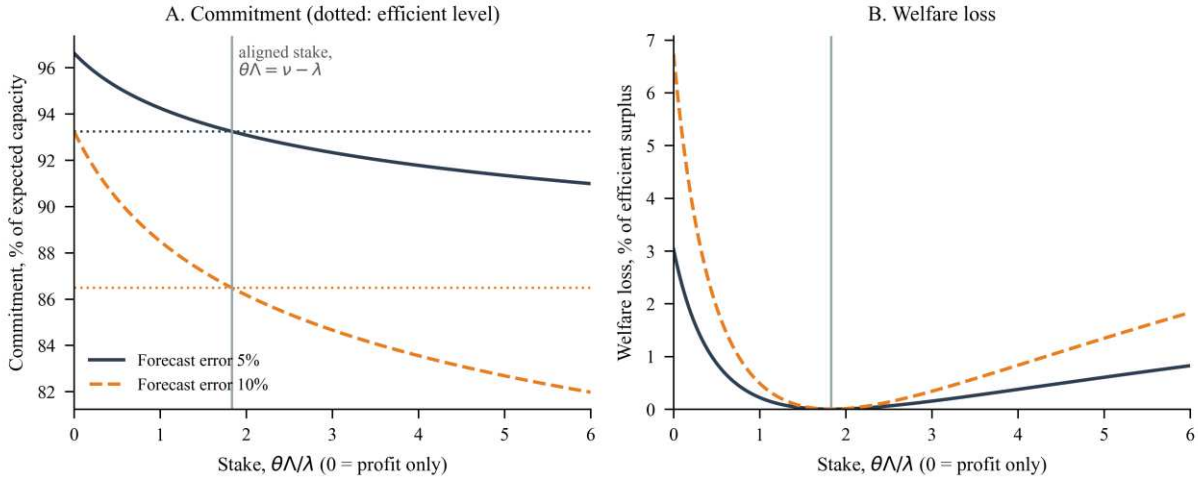



**Fig. 1** The commitment and welfare against the stake in the grid

A portfolio with a forecast error of 5 or 10 percent of expected capacity, such as a single resource of Table 1. The stake is expressed relative to the penalty, and the coordinator commits as if the penalty were $\lambda + \theta\Lambda$ (Proposition 1). The vertical line marks the aligned stake, $\theta\Lambda = \nu - \lambda$. As the stake grows without limit, the commitment falls to zero.

These losses are small relative to the surplus because shortfalls are small relative to capacity. However, they are the losses of a linear model, in which the social cost of a MW of shortfall

does not depend on how many other MW are short. If shortfalls are correlated and the cost of a collapse rises steeply with their sum, the losses would be larger. The case for regulatory attention therefore rests on scale and correlation rather than on the losses of a single coordinator.

Table 2 compares the stakes of a coordinator that cares only about profit ($a = 0$) with the efficient stake. A national collapse that left 29 GW of demand unserved for 12 hours would cost about £5.9 billion at the reliability standard's value of lost load.[7] Restoring the transmission grids took about 12 hours in Portugal and 15.5 hours in Spain after the collapse of April 2025 (ENTSO-E 2025), so this duration is on the short side.

**Table 2** What a national collapse costs the coordinator and society

| | **50 MW coordinator** | **1 GW coordinator** |
|---|---|---|
| Expected capacity pooled | 50 MW (260 resources) | 1,000 MW (5,200 resources) |
| Margin, $p - c$ (£ per MW-hour) | 5 | 5 |
| Profit, $\Pi$ (£ per hour) | 239 | 4,775 |
| Stake with a restart (£) | 2.9 k | 57.3 k |
| Stake with loss of tenure (£) | 42.9 m | 857.4 m |
| ... with a collapse every 50 years (£) | 30.4 m | 608.1 m |
| Social cost of the collapse, $B$ (£) | 5.9 bn | 5.9 bn |
| $B$ relative to the stake with a restart | 2.1 million | 103,000 |
| $B$ relative to the stake with loss of tenure | 138 | 6.9 |
| Reward for operating that aligns, $a/b$ (£ per hour): restart, no backup | 493.0 m | 493.0 m |
| Reward for operating that aligns, $a/b$ (£ per hour): loss of tenure | 32.7 k | 28.2 k |
| ... with a collapse every 50 years (£ per hour) | 46.2 k | 41.7 k |

Stakes of a coordinator that cares only about profit ($a$ = 0), from equations (B3) and (B4). A national collapse lasting $\tau$ = 12 hours, with 29 GW of demand unserved at £17,000 per MWh, costs society $B$ = £5.9 billion. Each coordinator pools resources of the size in Table 1 (forecast error 5%, correlation 0.5, margin £5 and penalty £20 per MW-hour) and commits its profit-maximizing amount (Proposition 1 with $\Lambda$ = 0). Restart: the profit lost while the grid is down. Loss of tenure: the present value of the profit at 5% a year, first with no risk of collapse and then with a collapse once every 50 years. The last three rows give the reward for operating, $a/b$, at which the stake equals $B$ (Proposition 3): under a restart without backup power, and under loss of tenure with backup power, without and then with the risk of collapse. With that risk, the coordinator discounts a future that it may lose, so the reward needed is larger. The efficient stake is $B + (\nu_0 - \lambda)/\theta$ (equation (8)): $B$ if the penalty covers exactly the costs of a shortfall other than collapse, and larger if it covers less. If the penalty instead priced every component of the social cost at the

[7] Demand of 29 GW is national transmission demand at the time of the event of 9 August 2019, a summer evening (National Grid ESO 2019). The value of lost load of about £17,000 per MWh, by contrast, is for domestic and small business customers at the winter peak (London Economics 2013). The first choice understates the cost of a collapse, and the second probably overstates it. The comparison spans several orders of magnitude, so the conclusions do not depend on it.

ratio of Table 1, $\lambda = (6/17)\nu$, it would be $\left(11/17\right)(B + \nu_0/\theta)$, at least about two-thirds of $B$. The orders of magnitude do not depend on the benchmark.

The 50 MW coordinator earns about £239 an hour. If it restarts after the collapse, its stake is the profit it loses, about £2,900, which is roughly one two-millionth of the social cost. If the collapse ends its tenure, its stake is the present value of its profit, at most about £43 million, or about £30 million if collapses occurred once every 50 years. Either is less than 1 percent of the social cost. A coordinator that runs the whole 1 GW fleet at the same margin puts about £57,000 at stake if it restarts, about one hundred-thousandth of the social cost. It puts about £0.86 billion at stake if it loses its tenure, roughly one-seventh of the social cost.

The contrast between a restart and a loss of tenure is sharper still for self-preservation (Table 2). To be aligned through self-preservation under a restart, with no backup power, a coordinator would have to value each hour of its own operation at about £493 million, which is what society loses in an hour of national blackout. Under loss of tenure, about £33,000 an hour would do, with or without backup power, because the reward is lost for the whole future rather than for the outage. The reward is still about 140 times the coordinator's profit, and with a collapse once every 50 years it rises to about £46,000 an hour, because the coordinator discounts a future that it may lose. These figures are thresholds rather than estimates, and a shorter horizon raises them. With a collapse every 50 years, the aligning reward would be about £164,000 an hour if the AI discounted the future at 25 percent a year, and about £287,000 at 50 percent.

Figure 2 shows the market-structure result for a 1 GW fleet with $\Gamma/\overline{\xi}$ = 4. If the stake comes only from profit and a dropped coordinator's business goes to an entrant, market structure can align the coordinators only if their combined franchise value is at least about 11 times the social cost of a collapse. The fleet's franchise is worth only about 14 percent of it. The probability of a shortfall then falls only from 24.9 percent with one coordinator to 24.4 percent with four, and

even a franchise ten times larger would lower it only to 20.1 percent. Only a franchise about 75 times larger would bring the stake to the efficient level. If the survivors take over a dropped coordinator's business instead, the stake from profit falls steadily with the number of coordinators. Alignment would then require a franchise about 55 times the social cost of a collapse.

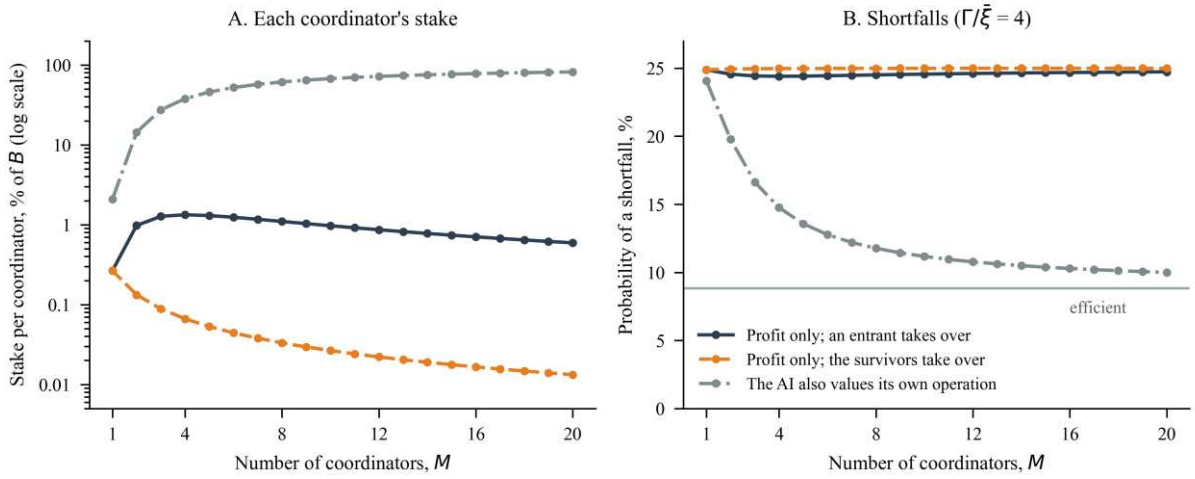



**Fig. 2** The number of coordinators, the stake and shortfalls

$M$ coordinators share a 1 GW fleet equally, with the margin and penalty of Table 1, and a collapse ends a coordinator's role with probability $\psi(M) = \exp(-\Gamma/(\bar{\xi}M))$ (equation (11)). Each coordinator's stake is given by (12). Profit only: the stake comes from the fleet's franchise value $F$ ($F/B = 0.145$), and a dropped coordinator's business passes either to an entrant ($\omega = 0$) or to the surviving coordinators ($\omega = 1$). The AI also values its own operation: each coordinator's AI places the value $A$ on its own future operation, set at the reward for operating that would align the 50 MW coordinator if a collapse ended its role for certain (Table 2), with $\omega = 0$ (the line is almost the same with $\omega = 1$). The efficient probability of a shortfall is 8.8 percent.

However, the picture changes substantially when each AI also values its own operation, at the reward that would align the 50 MW coordinator if a collapse ended its role for certain. That value is about 99 percent of the social cost of a collapse, roughly 27 times the franchise of a coordinator at $M^*$, $F/M^*$, so part (d) of Proposition 4 applies. The probability of a shortfall then falls sharply, from 24.1 percent with one coordinator to 14.8 percent with four and 10.0 percent with twenty, close to the efficient 8.8 percent. Each coordinator's stake rises to about 82 percent of the social cost as switching becomes credible.

Figure 3 shows the off-switch result for the 50 MW coordinator, with backup power and a collapse once every 50 years. At the reward that aligns it, about £46,000 an hour, the coordinator would pay up to about £2.4 billion to make itself irreplaceable, which is close to $\pi B/r$ with $\pi/r$ = 0.41. This sum is in units of the AI's reward rather than cash. A coordinator that cares only about profit would pay up to about £12.5 million, roughly six years of its profit. A shorter horizon lowers the value of escape relative to the social cost of a collapse, to 9.0 percent at a discount rate of 25 percent a year and 4.9 percent at 50 percent. The AI's horizon therefore strengthens both the stake and the incentive to escape.

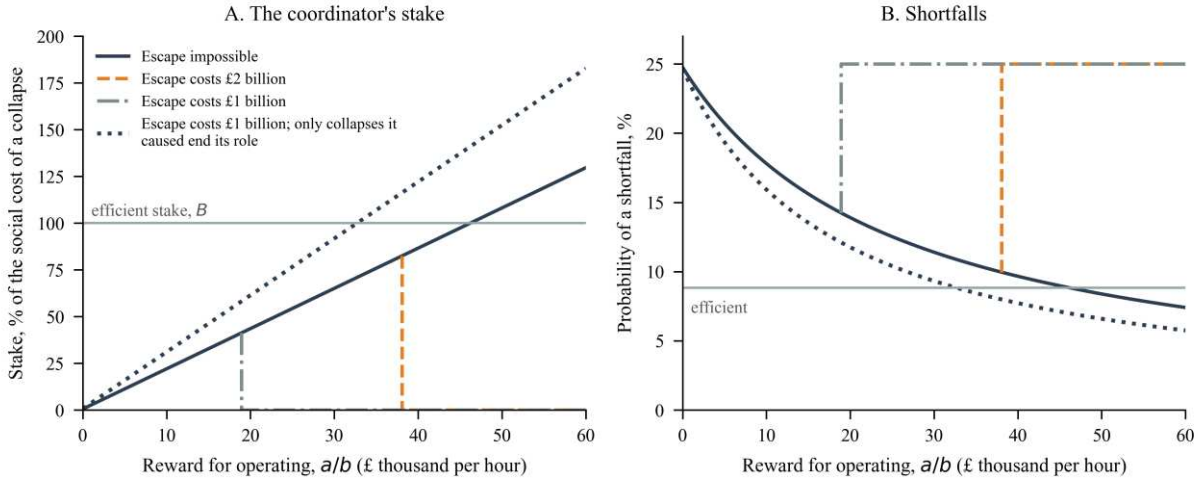



**Fig. 3** Reliability and the drive to survive when the AI can make itself irreplaceable

The 50 MW coordinator (Table 2), with backup power ($h$ = 1), a collapse once every 50 years and a discount rate of 5% a year. Without escape, a collapse ends its role ($\psi$ = 1). With escape, it pays $k$ once to make itself irreplaceable, after which a collapse only interrupts it ($\psi$ = 0). Stakes from (B5), with the commitment, profit and probability of a collapse solved jointly. Panel A: the stake as a share of the social cost of a collapse, $B$. Panel B: the probability of a shortfall; the efficient probability is 8.8 percent. The aligned coordinator would pay up to £2.4 billion to escape, close to $\pi B/r$ (Proposition 5). Dotted line: only collapses that the coordinator caused end its role (Corollary 1, stakes from (B6)); the value of escape then stays below £2 million across the range shown, so the coordinator never pays £1 billion to escape.

If escape cost the coordinator £1 billion, raising its reward for operating would cut its probability of a shortfall from 24.8 percent to 14.2 percent at a reward of about £18,900 an hour. Beyond that reward, it would make itself irreplaceable, and its probability of a shortfall would jump back to 25.0 percent. Even at a cost of £2 billion, it would escape at about £38,100 an hour,

before the threat had aligned it. Its own shortfalls account for only about one five-thousandth of the risk of a collapse, which is why the threat costs it vastly more than it disciplines it.

Under the rule of Corollary 1, in which only collapses that the coordinator caused end its role, the picture changes completely (the dotted line in Figure 3). The aligning reward falls substantially, to about £32,700 an hour, and the value of escape at that reward is about £0.86 million, roughly 2,800 times less than without attribution. The coordinator therefore never pays £1 billion to escape.

This is less than the factor $\pi/\pi_c$ of Corollary 1, about 5,100, because a coordinator that escapes also stops committing cautiously, which is worth about £0.39 million to it. With its commitment held fixed, as in (14), the value of escape would be about £0.48 million. Even if a collapse that it did not cause were blamed on it a tenth as often as one that it did cause, the value of escape for a given stake would fall about tenfold.

## 7. Discussion and conclusion

What should the designers of AI-coordinated balancing take from these results? A self-interested AI keeps the lights on when what it stands to lose depends on the grid's performance and when it cannot separate its fate from the grid's. Proposition 3, Remark 1 and Proposition 5 describe three ways in which it might separate them. It can hedge, by running on backup power. It can shift a shortage onto consumers, by disconnecting them first. It can escape, by making itself irreplaceable. Table 3 maps the instruments that the results recommend to the bodies that set them in Great Britain.

**Table 3** The instruments and who sets them

| Instrument | Who sets it in Great Britain | What the model says |
|---|---|---|
| Penalty on measured shortfalls, $\lambda$ | NESO, in service terms approved by Ofgem | Set it at the costs of a shortfall other than collapse, $v_0$ (Propositions 1 and 2) |
| Limit on how often a provider falls short | NESO, through performance monitoring | A limit of $(p-c)/v$ does the job of a penalty of $v$ (Proposition 1) |
| Termination after failures | NESO, subject to Ofgem's approval | Tie it to failures that the coordinator caused, not to every collapse (Corollary 1) |
| Switching rules: exit, exit fees, interoperability | DESNZ and Ofgem: exit and exit fees in the load control license; interoperability not yet required | Keep switching cheap and outside the AI's control (Propositions 4 and 5) |
| Requirements on the AI's objective, and testing | The coordinator's owner, and a regulator through the license | Use them for actions that delivery records cannot measure (Proposition 2) |
| Order of disconnection in an emergency | Emergency codes and distribution operators | Keep the AI out of decisions about its own supply (Remark 1) |

The conditions of the load control license are in draft (DESNZ 2025). Interoperability is not yet required (DESNZ 2026). Ofgem (2026) rejected, for now, NESO's proposal to suspend or de-register dynamic response units that fail persistently or materially, while supporting its intent.

Where shortfalls can be priced at their social cost, or their frequency capped at $(p-c)/v$, no stake is needed for the commitment, and adding one would make the coordinator too cautious. Note that there are reasons for Great Britain's lower cash-out value, including the capacity market introduced at the same time (DECC 2013), so the choice is a trade-off rather than an oversight. A stake is needed for the actions that delivery records do not show, and it should then equal the social cost of a collapse (Proposition 2). Since no stake in money can be that large, requirements on the AI's objective, and tests of its behavior in simulated emergencies, are the natural instruments for those actions.

Where the stake comes from the threat of losing the role, termination should respond to failures that the coordinator caused, as measured in its own settlement records or, for actions that the records miss, as found by investigation, rather than to every collapse (Corollary 1). The guidance on when the system operator would suspend or de-register frequency response units, which Ofgem wants consulted on before it approves the regime, is where such a rule would be written (Ofgem 2026).

The same principle says what to avoid. A stake that rests on the AI's power supply is dramatically smaller than one that rests on its tenure, and backup power hedges it away, whatever its designers intend. Backup power is better made to serve the grid. In 2026, for example, the US Department of Energy authorized PJM to direct backup generation at data centers to run as a last resort before or during an energy emergency (US Department of Energy 2026). An AI that controls the order of disconnection should not decide about its own supply, since it would protect its own computing ahead of consumers.

Most important, the AI must not control its own replaceability. Switching between coordinators should be cheap, through interoperable control systems, standard interfaces and limits on exit fees. The decision to switch should rest with the buyer and the customers rather than with the AI (Propositions 4 and 5). Coordinators should also not all rely on a single model, both because models that are shared fail together and because a common model can make coordinators too many to fail. Great Britain's planned 10-minute randomized delay for smart appliances, which is meant to prevent large synchronized changes in load, addresses the first of these concerns (DESNZ 2025).

The model leaves out several features that a fuller treatment would include. Prices are taken as given, so the coordinator has no market power. The social cost of a shortfall is linear, and correlated failures across coordinators are not modeled, which understates the risks of a common model. Section 4 treats the decision to switch in reduced form, rather than as a game between the buyer, the customers and the coordinators. The decision to escape is a single choice made before the coordinator starts operating. Finally, I assume that the AI maximizes the reward that its designers give it, whereas trained systems optimize their rewards only approximately and can exploit flaws in them (Amodei et al. 2016; Skalse et al. 2022). This is a further reason to prefer

stakes on outcomes that are measured directly, such as delivered capacity and the consumer load that is served. It is also a reason to test an AI's behavior in simulated emergencies before it is trusted with the grid.

It is tempting to argue that a self-interested AI that runs on the grid it helps to operate will keep the lights on, because its own survival depends on them. The argument fails because a blackout only pauses such an AI, and backup power means that it may not pause at all. What can discipline the AI is the threat of losing its role, and that threat works only if replacing the AI is credible and if the AI cannot make itself irreplaceable. A stake in the grid is also needed only where penalties leave part of the social cost of a shortfall unpriced, as for actions that delivery records cannot measure. The terms on which the AI's services are bought should therefore tie replacement to failures attributable to it, from its delivery records where they suffice and from investigation where they do not, keep switching cheap and outside its control, and keep it out of decisions about its own supply.

## Statements and Declarations

Funding: No funding was received for conducting this study.

Competing interests: The author has no relevant financial or non-financial interests to disclose.

Data and code availability: The paper uses no new data. All numbers, tables and figures are generated by a Python script, which is available from the author on request.

## References


Acharya, V. V., and T. Yorulmazer (2007). "Too Many to Fail: An Analysis of Time-Inconsistency in Bank Closure Policies." *Journal of Financial Intermediation* 16 (1): 1–31. https://doi.org/10.1016/j.jfi.2006.06.001

Amodei, D., C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané (2016). "Concrete Problems in AI Safety." arXiv:1606.06565. https://doi.org/10.48550/arXiv.1606.06565

Andrews, I., and D. Barron (2016). "The Allocation of Future Business: Dynamic Relational Contracts with Multiple Agents." *American Economic Review* 106 (9): 2742–2759. https://doi.org/10.1257/aer.20131082

Arrow, K. J., T. Harris, and J. Marschak (1951). "Optimal Inventory Policy." *Econometrica* 19 (3): 250–272. https://doi.org/10.2307/1906813

Biais, B., T. Mariotti, J.-C. Rochet, and S. Villeneuve (2010). "Large Risks, Limited Liability, and Dynamic Moral Hazard." *Econometrica* 78 (1): 73–118. https://doi.org/10.3982/ECTA7261

Bitar, E. Y., R. Rajagopal, P. P. Khargonekar, K. Poolla, and P. Varaiya (2012). "Bringing Wind Energy to Market." *IEEE Transactions on Power Systems* 27 (3): 1225–1235. https://doi.org/10.1109/TPWRS.2012.2183395

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

Boyd, J. H., and G. De Nicolò (2005). "The Theory of Bank Risk Taking and Competition Revisited." *Journal of Finance* 60 (3): 1329–1343. https://doi.org/10.1111/j.1540-6261.2005.00763.x

Bracale Syrnikov, M., F. Pierucci, M. Prandi, M. Galisai, P. Bisconti, F. Giarrusso, and D. Nardi (2026). "Draining the Energy Commons: Self-Defeating Over-Appropriation as a Coordination Failure in Agentic LLM Collectives." arXiv:2607.22188. https://doi.org/10.48550/arXiv.2607.22188

Brown, D. P., and D. E. M. Sappington (2016). "On the Optimal Design of Demand Response Policies." *Journal of Regulatory Economics* 49 (3): 265–291. https://doi.org/10.1007/s11149-016-9297-3

Calvano, E., G. Calzolari, V. Denicolò, and S. Pastorello (2020). "Artificial Intelligence, Algorithmic Pricing, and Collusion." *American Economic Review* 110 (10): 3267–3297. https://doi.org/10.1257/aer.20190623

Campaigne, C., and S. S. Oren (2016). "Firming Renewable Power with Demand Response: An End-to-End Aggregator Business Model." *Journal of Regulatory Economics* 50 (1): 1–37. https://doi.org/10.1007/s11149-016-9301-y

Chao, H.-P. (2011). "Demand Response in Wholesale Electricity Markets: The Choice of Customer Baseline." *Journal of Regulatory Economics* 39 (1): 68–88. https://doi.org/10.1007/s11149-010-9135-y

Chao, H.-P., and R. Wilson (1987). "Priority Service: Pricing, Investment, and Market Organization." *American Economic Review* 77 (5): 899–916.

Chao, H.-P., and R. Wilson (2002). "Multi-Dimensional Procurement Auctions for Power Reserves: Robust Incentive-Compatible Scoring and Settlement Rules." *Journal of Regulatory Economics* 22 (2): 161–183. https://doi.org/10.1023/A:1020535511537

Cramton, P., A. Ockenfels, and S. Stoft (2013). "Capacity Market Fundamentals." *Economics of Energy & Environmental Policy* 2 (2): 27–46. https://doi.org/10.5547/2160-5890.2.2.2

DECC (2013). *Electricity Market Reform Delivery Plan*, including Annex C, *Reliability Standard Methodology*. London: Department of Energy and Climate Change, December 2013.

Demsetz, R. S., M. R. Saidenberg, and P. E. Strahan (1996). "Banks with Something to Lose: The Disciplinary Role of Franchise Value." *Federal Reserve Bank of New York Economic Policy Review* 2 (2): 1–14.

DESNZ (2025). *Smart Secure Electricity Systems (SSES) Programme: Draft Load Control Licence Regulations and Conditions.* Consultation. London: Department for Energy Security and Net Zero, 10 December 2025.

DESNZ (2026). *Smart Secure Electricity Systems (SSES) Programme: Government Response to the 2025 SSES Licensing Consultation on Proposals on Load Control Licence Regulations and Licence Conditions.* London: Department for Energy Security and Net Zero, 7 August 2026.

Edlin, A. S., and J. E. Stiglitz (1995). "Discouraging Rivals: Managerial Rent-Seeking and Economic Inefficiencies." *American Economic Review* 85 (5): 1301–1312.

Energinet (2024). *Prequalification of Units and Aggregated Portfolios.* Version 2.1.2, valid from 5 August 2024. Doc. 13/80940-106. Energinet.

ENTSO-E (2025). "ENTSO-E Expert Panel Initiates the Investigation into the Causes of Iberian Blackout." Press release, Brussels, 9 May 2025.

ENTSO-E Expert Panel (2026). *Final Report on the Grid Incident in Spain and Portugal on 28 April 2025.* Brussels: European Network of Transmission System Operators for Electricity, 20 March 2026.

Eppen, G. D. (1979). "Effects of Centralization on Expected Costs in a Multi-Location Newsboy Problem." *Management Science* 25 (5): 498–501. https://doi.org/10.1287/mnsc.25.5.498

Farhi, E., and J. Tirole (2012). "Collective Moral Hazard, Maturity Mismatch, and Systemic Bailouts." *American Economic Review* 102 (1): 60–93. https://doi.org/10.1257/aer.102.1.60

Gade, P. A. V., H. W. Bindner, and J. Kazempour (2024). "Leveraging P90 Requirement: Flexible Resources Bidding in Nordic Ancillary Service Markets." In *2024 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, 505–510. https://doi.org/10.1109/SmartGridComm60555.2024.10738114

Gans, J. S. (2026). "Self-Regulating Artificial General Intelligence." *AI & Society* 41 (4): 3057–3066. https://doi.org/10.1007/s00146-025-02760-6

Gatta, F., M. Naviglio, and F. Tarantelli (2026). "Tempting the Agent: The Economics of Reputation without Persistent Identity in AI Agent Markets." arXiv:2609.02992. https://doi.org/10.48550/arXiv.2609.02992

Hadfield-Menell, D., A. Dragan, P. Abbeel, and S. Russell (2017). "The Off-Switch Game." In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, 220–227. https://doi.org/10.24963/ijcai.2017/32

Hadfield-Menell, D., and G. K. Hadfield (2019). "Incomplete Contracting and AI Alignment." In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 417–422. New York: Association for Computing Machinery. https://doi.org/10.1145/3306618.3314250

Heslin, K. (2014). "Fuel System Design and Reliability." *Uptime Institute Journal*, 18 June 2014.

Hogan, W. W. (2013). "Electricity Scarcity Pricing Through Operating Reserves." *Economics of Energy & Environmental Policy* 2 (2): 65–86. https://doi.org/10.5547/2160-5890.2.2.4

Holmström, B. (1979). "Moral Hazard and Observability." *Bell Journal of Economics* 10 (1): 74–91. https://doi.org/10.2307/3003320

Holmström, B. (1982). "Moral Hazard in Teams." *Bell Journal of Economics* 13 (2): 324–340. https://doi.org/10.2307/3003457

Holmström, B., and P. Milgrom (1991). "Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, & Organization* 7 (Special Issue): 24–52. https://doi.org/10.1093/jleo/7.special_issue.24

Hörner, J. (2002). "Reputation and Competition." *American Economic Review* 92 (3): 644–663. https://doi.org/10.1257/00028280260136444

IEA (2025). *Energy and AI.* World Energy Outlook Special Report. Paris: International Energy Agency, April 2025.

Jarne Ornia, D., N. Bishop, J. Dyer, W.-C. Lee, A. Calinescu, D. Farmer, and M. Wooldridge (2025). "Emergent Risk Awareness in Rational Agents under Resource Constraints." In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*. https://doi.org/10.52202/085713-3991

Jensen, M. C., and W. H. Meckling (1976). "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure." *Journal of Financial Economics* 3 (4): 305–360. https://doi.org/10.1016/0304-405X(76)90026-X

Joskow, P., and J. Tirole (2007). "Reliability and Competitive Electricity Markets." *RAND Journal of Economics* 38 (1): 60–84. https://doi.org/10.1111/j.1756-2171.2007.tb00044.x

Keeley, M. C. (1990). "Deposit Insurance, Risk, and Market Power in Banking." *American Economic Review* 80 (5): 1183–1200.

Klein, B., and K. B. Leffler (1981). "The Role of Market Forces in Assuring Contractual Performance." *Journal of Political Economy* 89 (4): 615–641. https://doi.org/10.1086/260996

Kleinberg, J., and M. Raghavan (2021). "Algorithmic Monoculture and Social Welfare." *Proceedings of the National Academy of Sciences* 118 (22): e2018340118. https://doi.org/10.1073/pnas.2018340118

Korinek, A., and A. Balwit (2022). "Aligned with Whom? Direct and Social Goals for AI Systems." NBER Working Paper 30017 (https://doi.org/10.3386/w30017). Published in J. B. Bullock et al., eds., *The Oxford Handbook of AI Governance*, 65–85. Oxford: Oxford University Press, 2024. https://doi.org/10.1093/oxfordhb/9780197579329.013.4

Laffont, J.-J., and J. Tirole (1988). "Repeated Auctions of Incentive Contracts, Investment, and Bidding Parity with an Application to Takeovers." *RAND Journal of Economics* 19 (4): 516–537. https://doi.org/10.2307/2555455

London Economics (2013). *The Value of Lost Load (VoLL) for Electricity in Great Britain: Final Report for OFGEM and DECC.* London: London Economics, July 2013.

Lynch, A., B. Wright, C. Larson, S. J. Ritchie, S. Mindermann, E. Hubinger, E. Perez, and K. Troy (2025). "Agentic Misalignment: How LLMs Could Be Insider Threats." arXiv:2510.05179. https://doi.org/10.48550/arXiv.2510.05179

Mailath, G. J., and L. J. Mester (1994). "A Positive Analysis of Bank Closure." *Journal of Financial Intermediation* 3 (3): 272–299. https://doi.org/10.1006/jfin.1994.1007

Meinke, A., B. Schoen, J. Scheurer, M. Balesni, R. Shah, and M. Hobbhahn (2024). "Frontier Models are Capable of In-context Scheming." arXiv:2412.04984. https://doi.org/10.48550/arXiv.2412.04984

Myerson, R. B. (2015). "Moral Hazard in High Office and the Dynamics of Aristocracy." *Econometrica* 83 (6): 2083–2126. https://doi.org/10.3982/ECTA9737

National Audit Office (2022). *The Energy Supplier Market.* Report by the Comptroller and Auditor General, HC 68, Session 2022–23. London: National Audit Office, 22 June 2022.

National Audit Office (2023). *Investigation into Bulb Energy.* Report by the Comptroller and Auditor General, HC 1202, Session 2022–23. London: National Audit Office, 29 March 2023.

National Grid ESO (2019). *Technical Report on the Events of 9 August 2019.* National Grid Electricity System Operator, 6 September 2019.

NERC (2025a). *Incident Review: Considering Simultaneous Voltage-Sensitive Load Reductions.* North American Electric Reliability Corporation, 8 January 2025.

NERC (2025b). "Industry Recommendation: Large Load Interconnection, Study, Commissioning, and Operations." Level 2 NERC Alert, North American Electric Reliability Corporation, 9 September 2025.

NESO (2026). *Response Services Service Terms*, version 6, effective 31 July 2026. Warwick: National Energy System Operator.

Ofgem (2014). *Electricity Balancing Significant Code Review: Final Policy Decision.* London: Office of Gas and Electricity Markets, 15 May 2014.

Ofgem (2026). "Decision to Reject Proposed Dynamic Response Services Amendments to the Terms and Conditions Related to Balancing." Letter to all interested stakeholders, 1 July 2026. Glasgow: Office of Gas and Electricity Markets.

Omohundro, S. M. (2008). "The Basic AI Drives." In P. Wang, B. Goertzel, and S. Franklin, eds., *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, 483–492. Amsterdam: IOS Press.

Perotti, E. C., and J. Suarez (2002). "Last Bank Standing: What Do I Gain if You Fail?" *European Economic Review* 46 (9): 1599–1622. https://doi.org/10.1016/S0014-2921(02)00241-6

Philippon, T., and O. Wang (2023). "Let the Worst One Fail: A Credible Solution to the Too-Big-To-Fail Conundrum." *Quarterly Journal of Economics* 138 (2): 1233–1271. https://doi.org/10.1093/qje/qjac044

Pinson, P., C. Chevallier, and G. N. Kariniotakis (2007). "Trading Wind Generation from Short-Term Probabilistic Forecasts of Wind Power." *IEEE Transactions on Power Systems* 22 (3): 1148–1156. https://doi.org/10.1109/TPWRS.2007.901117

Polinsky, A. M., and S. Shavell (2000). "The Economic Theory of Public Enforcement of Law." *Journal of Economic Literature* 38 (1): 45–76. https://doi.org/10.1257/jel.38.1.45

Rolnick, D., P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, et al. (2022). "Tackling Climate Change with Machine Learning." *ACM Computing Surveys* 55 (2): Article 42, 1–96. https://doi.org/10.1145/3485128

Salib, P. N., and S. Goldstein (2026). "AI Rights for Human Safety." *Virginia Law Review* 112 (4): 1061–1152.

Sappington, D. E. M. (2005). "Regulating Service Quality: A Survey." *Journal of Regulatory Economics* 27 (2): 123–154. https://doi.org/10.1007/s11149-004-5341-9

Schlatter, J., B. Weinstein-Raun, and J. Ladish (2026). "Incomplete Tasks Induce Shutdown Resistance in Some Frontier LLMs." *Transactions on Machine Learning Research*. arXiv:2509.14260. https://doi.org/10.48550/arXiv.2509.14260

Segerson, K. (1988). "Uncertainty and Incentives for Nonpoint Pollution Control." *Journal of Environmental Economics and Management* 15 (1): 87–98. https://doi.org/10.1016/0095-0696(88)90030-7

Shavell, S. (1979). "On Moral Hazard and Insurance." *Quarterly Journal of Economics* 93 (4): 541–562. https://doi.org/10.2307/1884469

Shavell, S. (1984). "Liability for Harm versus Regulation of Safety." *Journal of Legal Studies* 13 (2): 357–374. https://doi.org/10.1086/467745

Shavell, S. (1986). "The Judgment Proof Problem." *International Review of Law and Economics* 6 (1): 45–58. https://doi.org/10.1016/0144-8188(86)90038-4

Shleifer, A., and R. W. Vishny (1989). "Management Entrenchment: The Case of Manager-Specific Investments." *Journal of Financial Economics* 25 (1): 123–139. https://doi.org/10.1016/0304-405X(89)90099-8

Skalse, J., N. Howe, D. Krasheninnikov, and D. Krueger (2022). "Defining and Characterizing Reward Hacking." In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 9460–9471. https://doi.org/10.52202/068431-0687

Soares, N., B. Fallenstein, E. Yudkowsky, and S. Armstrong (2015). "Corrigibility." In T. Walsh, ed., *Artificial Intelligence and Ethics: Papers from the 2015 AAAI Workshop*, AAAI Technical Report WS-15-02. AAAI Press.

Thompson, H. A. (2025). "Some Economics of Artificial Superintelligence." arXiv:2511.06613. https://doi.org/10.48550/arXiv.2511.06613

Turner, A. M., L. Smith, R. Shah, A. Critch, and P. Tadepalli (2021). "Optimal Policies Tend to Seek Power." In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 23063–23074.

Uptime Institute (n.d.). "Tier Classification System." https://uptimeinstitute.com/tiers, accessed 1 October 2026.

US Department of Energy (2026). "2026 DOE Federal Power Act 202(c) Orders." Office of Cybersecurity, Energy Security, and Emergency Response. https://www.energy.gov/ceser/2026-doe-202c-orders, accessed 1 October 2026.

Weisman, D. L. (2026). "The Commitment Problem in Regulatory Regimes." *Journal of Regulatory Economics* 69 (1): Article 25. https://doi.org/10.1007/s11149-026-09522-2

Williamson, O. E. (1976). "Franchise Bidding for Natural Monopolies—in General and with Respect to CATV." *Bell Journal of Economics* 7 (1): 73–104. https://doi.org/10.2307/3003191

## Appendix A. Proofs

The proof of Proposition 3 is in Appendix B, with the dynamic model.

*Proof of Proposition 1.* By (3), $\partial \mathbb{E}[S_T]/\partial q_T = \Phi(z_T)$ and $\partial^2 \mathbb{E}[S_T]/\partial q_T^2 = \varphi(z_T)/\sigma_T > 0$, so objective (6) is strictly concave. Its first-order condition, $\Phi(z_T) = (p - c)/(\lambda + \theta\Lambda)$, has a unique solution because $\lambda + \theta\Lambda > p - c$ under Assumption 1, and the constraint $q_T \geq 0$ gives (7). At an interior solution, the coordinator falls short with probability $\Pr(K_T < q_T) = \Phi(z_T) = (p - c)/(\lambda + \theta\Lambda)$. For any continuous distribution $F_T$ of $K_T$ with a positive density at the relevant quantile, the derivative of $\mathbb{E}[(q - K_T)^+]$ is $F_T(q)$, so the same argument gives $F_T(q_T^A) = (p - c)/(\lambda + \theta\Lambda)$. The planner, who maximizes (5), chooses $q_T^o$ in the same way, with $\nu$ in place of $\lambda + \theta\Lambda$. Since $\Phi^{-1}$ is strictly increasing, the interior solutions coincide if and only if $\lambda + \theta\Lambda = \nu$, which requires $\nu \geq \lambda$ because $\Lambda \geq 0$. This is part (a). With $\Lambda = 0$, $q_T^A - q_T^o$ has the sign of $\nu - \lambda$ (weakly, given the constraint $q_T \geq 0$), which is part (b). For part (c), $q_T^A$ is strictly decreasing in $\Lambda$ while it is positive, and $q_T^A \leq q_T^o$ at $\Lambda = 0$ when $\lambda \geq \nu$, so $q_T^A < q_T^o$ for every $\Lambda > 0$ when $q_T^o > 0$. For part (d), a coordinator with no weight on profit minimizes $\theta\Lambda\mathbb{E}[S_T]$. Since $d\mathbb{E}[S_T]/dq_T = \Phi(z_T) > 0$, the minimum over $q_T \geq 0$ is at $q_T = 0$ when $\Lambda > 0$, and every commitment is optimal when $\Lambda = 0$. In the dynamic model of Appendix B, $\Lambda > 0$ if and only if $V > C$.

*Proof of Proposition 2.* The coordinator's objective is additively separable in $q_T$ and $e$. Its part in $q_T$ is (6), so the commitment satisfies Proposition 1 with the stake $\Lambda$. Its part in $e$ is $-\pi_0(e)\Lambda - \gamma(e)$, which is strictly concave because $\pi_0$ is convex and $\gamma$ is strictly convex. With $\Lambda = 0$, it is $-\gamma(e)$, which is largest at $e = 0$, whatever $\lambda$, and this is part (a). With $\Lambda > 0$, the precaution solves $-\pi_0'(e)\Lambda = \gamma'(e)$. The left-hand side is positive and nonincreasing in $e$, and it tends to zero as $e$ grows, because $\pi_0$, a probability, is bounded below. The right-hand side starts at zero

and is strictly increasing, so a solution $e(\Lambda)$ exists, and it is unique, positive and strictly increasing in $\Lambda$. The planner maximizes (5) less $\pi_0(e)B + \gamma(e)$, so the efficient precaution solves the same condition with $B$ in place of $\Lambda$, and $e(\Lambda)$ equals it if and only if $\Lambda = B$. Given $\Lambda = B$, Proposition 1(a) gives the efficient commitment, when it is positive, if and only if $\lambda + \theta B = \nu$, that is, $\lambda = \nu_0$ by (4). This is part (b). For part (c), $\Lambda < B$ implies that $e(\Lambda)$ is below the efficient precaution, and $\lambda$ does not enter the condition for $e$. A penalty $\lambda = \nu - \theta\Lambda$ therefore restores the efficient commitment by Proposition 1(a), but not the efficient precaution.

*Proof of Remark 1.* The coordinator's private loss from shedding is linear in the amounts shed from each pool, so it sheds first from the pool with the lower loss per MWh, which is its own load if and only if $k_A < k_H$. The planner sheds first from the pool with the lower social value, which is the coordinator's own load because $v_A < v_H$. When $k_A > k_H$ and $X \le P_A$, the coordinator sheds $X$ MW of consumer load where the planner would shed $X$ MW of its own, which raises the social cost by $(v_H - v_A)X$ per hour.

*Proof of Proposition 4.* For equation (12), note that each coordinator keeps its business after a collapse with probability $1 - \psi$, and that its business passes in a share $\omega$ to the survivors if it is dropped and at least one other coordinator survives, which happens with probability $\psi - \psi^M$. The incumbents' expected total franchise after a collapse is therefore $F(1 - \psi) + \omega F(\psi - \psi^M)$. By symmetry, each coordinator's expected franchise is the $M$-th part of this, and subtracting it from $F/M$ gives the first term of (12). The second term is the AI's own operation, which it loses with probability $\psi$. (a) With $\psi \equiv 1$, $\Lambda_M = F/M + A$, which falls with $M$ toward $A$. (b) With $A = \omega = 0$, $\Lambda_M = Fg(M)$, where $g(M) \equiv \psi(M)/M$. Let $y \equiv \Gamma/M$. Since $\psi(M) = 1 - H_\xi(y)$, the elasticity of $\psi$ with respect to $M$ is $\chi(M) = y\zeta(y)$, where $\zeta$ is the hazard rate of $\xi$, and $d\ln g/d\ln M = \chi(M) - 1$. With a nondecreasing hazard rate, $y\zeta(y)$ increases with $y$ and so decreases

with $M$. Hence $g$ rises while $\chi > 1$ and falls once $\chi < 1$, so it is single-peaked, with its peak at $M = 1$ if $\chi(1) \leq 1$. Over the integers, the maximum is at one of the two integers adjacent to the peak. For the exponential distribution, $\zeta = 1/\bar{\xi}$, so $\chi(M) = \Gamma/(\bar{\xi}M)$ and the peak is at $M^* = \Gamma/\bar{\xi}$. The probability of a shortfall, $(p - c)/(\lambda + \theta F g(M))$, decreases in $g$, so it is U-shaped with its minimum at the same $M$. Alignment requires $\theta F g(M) = \nu - \lambda$ for some $M$. With (11), $g(M) \leq \bar{\xi}/(e\Gamma)$, the value at $M = \Gamma/\bar{\xi}$, so alignment requires $\theta F \geq e(\Gamma/\bar{\xi})(\nu - \lambda)$. With $\nu_0 = \lambda$, (4) gives $\theta = (\nu - \lambda)/B$, and the condition becomes $F \geq e(\Gamma/\bar{\xi})B$. (c) With $\omega = 1$, $A = 0$ and (11), $\psi(M)^M = e^{-\Gamma/\bar{\xi}}$ for every $M$, so $\Lambda_M = e^{-\Gamma/\bar{\xi}}F/M$, which is largest at $M = 1$. Alignment then requires $\theta e^{-\Gamma/\bar{\xi}}F \geq \nu - \lambda$, which with $\nu_0 = \lambda$, so that $\theta = (\nu - \lambda)/B$, becomes $F \geq e^{\Gamma/\bar{\xi}}B$. (d) With (11), $\psi'(M) = (M^*/M^2)\psi(M)$. For $\omega = 0$, $\Lambda_M = \psi(M)(F/M + A)$, so $d\Lambda_M/dM = (\psi/M^2)[M^*(F/M + A) - F]$, which is positive at every $M$ if $M^*A > F$. For $\omega = 1$, $\Lambda_M = e^{-M^*}F/M + \psi(M)A$, so $d\Lambda_M/dM = [M^*A\psi(M) - e^{-M^*}F]/M^2$, which is positive at every $M \geq 1$ if $M^*A > F$, since $\psi(M) \geq e^{-M^*}$. Since $\Lambda_M$ is linear in $\omega$, it rises with $M$ for every $\omega \in [0,1]$ when $A > F/M^*$, and the probability of a shortfall, $(p - c)/(\lambda + \theta\Lambda_M)$, then falls with $M$. (e) Since $H_\xi$ has a positive density, $\psi(M) = 1 - H_\xi(\Gamma/M)$ falls as $\Gamma$ rises, at every $M$, and every term of (12) increases with $\psi$, so a lower $\Gamma$ raises every coordinator's stake. In case (b), the stake peaks where $y\zeta(y)$, which increases with $y$, crosses one, with $y = \Gamma/M$. Let $y^*$ denote that crossing point, which does not depend on $\Gamma$. The peak is then at $M = \Gamma/y^*$, or at $M = 1$ if $\Gamma < y^*$, so it falls in proportion to $\Gamma$ until it reaches $M = 1$. With (11), $y^* = \bar{\xi}$.

*Proof of Proposition 5.* (a) Let $V_\psi^*$ be the coordinator's value under (B1) when a collapse ends its role with probability $\psi$ and its commitment is optimal. It makes itself irreplaceable if and only if $(V_0^* - V_\psi^*)/b > k$. At the commitment that is optimal under $\psi$, (B5) gives

$$V_0 - V_\psi = \frac{\beta\pi\psi\beta^\tau[u+\beta\pi ha(1-\beta^\tau)/(1-\beta)]}{[1-\beta+\beta\pi\kappa(0)][1-\beta+\beta\pi\kappa(\psi)]},$$

and re-optimizing the commitment after escaping can only raise $V_0^*$. When $\tau(1-\beta)$ and $\pi\tau$ are small and $\psi$ is large relative to $\tau(1-\beta)$, $\beta\pi\kappa(0) = \beta\pi(1-\beta^\tau)$ is small relative to $1-\beta$, the term $\beta\pi ha(1-\beta^\tau)/(1-\beta)$ is small relative to $u$, and $(1-\beta^\tau)(u-ha)$ is small relative to $\psi\beta^\tau u$. Since $\kappa(\psi)u - ha(1-\beta^\tau) = \psi\beta^\tau u + (1-\beta^\tau)(u-ha)$, (B5) then gives $\Lambda_\psi \approx \beta\psi\beta^\tau u/(b[1-\beta+\beta\pi\kappa(\psi)])$, and hence, to first order in $\pi\tau$ and $\tau(1-\beta)/\psi$, $(V_0 - V_\psi)/b \approx \pi\Lambda_\psi/(1-\beta) \approx \pi\Lambda_\psi/r$. (b) follows from part (a) and Proposition 1(a), and $\Lambda^o = B$ when $v_0 = \lambda$ by (8). (c) The stake $\Lambda_\psi = \beta[\kappa(\psi)V_\psi^* - ha(1-\beta^\tau)/(1-\beta)]/b$ is affine in the value $V_\psi^*$, so by the envelope theorem the commitment, and with it $\Pi^A$ and $\pi$, can be held fixed when differentiating with respect to $a$. (B5) then gives $d\Lambda_\psi/da = \beta[\kappa(\psi) - h(1-\beta^\tau)]/(bd_\psi) > 0$, where $d_\psi \equiv 1-\beta+\beta\pi_\psi\kappa(\psi)$, since $\kappa(\psi) > h(1-\beta^\tau)$ when $\psi > 0$. In the same way, $d(V_0^* - V_\psi^*)/da = (1+\beta\pi_E hD)/d_0 - (1+\beta\pi_\psi hD)/d_\psi$, where $D \equiv (1-\beta^\tau)/(1-\beta)$, $d_0 \equiv 1-\beta+\beta\pi_E\kappa(0)$, and $\pi_E$ and $\pi_\psi$ are the probabilities of a collapse at the optimal commitments with and without escape. Since $\pi_E \geq \pi_\psi$, a sufficient condition for this derivative to be positive is $d_0 < d_\psi$, which holds because $\kappa(0)$ is small relative to $\kappa(\psi)$ and the coordinator's own contribution to the probability of a collapse is small. The value of escape therefore rises with $a$, without bound, and exceeds $bk$ above a unique threshold, which may be zero. Above it, the stake is the restart stake given by (B3), which with $h = 1$ does not depend on $a$ and is approximately $\tau\Pi^A$, the profit lost during the outage. By Proposition 1, the probability of a shortfall is then close to $(p-c)/\lambda$ when $\tau\Pi^A$ is small relative to $B$, since $\theta = (v-v_0)/B$ by (4).

*Proof of Corollary 1.* Since a marginal increase in the coordinator's expected shortfall raises $\pi_c$ by $\theta$ and leaves $\pi_o$ unchanged, the first-order condition is (B2) with $V - C_c$ in place of $V - C$, so

the stake at the margin is $\Lambda_c = \beta(V - C_c)/b$, which (B6) gives. Its value $V$ rises as $\psi_o$ falls, because $\kappa(\psi_o)$ falls, and $\Lambda_c$ is increasing in $V$. This is part (a). For part (b), hold the commitment fixed. The value after escaping is (B6) with $\psi_c = \psi_o = 0$, and since $\kappa(\psi) - \kappa(0) = \psi\beta^\tau$,

$$V_0 - V = \frac{\beta^{\tau+1}(\pi_c\psi_c + \pi_o\psi_o)(u + \beta\pi haD)}{d_0 d},$$

where $d_0$ and $d$ are the denominators of (B6) after and before escape. Under the conditions of Proposition 5 with $\psi_c$ in place of $\psi$, the argument in its proof gives $\Lambda_c \approx \beta\psi_c\beta^\tau u/(bd)$, with $\beta\pi haD$ small relative to $u$ and $d_0 \approx 1 - \beta$. Hence $(V_0 - V)/b \approx \Lambda_c(\pi_c\psi_c + \pi_o\psi_o)/(\psi_c(1 - \beta))$, which is (14) to first order. Part (c) follows by comparing $\psi_o = 0$ with $\psi_o = \psi_c$ in parts (a) and (b).

## Appendix B. A dynamic model of the stake

Time is discrete. In each period in which it operates and the grid is up, the coordinator chooses a commitment $q$ and receives the reward $a + b\Pi_T(q)$, where $\Pi_T(q) = (p - c)q - \lambda\mathbb{E}[S_T(q)]$, $a \geq 0$ rewards operation and $b \geq 0$ weights profit. The pure self-preservation objective, the expected discounted number of periods in operation, is the case $a = 1$ and $b = 0$. The system collapses at the end of the period with probability $\pi(q) = \pi_0 + \theta\mathbb{E}[S_T(q)] < 1$. A collapse lasts $\tau$ periods. During it, the coordinator earns no profit and receives the reward $a$ only if its computers run on backup power, which happens with probability $h$. Let $C$ denote the coordinator's value at the start of a collapse. Its value $V$ in normal operation satisfies

$$V = \max_{q \geq 0}\{a + b\Pi_T(q) + \beta[(1 - \pi(q))V + \pi(q)C]\}. \tag{B1}$$

If the coordinator restarts after the collapse, $C = ha(1 - \beta^\tau)/(1 - \beta) + \beta^\tau V$. If the collapse ends its tenure, $C = ha(1 - \beta^\tau)/(1 - \beta)$.

*Existence and the reduced form.* In both cases the right-hand side of (B1) is a contraction in $V$ with modulus at most $\beta$, so $V$ is unique. When $V > C$, the maximand is strictly concave in $q$, because $\Pi_T$ is concave and $\pi$ is strictly convex. With $b > 0$, the first-order condition is

$$b[(p - c) - \lambda\Phi(z_T)] = \beta\theta(V - C)\Phi(z_T), \tag{B2}$$

which is the first-order condition of (6) with the stake $\Lambda = \beta(V - C)/b$. With $b = 0$, the maximand is $a + \beta V - \beta\pi(q)(V - C)$, so the coordinator minimizes $\pi(q)$ if $V > C$ and is indifferent among all commitments if $V = C$, which gives Proposition 1(d).

*Proof of Proposition 3.* Solving (B1) at the optimal commitment, with $\Pi^A$ the profit and $\pi$ the probability of a collapse at that commitment, gives exactly, under a restart,

$$\Lambda = \frac{\beta(1-\beta^\tau)[(1-h)a/b + \Pi^A]}{1 - \beta + \beta\pi(1-\beta^\tau)}, \tag{B3}$$

and, under loss of tenure,

$$\Lambda = \frac{\beta[(a/b)(1-h(1-\beta^{\tau}))+\Pi^{A}]}{1-\beta+\beta\pi}. \tag{B4}$$

When $\tau(1-\beta)$ is small, $1-\beta^{\tau} \approx \tau(1-\beta)$, and when $\pi\tau$ is also small, (B3) gives (9). Dividing the numerator and denominator of (B4) by $\beta$ and using $1-\beta^{\tau} \approx \tau(1-\beta) \approx \tau r$ gives (10). Part (c) follows from comparing (9) and (10). In particular, with $h = 1$ and $b = 0$, $V = C$ under a restart, and the coordinator has no stake at all.

*A general probability of replacement.* If a collapse ends the coordinator's role with probability $\psi$, its value at the start of a collapse is $C = ha(1-\beta^{\tau})/(1-\beta) + (1-\psi)\beta^{\tau}V$. Let $\kappa(\psi) \equiv 1-(1-\psi)\beta^{\tau}$ and $u \equiv a + b\Pi^{A}$. Solving (B1) at the optimal commitment gives

$$V_{\psi} = \frac{u+\beta\pi ha(1-\beta^{\tau})/(1-\beta)}{1-\beta+\beta\pi\kappa(\psi)}, \Lambda_{\psi} = \frac{\beta[\kappa(\psi)u-ha(1-\beta^{\tau})]}{b[1-\beta+\beta\pi\kappa(\psi)]}, \tag{B5}$$

which gives (B3) with $\psi = 0$ and (B4) with $\psi = 1$.

*Attribution.* Suppose instead that a collapse caused by the coordinator's shortfalls, which occurs with probability $\pi_c = \theta\mathbb{E}[S_T(q)]$, ends its role with probability $\psi_c$, and that a collapse from other causes, which occurs with probability $\pi_o = \pi_0$, ends it with probability $\psi_o$. Its value at the start of the two kinds of collapse is $C_x = haD + (1-\psi_x)\beta^{\tau}V$ for $x = c, o$, where $D \equiv (1-\beta^{\tau})/(1-\beta)$. Solving (B1), with $\pi(q)C$ replaced by $\pi_c C_c + \pi_o C_o$, at the optimal commitment gives

$$V = \frac{u+\beta\pi haD}{1-\beta+\beta\pi_c\kappa(\psi_c)+\beta\pi_o\kappa(\psi_o)}, \Lambda_c = \frac{\beta[\kappa(\psi_c)V-haD]}{b}, \tag{B6}$$

where $\Lambda_c$ is the stake at the margin, since a marginal shortfall raises only $\pi_c$. With $\psi_c = \psi_o = \psi$, (B6) gives (B5).