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.

Governance of Unsafe Behavior in Electric Power Maintenance: A Three-Population Evolutionary Game with Endogenous Risk and a Conserved Quantity

Kaikai Man

1062627233@qq.com

North China University of Water Resources and Electric Power

Xue Yang

North China University of Water Resources and Electric Power

Article

Keywords:

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11616764/v1

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

Additional Declarations: No competing interests reported.

Governance of Unsafe Behavior in Electric Power Maintenance: A Three-Population Evolutionary Game with Endogenous Risk and a Conserved Quantity

Kaikai Man[1,*] & Xue Yang[1]

[1] School of Management and Economics, North China University of Water Resources and Electric Power, Zhengzhou 450046, Henan, China
*Corresponding author. E-mail: 1062627233@qq.com (K.M.); 18265786869@163.com (X.Y.)

Abstract

Maintenance crews work inside short windows on de-energized equipment, and the casualties that follow are few, severe and almost always preceded by a departure from a written procedure. Three agents decide whether that departure happens: the crew, the supervisor and the utility that pays for supervision. We model the layers as a three-population evolutionary game in which accident probability is endogenous and super-linear, penalties are proportional to fault, and the supervisor has a reason to look away. A vertex criterion classifies the eight pure-strategy states, and a Latin hypercube design with partial rank correlation ranks the parameters. Three closed-form thresholds split the risk sensitivity into four regimes, with no stable pure strategy between the last two; there the flow preserves volume in logit coordinates, the compliance face admits an explicit first integral, and the equilibrium is a nonlinear center of computable period. Costlier regulation destroys the center and makes two delegated-governance states bistable. The levers are the effort a shortcut saves, regulation cost, accountability and schedule pressure, not the penalties.

1. Introduction

Risk concentrates in short windows whenever crews work on equipment that is energized or that was de-energized minutes earlier. A switching sequence runs without verification. Crews push their working range past the permit. Somebody unbolts a guard rail and then falls. A temporary protective ground is left off before work begins. Accidents of this kind are few and severe, and their proximate cause is usually a departure from a written procedure that was available at the site[1,2]. Studies of power-sector casualties place procedure violations and missing on-site supervision consistently among the leading contributory factors[3–5]. Reliability studies reach the same phase from the other side, treating maintenance as the point where availability, cost, and human exposure are traded against one another[6–8]. Why does this count as a governance problem rather than an engineering one? The person who decides to skip a step and the person paid to notice the omission are different agents, and neither owns the residual loss.

Behavioral safety research has mapped layering. Unsafe acts are the proximal events through which latent organizational weaknesses become injuries[9,10], and the systems-theoretic reading of accidents locates enforcement, or its quiet relaxation, in the supervisory layer[11]. What a supervisor tolerates predicts what a crew does[12–14]. Safety-culture work describes the same mechanism through shared assumptions about which rules can be negotiated[15,16]. The picture is descriptive throughout. Supervision is shown to matter, but

the supervisor is never modeled as an agent who pays a cost for supervising and therefore has a reason of their own to look away.

Evolutionary game theory supplies the strategic layer that is missing. Maynard Smith and Price introduced the evolutionarily stable strategy for animal conflict[17], and Taylor and Jonker tied it to the replicator dynamics that move a population toward above-average payoffs[18–20]. Replicator dynamics now cover safety governance in coal mining[21–24], construction[25,26], food safety[27], chemical production[28], internal inspection[29], and railway transport[30]. Two features recur across these studies. Accident probability is normally a constant or a linear function of one behavioral variable; therefore, risk enters the payoff as a price instead of a state of the system. Meanwhile, penalties attach to the event rather than to the agent. Detect a violation, and every party to the transaction is fined, which overstates deterrence for the party that behaved correctly and blunts the incentive to monitor selectively.

The present study maintains the three-population structure and alters three modeling choices. What follows depends on these choices, not on the structure alone. First, accident probability is endogenous and super-linear. It rises with the intensity of unsafe behavior, with the leniency of the supervisor, and faster still with the product of the two, the formal counterpart of an everyday observation: violations are most dangerous when nobody is watching. Second, penalties are proportional to fault, and joint liability reaches a supervisor only when the duty was not performed; thus, the firm, supervisor, and crew form a chain of delegated duties rather than a common pool of liability. Third, the supervisory population has its own payoff. Strict duty costs effort and social capital, leniency saves both, and the firm's willingness to pay for supervision is itself a strategy.

The resulting replicator system is tractable without being trivial. Each equation depends only on the two other populations; therefore, the Jacobian is diagonal at every vertex, and the eight pure-strategy states can be classified in closed form. Three risk thresholds fall out of that classification, and between the last two lies an interval in which no pure strategy is stable. Applied work usually dismisses such an interval as a transition zone. Here, it is the object of study. The flow is volume-preserving in logit coordinates, which rules out an asymptotically stable interior state. On the invariant face of full compliance, the two-population reduction admits an explicit first integral, and its interior equilibrium is a nonlinear center whose period can be computed from the parameters. None of this is a technical curiosity. The center states formally the crackdown-relaxation cycle that regulators describe in words: a firm tightens supervision after an event, relaxes once the record improves, and tightens again when risk drifts back up.

A second family of thresholds concerns money rather than risk. Push the cost of strong regulation past a critical value, and the central equilibrium disappears. Bistability follows, between two states that both delegate governance away from the firm. In one, supervision is strict, while the crew takes shortcuts. In the other, the crew works safely, and supervision is dormant. Which state is reached depends on the initial conditions, a dependence that no single-trajectory simulation can reveal and that bears directly on where a regulator should intervene.

This paper addresses two questions: which parameters decide the governance configuration that a maintenance unit settles into, and whether the instruments regulators rely on are among them. Four results are new. One is a three-population safety game with endogenous, interactive accident risk, fault-proportional penalties, and a supervisor who is an agent rather than a constraint. Another is an exact classification of the pure-strategy states with closed-form risk and cost thresholds, which partitions the parameter space into four qualitatively different regions instead of two. A third characterizes the region that has no pure-strategy equilibrium: volume preservation in logit coordinates, an explicit first integral on the compliance face, a nonlinear center, and a relation between amplitude

and period. The fourth is empirical. A calibration protocol binds every parameter to an observable proxy, and a global uncertainty analysis then identifies the parameters that actually move the equilibrium. That answer is not the one penalty-centered policy would suggest.

The remainder of the paper is organized as follows. Section 2 sets out the model, Section 3 classifies the pure-strategy states and derives the risk thresholds, and Section 4 characterizes the interval in which no pure strategy survives. Section 5 traces the bistability that the cost of regulation creates. A hazard backlog that closes with a lag enters in Section 6. Section 7 reports the calibration and the numerical analysis. Sections 8 and 9 discuss the results and conclude.

2. Model and methods

2.1. Three Populations and Their Strategies

The model has three populations of agents, and each maintenance task draws one member of each. Crews are the first. A share x follows the written procedure, which means verified isolation, temporary grounding, attention to the permit boundary and to fall protection; the remaining share $1-x$ takes shortcuts. Supervisors are the second. A share y keeps strict duty: present at the work point, correcting on the spot, and closing rectification orders. The rest, share $1-y$, keep lenient duty, which can be remote, formal, or simply tolerant of the shortcuts the crew already takes. Utilities are the third, and a share z of their units runs strong regulation: safety staffing, video surveillance, unannounced inspection, and sanctions tied to performance. The remaining share $1-z$ runs weak regulation.

The three populations are coupled in a chain, and the coupling runs in one direction at each link. Regulatory intensity at the firm decides how likely a lenient supervisor is to be found out. Supervisory diligence decides how likely a shortcut is to be caught at the work point. Crew behavior decides the accident probability that all three end up bearing. What makes the chain worth modeling is the middle layer, because the supervisor is an agent with their own incentives. Strict duty costs time and social capital. The supervisor must be present, must interrupt work, must file rectification orders, and must absorb the resentment of a crew whose bonus depends on finishing inside the outage window. Leniency saves all of it. A supervisor modeled as a mechanical enforcer therefore removes the friction that makes supervision strategic, and it also removes the collusion problem that hierarchies are known to carry[31].

2.2. Endogenous Accident Probability and Payoffs

Let $u=1-x$ denote the intensity of unsafe behavior and $v=1-y$ denote the intensity of lenient duty. The probability that a task ends in an accident is

$$\lambda(u,v) = \lambda_0 + \beta[\delta u + \gamma v + \varphi\delta\gamma uv] \quad (1)$$

where $\lambda_0>0$ is the residual risk of a task performed correctly, $\delta>0$ and $\gamma>0$ scale the two behavioral intensities, $\beta>0$ is the risk sensitivity of the technology and the organization, and $\varphi\geq 0$ is the interaction coefficient. The interaction term is the formal statement that a shortcut is most dangerous when nobody is watching: the marginal effect of u on risk is $\beta\delta(1+\varphi\gamma v)$, which increases with leniency. When $\varphi=0$, the two channels are additive, and the model reduces to the standard linear specification; the qualitative results below do not depend on $\varphi>0$, but the quantitative thresholds do.

Losses are apportioned at the moment of the accident. A single accident costs D_W to the injured worker (injury, disability, lost earnings), D_S to the supervisor who is held accountable, and D_E to the utility (outage, equipment, fines, reputation). Detection is

probabilistic and depends on the two supervisory layers. A shortcut is detected with probability

$$p_W(y,z) = p_0 + p_y y + p_z z \quad (2)$$

where p_0 is the residual detection probability (colleague report, automatic video recognition), p_y is the marginal contribution of an on-site supervisor, and p_z is the marginal contribution of the firm's own inspection. A lenient supervisor is detected with probability $p_S(z)=s_0+s_z z$.

Penalties are proportional to the fault. A detected violation costs the violator a fine K_W and raises the reputational loss from L to θL, with $\theta>1$, so that the expected penalty of a worker is

$$P_W(y,z) = L_W + p_W(y,z)[(\theta-1)L_W + K_W] \quad (3)$$

and that of a supervisor is

$$P_S(z) = L_S + p_S(z)[(\theta-1)L_S + K_S] \quad (4)$$

Reputational loss L is borne whether or not the violation is detected, because it accumulates through the safety record; detection adds the fine and multiplies the reputational component. Two consequences follow. The party that behaved correctly is never fined for the other party's violation, and the deterrence of a violation depends on the number of supervisory layers that are active rather than on a single penalty parameter.

Joint liability is treated separately. When a crew takes a shortcut and an accident follows, the supervisor bears a fraction η of the management-side loss, consistently with the "one post, two responsibilities" principle of Chinese work-safety law[2]. Liability is incurred only in the lenient state: a supervisor who performed the duty and was overruled is not charged with the consequence. The expected payoffs of the three populations in the eight pure-strategy combinations are developed below.

Each departure carries private gain and public cost. A crew that takes shortcuts saves effort worth δe_W; a supervisor who performs lenient duty saves effort and social cost worth γe_S. A firm that regulates strongly bears the cost C and reduces the accident probability by a proportion ψ; a firm that regulates weakly bears an accountability loss B per unit of tolerated departure, which represents the sanction of the parent company and of the energy regulator. Strong regulation also slows the outage schedule, because the work that was displaced by shortcuts must be done properly; this progress loss is $g[\delta u+\gamma v]$. Finally, strong regulation raises the detection of departures, which returns fines K_W and K_S to the firm at the rate p_z.

The parameter set, its interpretation, and the baseline values are listed in Table 1.

Table 1 | Parameters of the model and baseline values. Monetary quantities are expressed in units of the worker's loss DW; Section 7.1 describes how each value is anchored to an observable proxy.

Parameter	Meaning	Baseline
δ	Intensity of unsafe behavior	0.5
γ	Intensity of lenient duty	0.5
e_W	Effort saved by a shortcut	15
e_S	Effort and social cost saved by leniency	14
λ_0	Residual accident probability	0.02
β	Risk sensitivity (scenario parameter)	0.01 / 0.06 / 0.30
φ	Interaction of the two behavioral channels	1.0

Parameter	Meaning	Baseline
D_W	Loss borne by the injured worker	120
D_S	Loss borne by the accountable supervisor	40
D_E	Loss borne by the utility	200
p_0	Residual detection probability	0.10
p_y	Marginal detection by the supervisor	0.30
p_z	Marginal detection by the firm	0.20
s_0	Residual detection of lenient duty	0.10
s_z	Marginal detection of lenient duty	0.35
K_W	Fine for a detected shortcut	1.5
K_S	Fine for detected leniency	3
L_W	Reputational loss of the worker	2
L_S	Reputational loss of the supervisor	4
θ	Reputational multiplier upon detection	1.5
η	Joint-liability coefficient	0.4
B	Accountability loss under weak regulation	10
ψ	Preventive efficacy of strong regulation	0.3
g	Schedule-pressure coefficient	3
C	Cost of strong regulation	4.5

With these conventions, the expected payoff of each population in each of the eight pure-strategy combinations follows from Equations (1)–(4). Table 2 lists the payoffs of the crew and the supervisor.

Table 2 | Payoff matrix of the crew and of the supervisor. $\lambda uv=\lambda(u,v)$ is the accident probability of Equation (1); PW and PS are the expected penalties of Equations (3) and (4). The term $-\eta\beta\delta DS$ appears only in the two rows in which the supervisor is lenient and the crew is unsafe.

Situation (crew; supervisor; firm)	U_W (crew)	U_S (supervisor)
Safe; strict; strong	$R_W-\lambda_{00}D_W$	$R_S-\lambda_{00}D_S$
Safe; strict; weak	$R_W-\lambda_{00}D_W$	$R_S-\lambda_{00}D_S$
Safe; lenient; strong	$R_W-\lambda_{01}D_W$	$R_S+\gamma e_S-\lambda_{01}D_S-P_S(1)$
Safe; lenient; weak	$R_W-\lambda_{01}D_W$	$R_S+\gamma e_S-\lambda_{01}D_S-P_S(0)$
Unsafe; strict; strong	$R_W+\delta e_W-\lambda_{10}D_W-P_W(1,1)$	$R_S-\lambda_{10}D_S$
Unsafe; strict; weak	$R_W+\delta e_W-\lambda_{10}D_W-P_W(1,0)$	$R_S-\lambda_{10}D_S$
Unsafe; lenient; strong	$R_W+\delta e_W-\lambda_{11}D_W-P_W(0,1)$	$R_S+\gamma e_S-\lambda_{11}D_S-P_S(1)-\eta\beta\delta D_S$
Unsafe; lenient; weak	$R_W+\delta e_W-\lambda_{11}D_W-P_W(0,0)$	$R_S+\gamma e_S-\lambda_{11}D_S-P_S(0)-\eta\beta\delta D_S$

The payoff of the firm in row (x,y,z) is

$$U_E = \Pi - \lambda(u,v)(1-\psi z)D_E$$
$$-z[C + g(\delta u + \gamma v) - p_z(K_W\delta u + K_S\gamma v)] - (1-z)B(\delta u + \gamma v) \quad (5)$$

with $u=1-x$ and $v=1-y$, and it is omitted from Table 2 because only its difference between the strong and weak columns enters the dynamics; that difference is Equation (8) below.

2.3. Replicator Dynamics

The three populations adjust their strategies by imitation. Let F_W be the payoff advantage of safe over unsafe behavior, F_S the advantage of strict over lenient duty, and F_E the advantage of strong over weak regulation. Subtracting the payoffs in Table 2 yields

$$F_W(y,z) = -\delta e_W + L_W + a_W(y,z)b_W + \beta\delta[1+\varphi\gamma(1-y)]D_W \tag{6}$$

$$F_S(x,z) = L_S + a_S(z)b_S + \beta\gamma[1+\varphi\delta(1-x)]D_S + \eta\beta\delta(1-x)D_S - \gamma e_S \tag{7}$$

$$F_E(x,y) = p_z[K_W\delta(1-x) + K_S\gamma(1-y)] + \psi\lambda(1-x,1-y)D_E + (B-g)[\delta(1-x)+\gamma(1-y)] - C \tag{8}$$

with the shorthand $a_W(y,z)=p_0+p_y\ y+p_z\ z$, $a_S(z)=s_0+s_z\ z$, $b_W=(\theta-1)L_W+K_W$ and $b_S=(\theta-1)L_S+K_S$. The replicator equations are

$$\frac{dx}{dt} = x(1-x)F_W(y,z), \frac{dy}{dt} = y(1-y)F_S(x,z), \frac{dz}{dt} = z(1-z)F_E(x,y) \tag{9}$$

The unit cube is invariant, and each face is invariant because the corresponding factor $x(1-x)$ vanishes there. Time is measured in units of one learning period, and the coefficients represent the rate at which the three populations revise behavior. For notational convenience, the three rates are normalized to one; unequal rates would change the transient speed and the shape of the orbits, but not the location of the equilibria.

2.4. Structural Properties

Three structural properties of (9) perform analytical work. The first is that F_W is independent of x, F_S of y, and F_E of z: the variable a population controls never enters its own payoff difference, so a crew's incentive to be safe depends on whether it is being watched, not on how many crews are already safe. The second concerns the cross-partial derivatives, which are

$$\frac{\partial F_W}{\partial y} = p_y b_W - \beta\delta\varphi\gamma D_W, \frac{\partial F_S}{\partial x} = -\beta\gamma\varphi\delta D_S - \eta\beta\delta D_S, \frac{\partial F_E}{\partial x} < 0, \frac{\partial F_E}{\partial y} < 0 \tag{10}$$

Supervision and compliance are strategic substitutes at both links of the chain, since more of one lowers the marginal return to the other. The signs divide cleanly. The first derivative has an ambiguous sign and is taken up in Proposition 3; the other three are negative without qualification. The third property follows from the first. Because no payoff difference depends on its own variable, the Jacobian of (9) is diagonal at every vertex of the cube, which permits an exact classification rather than a numerical one.

3. Pure-Strategy Stability

3.1. A Vertex Criterion

The eight pure-strategy profiles are the vertices of the cube. Because F_W does not depend on x, the first equation of (9) vanishes identically on the faces $x=0$ and $x=1$, and its derivative with respect to y carries the factor $x(1-x)$; the same holds cyclically. The Jacobian is therefore diagonal at every vertex, and the eigenvalues can be read off without linearizing the payoff functions.

Proposition 1. *At a vertex $(x,y,z)\in\{0,1\}^3$ the Jacobian of (9) is diagonal with entries*

$$\lambda_1 = (1-2x)F_W(y,z), \lambda_2 = (1-2y)F_S(x,z),$$
$$\lambda_3 = (1-2z)F_E(x,y) \quad (11)$$

and the vertex is an evolutionarily stable strategy if and only if $\lambda_1<0$, $\lambda_2<0$, and $\lambda_3<0$.

Proof of Proposition 1. On the face $x=0$, the first equation reduces to $dx/dt=0$, and the derivative of dx/dt with respect to y carries the factor $x(1-x)$ and vanishes there; on $x=1$, the same factor vanishes. Hence, $\partial(dx/dt)/\partial y=\partial(dx/dt)/\partial z=0$ at every vertex, and the remaining off-diagonal entries vanish for the same reason. The Jacobian is diagonal, its eigenvalues are the diagonal entries, and the claim follows from Lyapunov's first method. Because the vertex is hyperbolic whenever the three entries are non-zero, local asymptotic stability and evolutionary stability coincide. □

3.2. Classification of the Eight Vertices

Proposition 1 reduces the classification to the signs of three affine functions. It is convenient to solve each of them for the parameter that determines its sign. Define

$$\beta_W(y,z) = \frac{\delta e_W - L_W - a_W(y,z)b_W}{\delta[1+\varphi\gamma(1-y)]D_W} \quad (12)$$

$$C_E(x,y) = p_z[K_W\delta(1-x) + K_S\gamma(1-y)]$$
$$+\psi\lambda(1-x,1-y)D_E + (B-g)[\delta(1-x)+\gamma(1-y)] \quad (13)$$

All denominators in (12) are positive; therefore, $F_W(y,z)>0$ if and only if $\beta>\beta_W(y,z)$, and $F_E(x,y)>0$ if and only if $C<C_E(x,y)$. Replacing each eigenvalue condition of Proposition 1 by the corresponding inequality in β or C yields Table 3, which classifies the eight pure-strategy profiles. Proposition 4 rules out interior states by a volume-preservation argument, and Lemma 1 applies the same argument to each face; apart from the edge-interior states discussed in Section 4.1, the eight vertices are therefore the only candidates for asymptotic stability.

Table 3 | Necessary and sufficient conditions for each pure-strategy profile to be an evolutionarily stable strategy, obtained by substituting Equations (12) and (13) into Proposition 1. With the baseline values of Table 1 and β=0.06 the thresholds are $\beta_W(0,0)=0.058$, $\beta_W(0,1)=0.053$, $\beta_W(1,0)=0.075$, $\beta_W(1,1)=0.067$, $\beta_S(0,0)=0.066$, $\beta_S(0,1)=0.020$, $\beta_S(1,0)=0.125$, $\beta_S(1,1)=0.038$, $C_E(0,0)=13.15$, $C_E(0,1)=6.65$, $C_E(1,0)=6.80$ and $C_E(1,1)=1.20$; the cost thresholds increase with β, which is why the cost boundaries of Sections 5 and 7 are slanted rather than horizontal.

Vertex (x,y,z)	$\lambda_1<0$	$\lambda_2<0$	$\lambda_3<0$
(0,0,0)	$\beta<\beta_W(0,0)$	$\beta<\beta_S(0,0)$	$C>C_E(0,0)$
(0,0,1)	$\beta<\beta_W(0,1)$	$\beta<\beta_S(0,1)$	$C<C_E(0,0)$
(0,1,0)	$\beta<\beta_W(1,0)$	$\beta>\beta_S(0,0)$	$C>C_E(0,1)$
(0,1,1)	$\beta<\beta_W(1,1)$	$\beta>\beta_S(0,1)$	$C<C_E(0,1)$
(1,0,0)	$\beta>\beta_W(0,0)$	$\beta<\beta_S(1,0)$	$C>C_E(1,0)$
(1,0,1)	$\beta>\beta_W(0,1)$	$\beta<\beta_S(1,1)$	$C<C_E(1,0)$
(1,1,0)	$\beta>\beta_W(1,0)$	$\beta>\beta_S(1,0)$	$C>C_E(1,1)$
(1,1,1)	$\beta>\beta_W(1,1)$	$\beta>\beta_S(1,1)$	$C<C_E(1,1)$

Only one row of Table 3 is empty. The state (1,0,1) requires $\beta>\beta_W(0,1)$ and $\beta<\beta_S(1,1)$, but the risk sensitivity raises the worker's incentive to be safe through the accident term and lowers the supervisor's incentive to be strict through the same term, so that at the calibrated values $\beta_S(1,1)=0.038<0.053=\beta_W(0,1)$ and no β satisfies both inequalities. The row (1,1,1) is not empty but requires $C<C_E(1,1)=1.20$, a regulation technology so cheap that the

firm would regulate even when nobody departs; it lies outside the calibrated range, and Corollary 1 records where it appears.

3.3. The Risk Thresholds and the Four Regimes

Substituting the baseline values into Table 3 shows that the ordering

$$0 < \beta_S(0,1) < \beta_W(1,1) < \beta_S(1,0) \tag{14}$$

holds, and that for $C_E(1,1)<C<\min\{C_E(0,1),C_E(1,0)\}$, that is, for $1.20<C<6.65$, the admissible set of Table 3 collapses to at most one vertex in each of four intervals of β. The result is stated as Proposition 2.

Proposition 2. *Let the parameters satisfy (14) and $1.20<C<6.65$. Then (9) has exactly one evolutionarily stable pure-strategy profile when $\beta \notin (\beta_W(1,1),\beta_S(1,0))$ and none when β lies in that interval: (0,0,1) for $0<\beta<\beta_S(0,1)$, (0,1,1) for $\beta_S(0,1)<\beta<\beta_W(1,1)$, and (1,1,0) for $\beta>\beta_S(1,0)$.*

Proof of Proposition 2. By Proposition 1, a vertex is stable exactly when the three inequalities of its row in Table 3 hold. Under (14) and $1.20<C<6.65$, the third inequality fails for four rows, for four different reasons: (0,0,0) needs $C>C_E(0,0)=13.15$, (0,1,0) needs $C>C_E(0,1)=6.65$, (1,0,0) needs $C>C_E(1,0)=6.80$, and (1,1,1) needs $C<C_E(1,1)=1.20$. The row (1,0,1) cannot be satisfied at all, because it needs $\beta>\beta_W(0,1)=0.053$ and at the same time $\beta<\beta_S(1,1)=0.038$. Three rows remain. For (0,0,1) the first two inequalities hold on $(0,\beta_S(0,1))$, for (0,1,1) on $(\beta_S(0,1),\beta_W(1,1))$ and for (1,1,0) on $(\beta_S(1,0),\infty)$; on the remaining interval $(\beta_W(1,1),\beta_S(1,0))$ no row satisfies all three. The thresholds are the numbers given in the caption of Table 3. □

Table 4 | The four regimes of Proposition 2 at the baseline values of Table 1 with C=4.5. Regimes I and II differ in whether the supervisor is active, regime III is analyzed in Section 4, and regime IV is the only one in which compliance survives the withdrawal of the firm.

Regime	β (calibrated)	Attracting state	Reading
I	$0<\beta<0.020$	(0,0,1)	Risk is imperceptible: crews take shortcuts, supervisors are lenient, and even permanent strong regulation cannot change the outcome
II	$0.020<\beta<0.067$	(0,1,1)	Supervision is awake and the firm regulates, yet crews still cut corners: the equilibrium rests on supervision alone
III	$0.067<\beta<0.125$	none	No pure strategy survives: supervision and regulation oscillate around each other
IV	$\beta>0.125$	(1,1,0)	Risk is salient: crews comply without coercion and the firm can withdraw from permanent strong regulation

Figure 1 shows the trajectories of three of the four regimes from three initial conditions. In each panel, the three populations reach the vertex predicted by Proposition 2 within a few learning periods, and the three initial conditions produce indistinguishable paths, as a pure-strategy attractor requires.

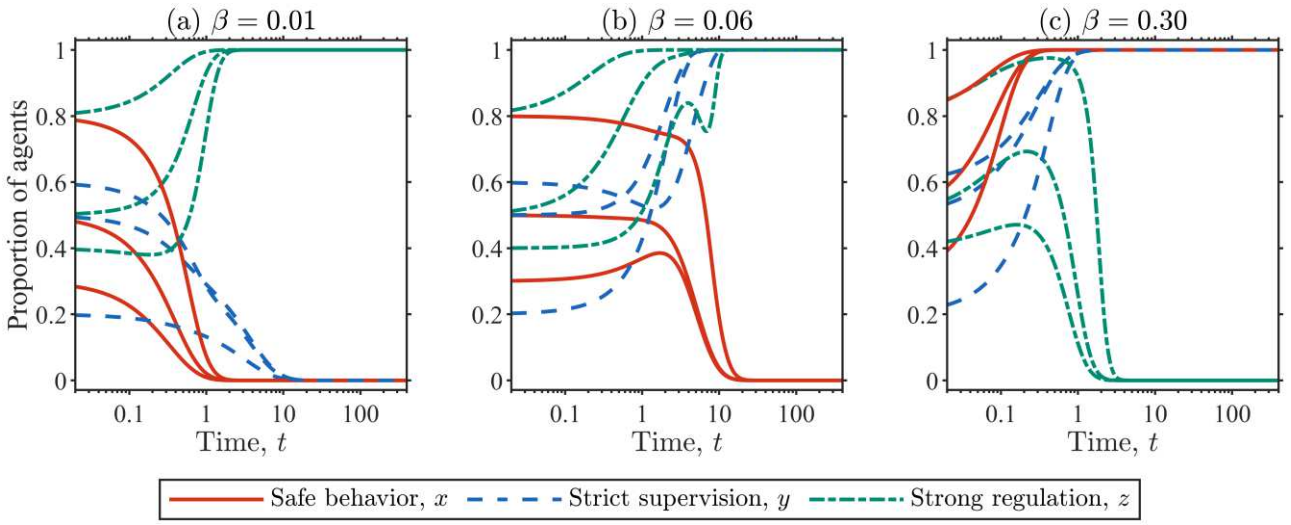

Figure 1 | Evolutionary trajectories of the three populations for three of the four risk regimes of Proposition 2: (a) β=0.01, (b) β=0.06 and (c) β=0.30, each started from (0.5,0.5,0.5), (0.3,0.2,0.8) and (0.8,0.6,0.4). The three initial conditions overlap in every panel.

Figure 2 transforms Proposition 2 into a bifurcation diagram, plotting the attracting vertex against β over the interval [0,0.4]. The diagram has three thresholds rather than one, and the interval labeled none between the second and third is the transition region analyzed in Section 4; its width, 0.058 at the calibrated values, is sufficiently large to be observed in a sequence of annual safety reports.

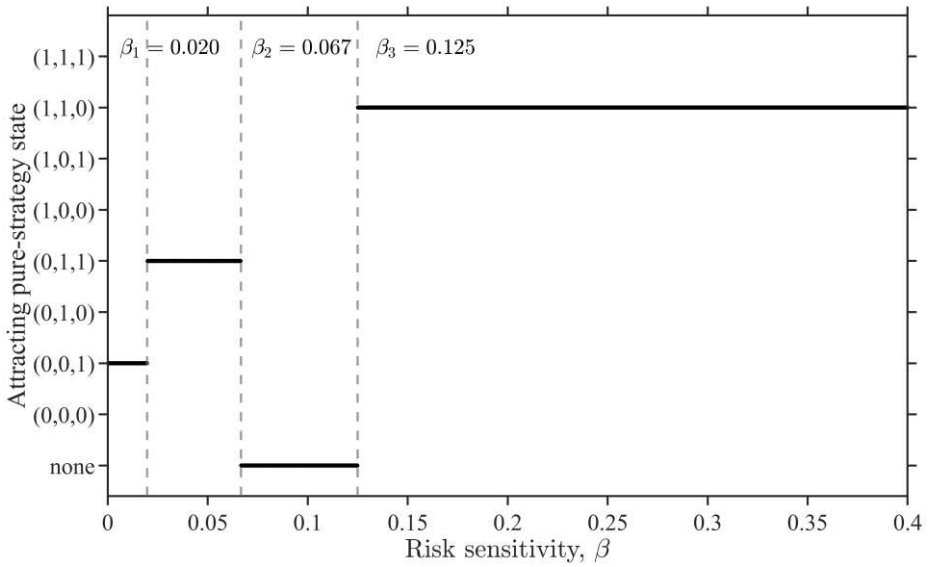

Figure 2 | Attracting pure-strategy state as a function of the risk sensitivity β with the other parameters at their baseline values. The vertical dashed lines mark $\beta_S(0,1)$=0.020, $\beta_W(1,1)$=0.067 and $\beta_S(1,0)$=0.125; the horizontal segment labeled none is the transition region of Proposition 2.

Table 4 summarizes the four regimes and the reading that the calibration supports.

The economic content of Proposition 2 is that the accident-risk sensitivity β, and not the penalty parameters, decides which of the four regimes is reached. The thresholds are not free; each is a ratio of the net benefit of the departure to its expected accident cost. $\beta_S(0,1)$ equates the supervisor's saving from leniency, $\gamma e_S - L_S - (s_0+s_z)b_S$, with the accident and liability exposure created by a lenient supervisor when the firm already regulates strongly, $\gamma D_S(1+\varphi\delta)+\eta\delta D_S$. In Regime I, the exposure is too small to make duty worth its

cost; in Regime IV, the worker's own exposure is large enough that compliance no longer needs to be bought.

3.4. Risk Compensation

The first derivative in (10) has a sign that depends on the parameters, and the sign carries a substantive warning about monitoring.

Proposition 3. *Let $\beta_{RC}=p_y b_W/(\delta\varphi\gamma D_W)$. If $\beta>\beta_{RC}$ then $\partial F_W/\partial y<0$: raising the intensity of supervision lowers the payoff advantage of safe behavior at the margin. With the baseline values $\beta_{RC}=0.025$, so the effect is present in regimes II to IV.*

Proof of Proposition 3. Differentiating F_W of (6) with respect to y yields $\partial F_W/\partial y=p_y b_W-\beta\delta\varphi\gamma D_W$, which is negative exactly when $\beta>p_y b_W/(\delta\varphi\gamma D_W)$. The numerical value follows from $p_y b_W=0.75$ and $\delta\varphi\gamma D_W=30$. □

The mechanism is the one described by Peltzman for automobile safety regulation[32]: an instrument that lowers the accident probability also lowers the private cost of carelessness, and the two effects partly cancel. In the present model the cancellation is exact at the margin: an increase in y raises detection by p_y and lowers the accident term by $\beta\delta\varphi\gamma D_W$. The practical consequence is that supervision which is not matched by an increase in detection certainty can make the crew bolder rather than safer. Proposition 3 is a statement about a derivative, not about the equilibrium: the comparative statics of Section 3.2 show that raising β still moves the system toward compliance, because it operates on all three thresholds at once. What the proposition rules out is the reading of supervision as a monotone instrument.

The remaining derivatives in (10) have no such ambiguity. $\partial F_S/\partial x<0$ identically: a supervisor who expects the crew to be safe has less reason to be present, because both the accident exposure and the detection return are lower. $\partial F_E/\partial x<0$ and $\partial F_E/\partial y<0$: the firm regulates less when the other two layers are already doing the work. The signs of the remaining derivatives describe a chain in which each layer substitutes for the next, and they explain why the equilibria of Section 3.2 are corner solutions rather than interior compromises.

4. Dynamics in the Transition Region

4.1. No Interior Attractor

Regime III of Proposition 2 has no pure-strategy equilibrium, and the natural question is what the flow does there. The replicator system (9) answers this exactly: in the coordinates that linearize the logistic factor, the flow preserves volume, and this rules out every attractor except on the boundary.

Lemma 1. *On each face of the cube the induced flow is area preserving in the corresponding logit coordinates. Consequently no state in the relative interior of a face is asymptotically stable, and every evolutionarily stable state is either a vertex or lies on an edge.*

Proof of Lemma 1. On the face $x=1$, we put $v=\ln[y/(1-y)]$ and $w=\ln[z/(1-z)]$; then $dv/dt=F_S(1,z)$ and $dw/dt=F_E(1,y)$, and the divergence $\partial F_S/\partial w+\partial F_E/\partial v$ vanishes because F_S does not depend on y and F_E does not depend on z. The same computation applies to the faces $y=1$ and $z=1$. An asymptotically stable equilibrium of an area-preserving planar flow would require a neighborhood whose area contracts, which is impossible, and an evolutionarily stable state is asymptotically stable. Edge-interior states would require an exact indifference condition such as $F_S(1,0)=0$, which holds only on a measure-zero set of parameters and does not hold at the calibrated values. □

Proposition 4. *Let $u=\ln x/(1-x)$, $v=\ln y/(1-y)$ and $w=\ln z/(1-z)$ on the open cube. In these coordinates the vector field of (9) is (F_W,F_S,F_E) and its divergence vanishes identically. Consequently (i) the flow preserves the measure du dv dw, (ii) no interior equilibrium is asymptotically stable, and (iii) no proper invariant set of the open cube has a basin of positive measure.*

Proof of Proposition 4. Since $du/dt=(dx/dt)/(x(1-x))=F_W(y,z)$ and cyclically, the field in (u,v,w) is $(F_W(y,z),F_S(x,z),F_E(x,y))$. Its first component does not depend on u, the second does not depend on v and the third does not depend on w; hence, $\nabla\cdot(F_W,F_S,F_E)=0$. Liouville's theorem then gives measure preservation, and an asymptotically stable equilibrium would require a neighborhood whose volume contracts to zero, which is impossible. A hyperbolic interior equilibrium has eigenvalues summing to the trace, which is zero; hence, it is a center or a saddle, and in either case, not attracting; a non-hyperbolic equilibrium with zero trace is likewise not attracting. □

Proposition 4 states that permanent convergence to a mixed configuration is impossible: a trajectory that starts inside the cube cannot converge to an interior point and either remains on a bounded invariant set in the interior or approaches the boundary. The numerical experiments in Section 7 confirm the second alternative in regime III, where the trajectories reach the face $x=1$ and then circulate on it. The mathematics of this face is the subject of the next subsection. The result also explains why the applied literature, which often reports a single interior stable point for three-population safety games, must introduce an additional payoff asymmetry to obtain one: without it, the replicator structure forbids the outcome.

4.2. The Invariant Face and Its First Integral

Suppose that $\beta>\beta_W(1,1)$. Since $\partial F_W/\partial y<0$ for $\beta>\beta_{RC}$ by Proposition 3, $F_W(y,z)\geq F_W(1,z)>0$ for every (y,z), and the face $x=1$ is attracting from the interior. On that face, the dynamics reduce to two equations, and because F_S does not depend on y and F_E does not depend on z, the reduction is separable. Write

$$A(z)=F_S(1,z)=A_0+A_1z,$$
$$A_0=L_S+s_0b_S+\beta\gamma D_S-\gamma e_S,\ \ A_1=s_zb_S \tag{15}$$

$$B(y)=F_E(1,y)=B_1(1-y)-B_0,$$
$$B_1=\gamma[p_zK_S+B-g+\psi\beta D_E],\ \ B_0=C-\psi\lambda_0D_E \tag{16}$$

where $dy/dt=y(1-y)A(z)$ and $dz/dt=z(1-z)B(y)$. The interior equilibrium of the reduction is $z^*=-A_0/A_1$ and $y^*=1-B_0/B_1$, which lies in the open square when $-A_1<A_0<0$ and $0<B_0<B_1$; at the baseline values, these conditions read $0.0375<\beta<0.125$ and $1.20<C<5.0+30\beta$.

Proposition 5. *Let $-A_1<A_0<0$ and $0<B_0<B_1$. Then*

$$H(y,z)=(B_1-B_0)ln\ y+B_0ln(1-y)$$
$$-A_0ln\ z+(A_0+A_1)ln(1-z) \tag{17}$$

is the first integral of the reduction on the open face $x=1$: H is constant along trajectories. Its unique critical point in the square is (y^*,z^*), the Hessian there is negative definite, and consequently every orbit that stays in the open square is a closed level curve of H around (y^*,z^*); orbits that reach $y=0$ or $z=0$ leave the face. The equilibrium is a nonlinear center.

Proof of Proposition 5. The partial derivatives of (17) are $\partial H/\partial y=[B_1(1-y)-B_0]/(y(1-y))=B(y)/(y(1-y))$ and $\partial H/\partial z=[-A_0(1-z)-(A_0+A_1)z]/(z(1-z))$; the second simplifies to $-[A_0+A_1z]/(z(1-z))=-A(z)/(z(1-z))$. Hence, along a trajectory, $dH/dt=B(y)/(y(1-y))\ y(1-y)A(z)-A(z)/(z(1-z))\ z(1-z)B(y)=0$. The critical point solves $A(z^*)=0$ and $B(y^*)=0$, which gives the stated coordinates. The second derivatives at the critical point are $\partial^2H/\partial y^2=-(B_1-B_0)/y^{*2}-B_0/(1-y^*)^2<0$ and $\partial^2H/\partial z^2=-A_1/z^*-A_1/(1-z^*)<0$, with vanishing cross derivative; the Hessian is therefore negative definite and H has a strict local maximum. Level sets of a smooth function with a nondegenerate maximum are closed curves surrounding it, and since H is conserved, they are invariant. □

The first integral is the mathematical form of a familiar administrative cycle. Because $A_0<0$ and $A_1>0$, $A(z)$ increases in z: a firm that regulates strongly raises the return to strict supervision, which is why supervision rises after a crackdown. Because $B_1>0$ and $B_0>0$, $B(y)$ decreases in y: a supervisor who is strict lowers the return to regulation, which is why the firm relaxes once the record improves. The two monotone responses form a negative feedback loop, and the level sets of H are nested around the equilibrium. The amplitude of the cycle is set by initial conditions, not by the parameters; therefore, two plants with identical economics but different histories can display permanently different amplitudes.

4.3. Center Coordinates and the Period

The linearization of the reduction at (y^*,z^*) has eigenvalues $\pm i\omega$ with

$$\omega^2 = y^*(1-y^*)A_1 z^*(1-z^*)B_1 \tag{18}$$

so that small oscillations have a period $T_0=2\pi/\omega$. Table 5 reports the center and period for five values of the risk sensitivity in regime III.

Table 5 | Center of the reduction on the face x=1 and the period of small oscillations, computed from Equations (15), (16) and (18) with the baseline values of Table 1. The center migrates toward the boundary z=0 as β approaches $\beta_S(1,0)=0.125$, and the period diverges there because the center reaches the boundary of the admissible square.

β	$y^\bullet$	$z^\bullet$	ω	$T_0=2\pi/\omega$
0.07	0.441	0.629	0.771	8.15
0.08	0.468	0.514	0.821	7.65
0.09	0.492	0.400	0.826	7.61
0.10	0.515	0.286	0.779	8.07
0.12	0.554	0.057	0.415	15.13

Figure 3 shows the corresponding phase portraits. Each panel contains three orbits started at different distances from the center, and in each, the orbits are closed and nested, as Proposition 5 requires. At $\beta=0.12$, the center is close enough to the boundary $z=0$ that the largest orbit is cut by the boundary; this is the geometric counterpart of the period growth in the last row of Table 5, where T_0 tends to infinity as $\beta\to0.125$.

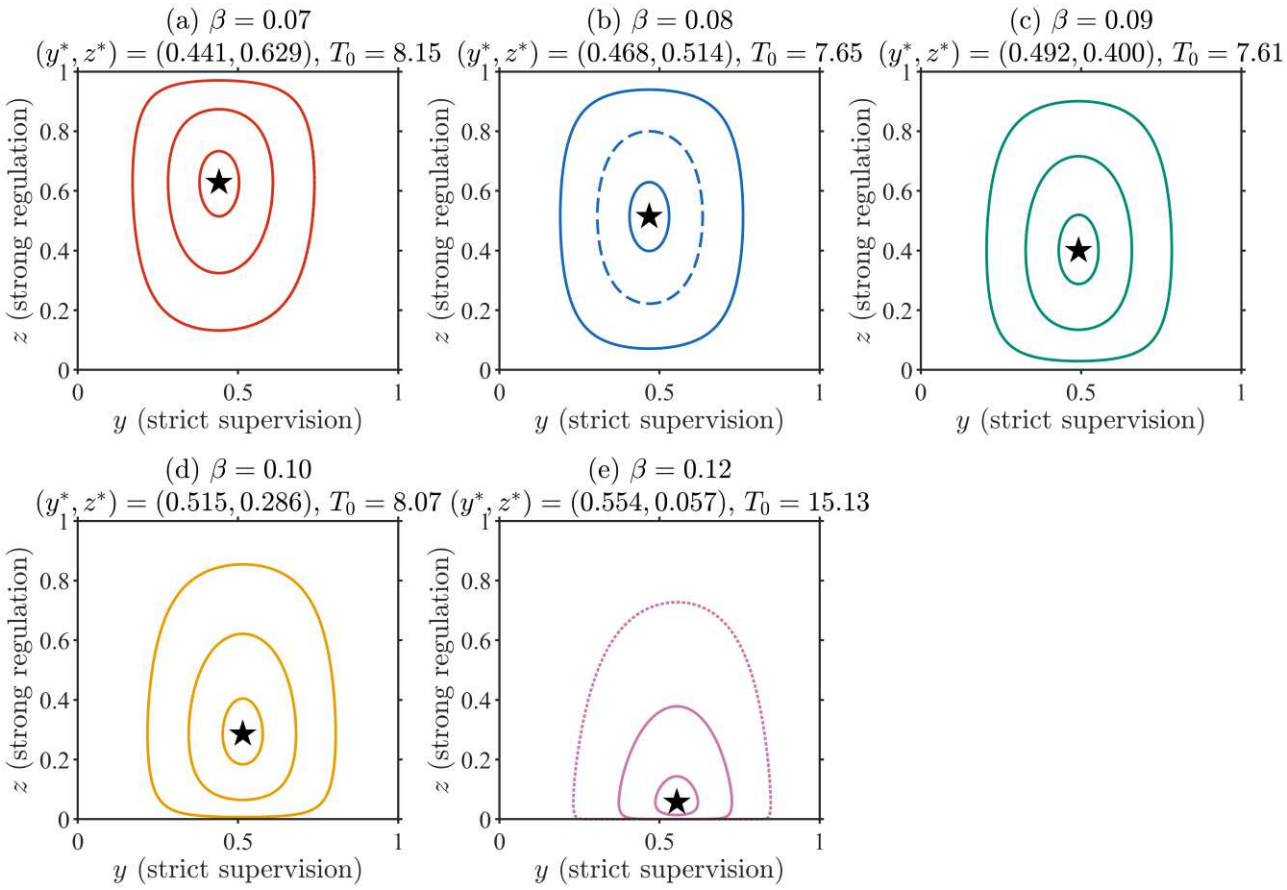

Figure 3 | Phase portraits of the reduction on the face x=1 for five values of β. In each panel the star marks the analytic center (y∗,z∗) of Table 5 and the three curves are orbits started at different distances from it.

Figure 4 shows the same dynamics as a time series at β=0.09. Compliance reaches one within two learning periods, while supervision and regulation oscillate with a period of 7.6, in agreement with Table 5. The period is a property of the parameters; the amplitude is a property of the initial condition because H is conserved along the orbit.

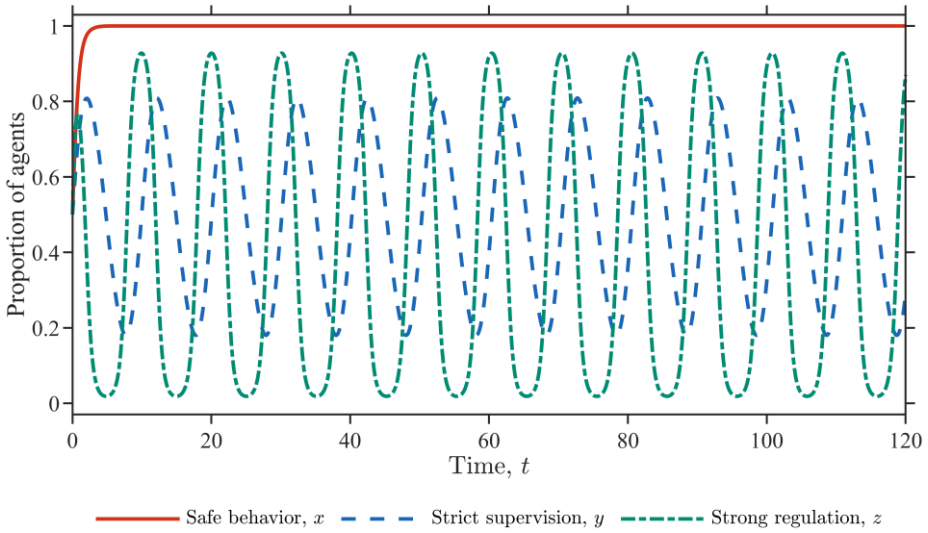

Figure 4 | Crackdown-relapse cycle on the invariant face x=1 at β=0.09. The crew's compliance level reaches one after a transient of about two learning periods, while supervision and regulation oscillate with the period T0=7.6 predicted by Equation (18).

Table 5 carries two features that matter for measurement. As risk becomes more salient, y^* increases and z^* decreases: the crew needs less supervision and the firm needs less regulation, but the two movements are not symmetric, and the center reaches the boundary $z=0$ exactly when β reaches $\beta_S(1,0)=0.125$. The period itself is not monotone. It falls slightly between $\beta=0.07$ and $\beta=0.09$ and then rises sharply as the center approaches the boundary because the product $z^*(1-z^*)$ vanishes. An observer who measured only the period of the crackdown-relapse cycle would therefore infer the wrong value of β at the ends of the interval.

For orbits of finite amplitude, the period exceeds T_0 and increases with the amplitude, a property inherited from the first integral: the level curves of H are elongated in the direction in which the restoring force is weak. At $\beta=0.09$, the period rises from 7.64 for an amplitude of 0.02 to 11.16 for an orbit that reaches the boundary, against $T_0=7.61$.

5. Costly Regulation and Bistability

5.1. The Cost Thresholds

Proposition 2 was stated for $1.20<C<6.65$, the range that holds at $\beta=0.06$, and both endpoints are thresholds: the lower one is $C_E(1,1)=1.20$ and the upper one is $C_E(0,1)$, which equals 6.65 at $\beta=0.06$ and rises by 30 for each unit of β. The four thresholds are listed in Table 6. Varying C reorganizes the classification of Table 3, and because C is the parameter a utility controls most directly, the reorganization is the policy-relevant part of the model.

Corollary 1. *At the baseline values of Table 1 the following hold. (i) For $1.20<C<C_E(0,1)$, which equals 6.65 at $\beta=0.06$, the attracting state is the one described in Proposition 2. (ii) At $C=C_E(0,1)$ the state (0,1,1) disappears; for $0.066<\beta<0.075$ the state (0,1,0) then exists, and it is the only attractor only on the narrow strip $C_E(0,1)<C<C_E(1,0)$. (iii) At $C=C_E(1,0)$, which equals 6.80 at $\beta=0.06$, the state (1,0,0) appears, and above that level the two states (0,1,0) and (1,0,0) coexist for $0.066<\beta<0.075$. (iv) At $C=C_E(0,0)$, which equals 13.15 at $\beta=0.06$, the state (0,0,0) appears for $\beta<0.058$. (v) Below $C=C_E(1,1)=1.20$ the state (1,1,1) replaces (1,1,0) for $\beta>0.067$.*

Proof of Corollary 1. The four cost thresholds are the values of C_E at the argument pairs (1,1), (0,1), (1,0), and (0,0), evaluated with the parameter values of Table 1. Substituting them into Table 3 and repeating the bookkeeping of Proposition 2 yields the five statements; the coexistence in (iii) follows because the row (0,1,0) requires $\beta<\beta_W(1,0)=0.075$ and $C>6.65$, while the row (1,0,0) requires $\beta>\beta_W(0,0)=0.058$ and $C>6.80$. □

The interpretation of the thresholds is asymmetric. Raising C does not simply weaken regulation; it removes the only state in which all three populations do their part and replaces it with states in which governance is delegated. In (0,1,0), the firm has withdrawn from regulation, and the supervisor is the sole barrier to an accident, so that compliance depends entirely on the diligence of one agent who bears no accident loss beyond D_S. In (1,0,0), the crew works safely, but no supervisory layer is active: compliance is sustained by the crew's own exposure to risk, and any shock that lowers the perceived exposure removes the only reason to comply. Neither state is a governance success, although the second looks like one if only the compliance rate is observed.

Table 6 | Cost thresholds of Corollary 1 and the changes they produce in the attracting set, evaluated with the baseline values of Table 1.

Threshold	Expression	Value	What changes at the threshold
$C_E(1,1)$	$\psi\lambda_0 D_E$	1.20	Strong regulation survives when nobody departs
$C_E(0,1)$	$p_z K_W\delta+(B-g)\delta+\psi\lambda(1,0)D_E$	6.65	The central state (0,1,1) disappears
$C_E(1,0)$	$p_z K_S\gamma+(B-g)\gamma+\psi\lambda(0,1)D_E$	6.80	The self-regulating state (1,0,0) appears
$C_E(0,0)$	$p_z(K_W\delta+K_S\gamma)+(B-g)(\delta+\gamma)+\psi\lambda(1,1)D_E$	13.15	The all-violation state (0,0,0) appears

5.2. Two Delegated-Governance States

The coexistence in part (iii) of Corollary 1 is not a technicality; it is the only place in the parameter space where the model predicts that two different governance configurations are both stable for the same economy. Which one is reached depends on where the system starts, and Figure 5 shows the basins at β=0.07 and C=8.0, the point at which the two windows overlap.

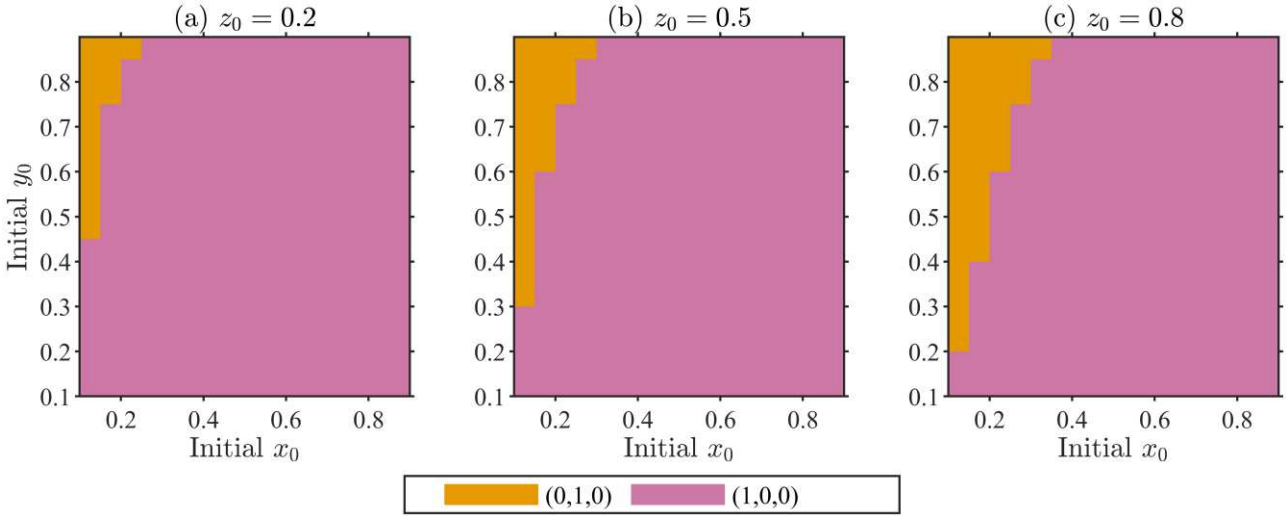

Figure 5 | Basins of attraction at β=0.07 and C=8.0 for three initial values of the regulation share: (a) z0=0.2, (b) z0=0.5 and (c) z0=0.8. Orange: convergence to (0,1,0), in which supervision compensates for risky work. Pink: convergence to (1,0,0), in which the crew self-regulates while supervision is dormant. The boundary shifts toward higher initial compliance as z0 increases.

The basin of (0,1,0) is small and shrinks as the initial share of strong regulation rises: over the 289 initial conditions of each panel, 94%, 91%, and 86% converge to (1,0,0) for z_0=0.2,0.5,0.8, and the remainder converge to (0,1,0). The boundary itself is not a level set of a single variable. A task force that starts with high compliance but low supervision can still end in the supervision-compensating state, because the trajectory crosses the boundary while the crew's behavior is evolving, which is why the three panels differ by more than the initial condition alone suggests.

For calibration purposes, the practical message is that the same utility can observe two different equilibria in two of its units without either unit being irrational. The observation that one substation shows disciplined crews and inattentive supervisors while another shows the reverse is, in this model, a consequence of history rather than of differences in the safety culture of the two sites.

6. Hazard Backlog and the Closure Cycle

The model thus far assumes that the consequence of an unsafe act is felt at the moment of the act. In practice, the consequence accumulates: a defect that is not rectified remains in the equipment, and the accident probability of a task depends on the stock of unresolved hazards as well as on the behavior observed during the task. This section adds that stock and shows that the addition turns the neutral center of Section 4 into a dissipative system whose cycle survives a wide range of closure rates.

Let $S(t)\geq 0$ be the backlog of unresolved hazards, measured in units of one accident probability. Hazards are created by departures and closed at a rate that depends on regulation:

$$\frac{dS}{dt} = \alpha[\delta u + \gamma v + \varphi\delta\gamma uv] - (\kappa_0 + \kappa_1 z)S \tag{19}$$

where $\alpha>0$ is the hazard-generation rate per unit of departure, $\kappa_0>0$ the natural closure rate, $\kappa_1\geq 0$ the marginal effect of strong regulation on closure, and $\sigma\geq 0$ the contribution of the backlog to the accident probability. The experiments reported in this section use α=1.2, σ=0.12 and κ_1=0.5, with κ_0 varied between 0.03 and 6 learning periods. The mean closure time is $\tau=1/(\kappa_0+\kappa_1 z)$, which is the quantity an operator observes as the average age of an open defect.

The extended system has four state variables, and the face x=1 remains invariant when $\beta>\beta_W(1,1)$. On that face, with $w=\gamma(1-y)$, the reduction is

$$\frac{dy}{dt} = y(1-y)A(z),$$

$$\frac{dz}{dt} = z(1-z)[B_1(1-y) - B_0 + \psi\sigma D_E S] \tag{20}$$

Its divergence in (y,z,S) is $-(\kappa_0+\kappa_1 z)<0$, so the volume-preserving property of Proposition 4 is lost as soon as the backlog is added: the lag makes the system dissipative. The interior equilibrium of (20) is

$$z^* = -\frac{A_0}{A_1},\ 1-y^* = \frac{B_0 - \psi\sigma D_E S^*}{B_1},\ S^* = \frac{\alpha\gamma(1-y^*)}{\kappa_0 + \kappa_1 z^*} \tag{21}$$

Proposition 6. *The interior equilibrium (21) is locally asymptotically stable if and only if*

$$\kappa_1 S^*(\kappa_0 + \kappa_1 z^*) > y^*(1-y^*)A_1\alpha\gamma \text{ iff } \kappa_1 > y^*(\kappa_0)A_1 \tag{22}$$

and at these values the inequality fails for every κ_0, so the equilibrium is unstable throughout the admissible range, and the attractor of (20) is a relaxation oscillation.

Proof of Proposition 6. Writing $a=y^*(1-y^*)$, $b=z^*(1-z^*)$, $c=\kappa_0+\kappa_1 z^*$ and $E=\psi\sigma D_E$, the Jacobian of (20) at (21) is $[[0,aA_1,0],[-bB_1,0,bE],[-\alpha\gamma,-\kappa_1 S^*,-c]]$. Expanding the determinant of $J-\lambda I$ gives the cubic $\lambda^3+c\lambda^2+(bE\kappa_1 S^*+abA_1B_1)\lambda+abA_1(B_1c+E\alpha\gamma)=0$. All coefficients are positive; the Routh-Hurwitz condition reduces to $\kappa_1 S^* c > y^*(1-y^*)A_1\alpha\gamma$. Substituting $S^*c=\alpha\gamma(1-y^*)$ and cancelling $\alpha\gamma(1-y^*)>0$ gives $\kappa_1>y^*A_1$. At β=0.09 and C=3.0, the values used in Figure 6, A_1=1.75 and $y^*(\kappa_0)\in[0.75,0.93]$ for $\kappa_0\in[0.03,6]$, so the condition would require κ_1 between 1.31 and 1.63 whereas the calibrated value is κ_1=0.5; the inequality fails for every κ_0 in the range. □

The instability is informative rather than disappointing. Proposition 4 showed that the cycle of Section 4 is neutral; therefore, its amplitude is a property of the initial condition and cannot be predicted from the parameters. Adding the backlog removes this degeneracy: the cycle becomes an attractor, its amplitude is determined by the parameters, and the closure rate κ_0 controls the amplitude through the backlog rather than through the position of the center. Figure 6 illustrates the mechanism.

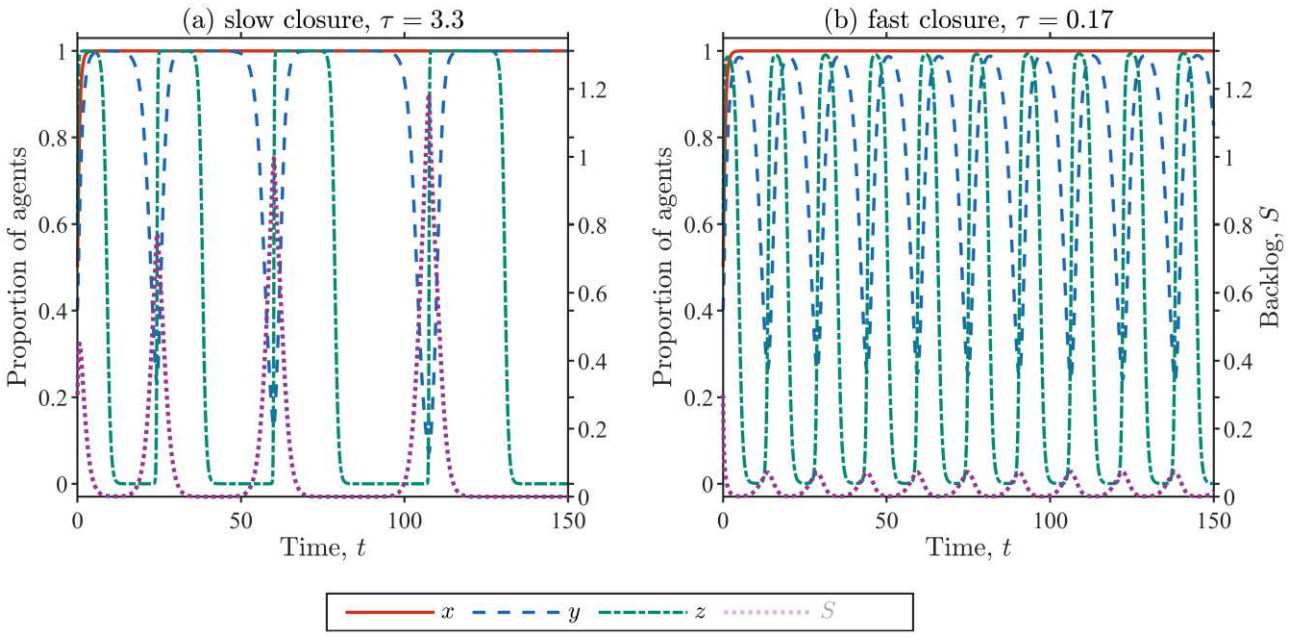

Figure 6 | Effect of the hazard-closure cycle on the extended system at β=0.09: (a) slow closure, τ=3.3; (b) fast closure, τ=0.17. Faster closure shrinks the backlog and shortens the excursions of the supervision share, but it does not remove the crackdown-relapse cycle.

The numerical amplitude of the backlog falls from 1.47 at κ_0=0.03 to 0.089 at κ_0=6, a factor of sixteen, while the amplitude of the regulation share remains at one in both cases. The managerial reading is therefore narrower than the one usually offered: accelerating the closure of defects reduces the stock of latent risk and the amplitude of the supervisory cycle, but it does not by itself produce a stable configuration. A stable configuration requires the risk sensitivity to be moved out of the transition region, which is the subject of Section 7.

7. Calibration and Numerical Analysis

7.1. Calibration Protocol

The propositions above are qualitative in β and C and quantitative in the thresholds; therefore, the numerical section has two tasks: to verify the classification and to establish that the conclusions survive uncertainty in the 22 parameters other than β, φ and λ_0. Both tasks require a calibration protocol because a three-population game with arbitrary numbers invites the objection that the thresholds are artifacts of the values chosen.

The protocol rests on three rules. All monetary quantities are expressed in units of the worker's expected loss D_W, so that D_W=120 is a normalization and the remaining monetary parameters are ratios that can be measured directly. Every parameter is bound to an observable proxy and assigned to one of five evidence classes: A, statutes and mandatory standards; B, industry statistics and accident investigation reports; C, firm records and production systems; D, questionnaires and expert elicitation; E, published estimates. Parameters supported only by classes D and E are flagged, and we measure their influence in the global analysis of Section 7.4 rather than assume it away. Table 7 presents the resulting assignment.

Table 7 | Calibration of the model. A: statutes and standards, including GB 26859-2011 and the 2021 Work Safety Law[1,2]; B: industry statistics and accident investigations; C: firm records; D:

questionnaire or expert elicitation; E: published estimates. Monetary parameters are expressed in units of DW.

Parameter	Observable proxy	Class	Baseline
δ	Share of work permits with a recorded violation	C	0.5
γ	One minus the on-site supervision coverage rate	C	0.5
e_W	Time saved by the shortcut times the hourly cost	C, D	15
e_S	Extra hours of strict duty plus the social-cost scale	C, D	14
λ_0	Accidents per maintenance task, residual	B	0.02
β	Marginal effect of a violation on the accident odds	B	0.01/0.06/0.30
φ	Joint-to-single-factor accident rate ratio	B	1.0
D_W	Weighted injury and fatality compensation	A	120
D_S	Fines, performance loss and career effect for managers	A, C	40
D_E	Fines, outage cost, repair and reputation	A, C	200
p_0	Video coverage times recognition accuracy plus reports	C	0.10
p_y	Detection rate with supervision minus without	C	0.30
p_z	Detection increment of unannounced inspection	C	0.20
s_0	Residual detection rate of lenient duty	C	0.10
s_z	Detection increment of lenient duty under inspection	C	0.35
K_W	Fine per recorded violation	C	1.5
K_S	Deduction per recorded supervisory failure	C	3
L_W	Stated importance of a clean safety record	D	2
L_S	Stated career cost of a supervisory failure	D	4
θ	Reputation multiplier upon notification	D	1.5
η	Share of management liability borne by the supervisor	A, C	0.4
B	Parent-company and regulator sanction per tolerated departure	A, C	10
ψ	Fall in the accident rate after hazard rectification	C, E	0.3
g	Value of the schedule lost by strict regulation	C	3
C	Incremental safety expenditure per task	A, C	4.5

Two parameters are deliberately not calibrated. The risk sensitivity β is the scenario variable of the paper and varies over [0,0.4], because its value differs by an order of magnitude between a mature transmission utility with a strong reporting culture and a contractor-operated distribution unit. The interaction coefficient φ is fixed at one; the propositions hold for every $\varphi \geq 0$, and because φ only scales the interaction term of (1), changing it shifts the thresholds monotonically without changing the classification.

7.2. Verification of the Classification

At β=0.06, the eight vertices have the eigenvalue triples of Proposition 1, and only (0,1,1) has three negative entries, with $(\lambda_1,\lambda_2,\lambda_3)$=(−0.40,−1.53,−2.15). At β=0.01 the stable vertex is (0,0,1) with (−3.85,−0.37,−4.90), and at β=0.30 it is (1,1,0) with (−13.50,−3.50,−3.30). We therefore recovered the four regimes of Proposition 2, and the three transitions occur at the analytic thresholds 0.020, 0.067, and 0.125 rather than at grid points, which is why Figure 2 uses a step of 4×10^{-4} in β.

The transition region behaves as predicted in Section 4. For $\beta \in (0.067, 0.125)$, no vertex has three negative eigenvalues, and in integrations from 27 initial conditions, every trajectory reaches the face $x=1$, with the two remaining coordinates circulating around the analytic center of Table 5. The invariance of the face is exact in the integration because $F_W(y,z)>0$ on it; the observed drift of x below one is zero to machine precision, as it must be when dx/dt vanishes identically on the face.

7.3. Two-Parameter Maps

Figures 7 and 8 report the attracting state over two-parameter slices. We computed the maps by evaluating the eight vertex conditions of Table 3 on grids of 90×90 (Figure 7) and 70×70 (Figure 8); therefore, they are exact up to the resolution of the grid and do not depend on any integration.

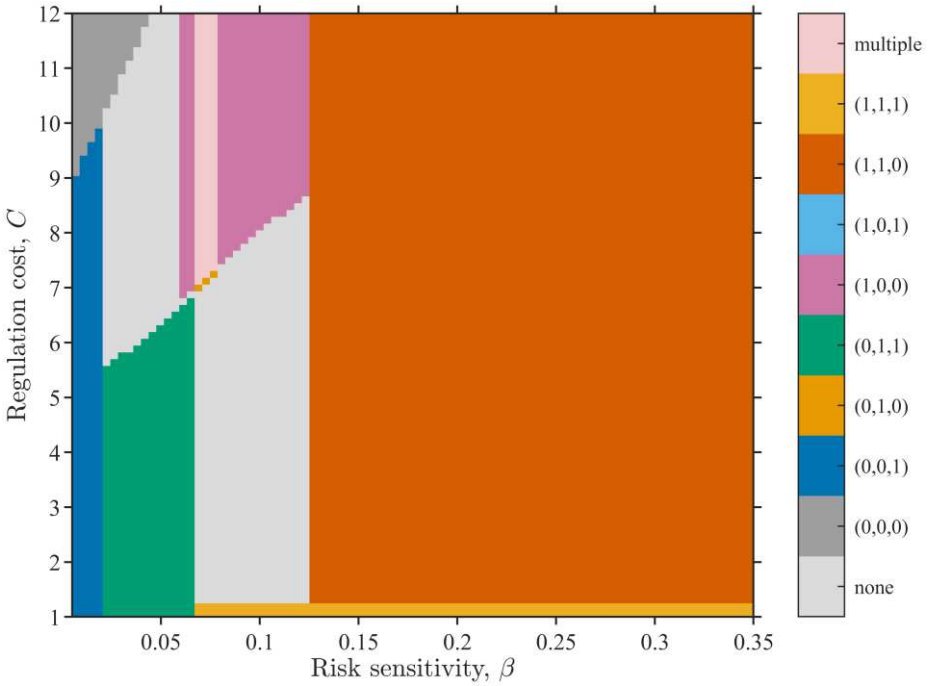

Figure 7 | Attracting state in the (β,C) plane with the other parameters at their baseline values. The central region (0,1,1) is bounded on the left by $\beta_S(0,1)=0.020$, on the right by $\beta_W(1,1)=0.067$ and above by the slanted line $C=C_E(0,1)$, which stands at 6.65 when β=0.06; to its right lies the transition region and then the compliant state (1,1,0); above it, the firm withdraws and the self-regulating state (1,0,0) appears.

Figure 7 makes the structure of Corollary 1 visible. The green region (0,1,1) is bounded on the left by $\beta_S(0,1)=0.020$, on the right by $\beta_W(1,1)=0.067$, and above by the line on which $C=C_E(0,1)$; because the accident term of C_E grows with β, that upper boundary is slanted rather than horizontal, and it stands at 6.65 when $\beta=0.06$. To its right the purple state (1,0,0) appears as soon as β exceeds $\beta_W(0,0)=0.058$ and the cost passes $C_E(1,0)$. Where both cost conditions hold, the pink band $\beta \in (0.066, 0.075)$ carries the two stable states (0,1,0) and (1,0,0); its width in β is 0.009, and it is the bistability of Corollary 1. To the right of $\beta=0.125$, the orange state (1,1,0) fills the plane except for the strip $C<1.20$, where (1,1,1) survives.

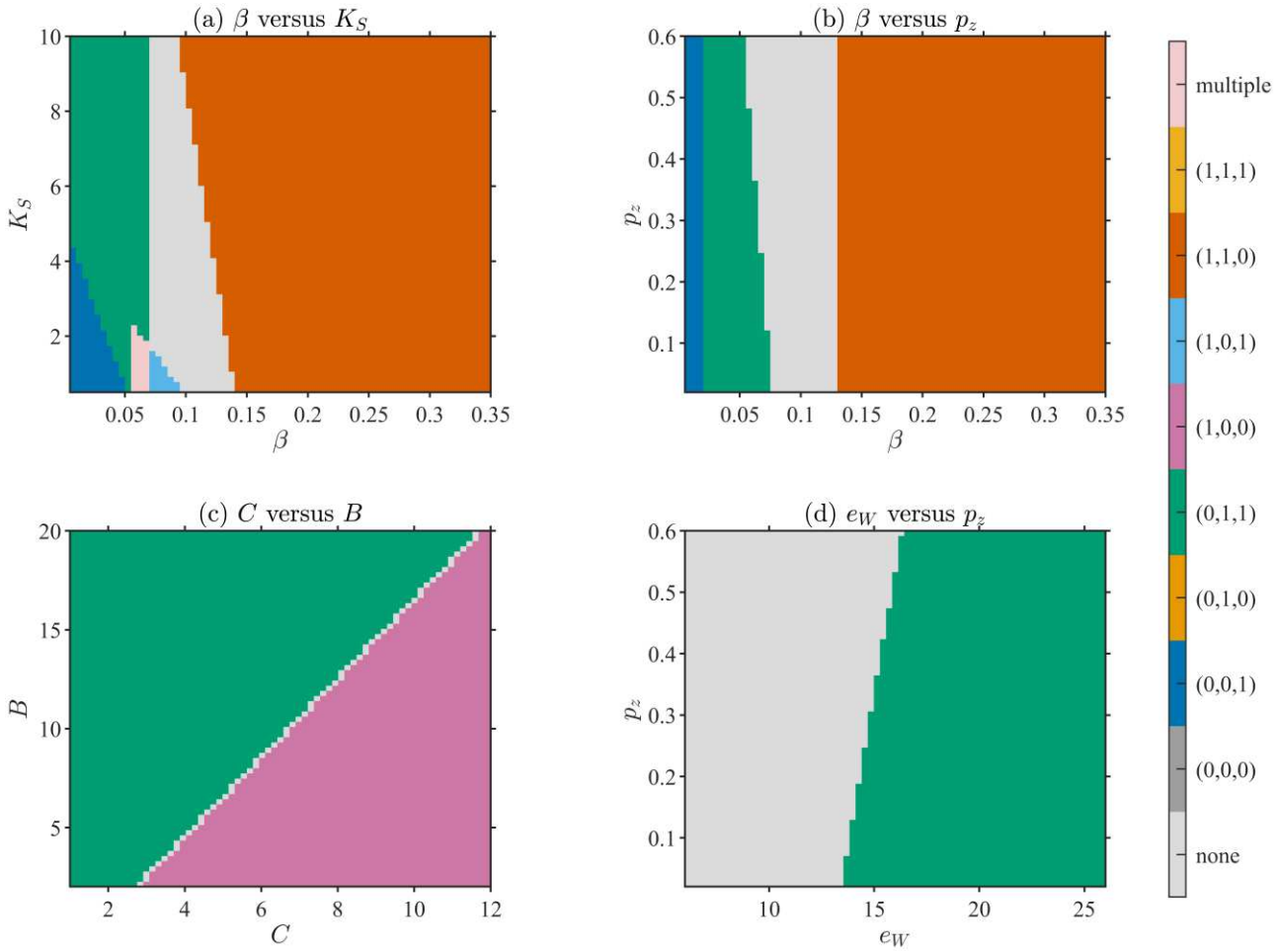

Figure 8 | Attracting state on four two-parameter slices: (a) β versus KS, (b) β versus pz, (c) C versus B and (d) eW versus pz. Panel (a) shows that the supervisor's fine barely moves the boundary of the central region, panel (b) that a higher detection rate shifts the compliance threshold to lower β, and the diagonal in (c) that regulation cost and accountability are substitutes.

Figure 8 is the most policy-relevant map. Panel (a) varies the supervisor's fine K_S over more than an order of magnitude and leaves the boundary of the central region unchanged, a property that is proved in Section 7.5. Panel (b) shows the opposite for the firm's detection rate: the central region narrows as p_z rises, because $\beta_W(1,1)$ falls from 0.071 at p_z=0.10 to 0.050 at p_z=0.60, while the region (0,0,1) is unaffected, since $\beta_S(0,1)$ does not depend on p_z. Panel (c) shows that C and B are substitutes along a diagonal: a firm that faces stronger accountability can afford more expensive regulation and still remain in the central state. Panel (d) shows the interaction between the crew's effort saving and the firm's detection: the central region requires e_W>13.7 at p_z=0.10 and e_W>16.2 at p_z=0.60, so the threshold rises by less than a fifth across the panel.

7.4. Global Uncertainty and Sensitivity

The calibration of Table 7 assigns six of the 25 parameters to classes D and E, and the values of those parameters are the weakest link in any quantitative claim. The response adopted here is a global uncertainty analysis rather than a defense of point values. We sample each of the 22 parameters other than β, φ and λ_0 over ±30% of its baseline by Latin hypercube sampling (LHS) with N=5000 draws[33], evaluate the stability margin of the target equilibrium for every draw, and compute partial rank correlation coefficients (PRCC) between the margin and each parameter[34]. The stability margin is m=min $\{-\lambda_1,-\lambda_2,-\lambda_3\}$ at the target vertex, so that m>0 reproduces the ESS condition of Proposition 1 and m is

continuous in the parameters, which the rank correlation requires. Table 8 reports the outcome.

Table 8 | Sensitivity of the two calibrated targets over ±30% of the 22 non-scenario parameters, N=5000 Latin hypercube draws. The columns report the share of draws in which the target is stable, in which it is the only stable vertex, in which no vertex is stable, and in which two or more vertices are stable.

Scenario	*P*(target is ESS)	Unique ESS	No ESS	Two or more ESS
β=0.06, target (0,1,1)	41.2%	77.5%	18.2%	4.3%
β=0.30, target (1,1,0)	97.5%	97.7%	2.3%	0.0%

The two scenarios behave very differently, and the difference is itself a result. Regime IV holds: in 97.5% of the draws, the target (1,1,0) remains a stable state, so a policy that raises β beyond 0.125 does not need precise values for the other parameters. Regime II does not: its target survives in only 41.2% of the draws, and in 18.2% of them, no vertex is stable at all. The fragility is not a defect of the calibration but a property of the regime, and it is consistent with the narrowness of the interval ($\beta_S(0,1)$,$\beta_W(1,1)$)=(0.020,0.067) in which the state exists.

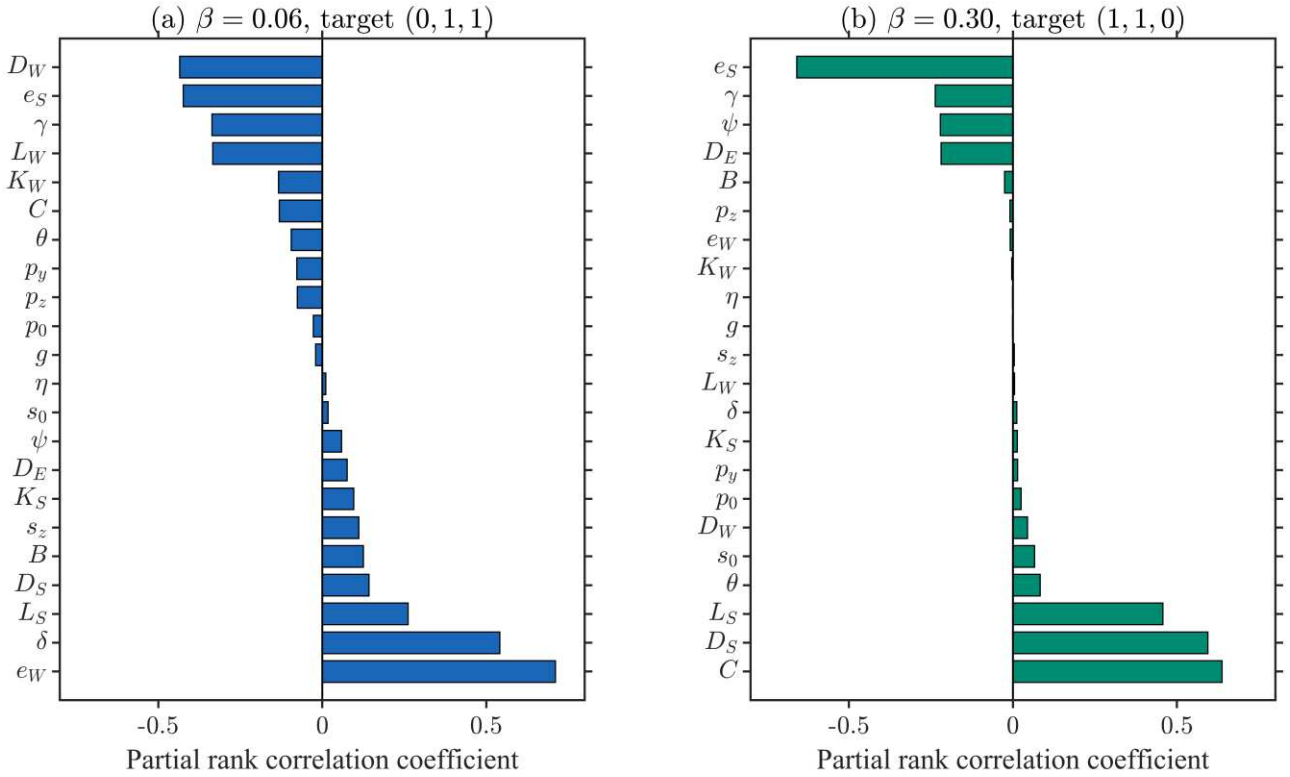

Figure 9 | Partial rank correlation coefficients between the stability margin of the target equilibrium and the 22 sampled parameters: (a) β=0.06 with target (0,1,1); (b) β=0.30 with target (1,1,0). Bars are sorted within each panel.

The rank correlations in Figure 9 identify the parameters that carry the conclusions. In regime II, the margin is driven by the crew's effort saving e_W (+0.71), the intensity of unsafe behavior δ (+0.54), the worker's loss D_W (−0.44), the supervisor's effort saving e_S (−0.42), and the leniency intensity γ (−0.34). In regime IV, the ranking changes: the regulation cost C (+0.64), the supervisor's loss D_S (+0.59), and the supervisor's reputational loss L_S (+0.46) dominate, and the supervisor's effort saving e_S enters with −0.66. The fine K_S and

the joint-liability coefficient η have coefficients below 0.10 in both panels, which is the numerical counterpart of the invariance proved in Section 7.5.

7.5. Which Levers Move the Equilibrium

The maps and rank correlations agree that the parameters usually proposed as policy instruments do not move the equilibrium. To make this point exact, each parameter was varied over a wide interval at β=0.06, and the critical value at which the target (0,1,1) loses stability was computed by bisection. Table 9 presents the results.

Table 9 | Critical values of the parameters at β=0.06, obtained by bisection on the stability of (0,1,1), and the change produced when the threshold is crossed. The ratio is the critical value divided by the baseline. In the last two rows no critical value exists inside the scanned interval; for sz and ψ the value given is the level below which the central state disappears, so the state is unchanged above it.

Parameter	Baseline	Critical value	Ratio	Effect of crossing the threshold
C	4.50	6.65	1.48	The central state is replaced by (0,1,0)
e_W	15.0	14.20	0.95	The crew abandons the shortcut
e_S	14.0	17.06	1.22	The supervisor abandons strict duty
B	10.0	5.70	0.57	The firm abandons strong regulation
p_z	0.20	0.360	1.80	Higher detection destabilizes the target
g	3.00	7.30	2.43	Schedule pressure displaces regulation
K_S	3.00	none on [0.5,10]	–	No change of the attracting state
s_z	0.35	0.044	0.13	No change above 0.044
η	0.40	none on [0.05,1]	–	No change of the attracting state
ψ	0.30	0.085	0.28	No change above 0.085

The levers that move the state are those that change the return to a departure rather than the penalty for it: e_W, e_S, C, B and g. A utility that wants the central state must reduce the effort saved by a shortcut, keep the cost of regulation inside 1.48 times the baseline, and hold accountability above 0.57 times the baseline. The levers that leave it where it is are the ones the industry relies on: over a twenty-fold range of K_S, a range of η from 0.05 to 1, and a range of ψ from 0.09 to 0.75, the attracting state does not change.

Two of the thresholds run against intuition. Raising the detection rate p_z above 0.36 also destabilizes the central state, but for a benign reason: detection revenue makes compliance profitable for the crew, so the premise of the target disappears and the system enters the transition region rather than moving to another pure strategy. The opposite threshold operates on e_W: lowering the effort saved below 14.2 makes the shortcut unprofitable and the crew safe, because a large effort saving is exactly what makes the departure worth its accident exposure.

The invariance of the fine K_S has an algebraic explanation. In F_S of (7), the fine enters only through $b_S=(\theta-1)L_S+K_S$ multiplied by $a_S(z)=s_0+s_z z$, and at the vertex (0,1,1) the condition $F_S(0,1)>0$ reads $\beta>\beta_S(0,1)$ with β_S linear in b_S. A higher fine raises β_S, which makes the condition harder to satisfy; at the calibrated values, the shift is smaller than the gap to β=0.06, and no value of K_S up to 10, twenty times the baseline, removes the state. The same argument applies to η, which enters F_S additively through $\eta\delta(1-x)D_S$ and is swamped by the accident term.

8. Discussion

8.1. What the Model Says

Which parameter decides the governance configuration? According to Sections 3 and 7, it is the accident-risk sensitivity, not the penalty. The thresholds of Proposition 2 are the ratios of the net private benefit of a departure to its expected accident cost. The fines and detection probability enter that ratio through the expected penalty; thus, raising either one lowers the threshold and pushes the configuration toward compliance. Penalties are then not a monotone instrument. Inside the transition regime, no pure strategy is stable, and the fines examined in Section 7.5 leave the attracting set where they were found.

The transition region is not featureless. Applied work usually reads the absence of a pure-strategy equilibrium as a modeling failure or a numerical nuisance. Here, the interval is 0.058 wide in β, the flow preserves volume inside it, the compliance face carries an explicit first integral, and the supervisory cycle is a nonlinear center. The amplitude is set by history. The cycle persists as long as β remains in the interval.

Regulation cost divides the space a second time. Below $C_E(0,1)=6.65$, the firm can act as a guarantor of safety. Above it, governance passes to one of the other two layers, and the system becomes bistable between a supervised and a self-regulating configuration. These two are not equivalent. In the supervised configuration, the crew keeps taking shortcuts, and the supervisor is the only barrier. In the self-regulating configuration, the crew complies because risk is salient, and no supervisory layer is active. The second state reads better in a compliance report and is the more fragile of the two, as it rests on the crew's perceived exposure rather than on an institution.

Global analysis changes what can be claimed. The compliant regime survives a ±30% band on all 22 sampled parameters. The supervised regime survives in less than half of the draws. A point calibration alone would have presented the second regime as a stable prediction. It is a conditional one, and the rank correlations name the four parameters that decide whether it holds.

8.2. Implications for Practice

Consider the crew first. The lever that matters is the effort a shortcut saves, eW, which carries the largest rank correlation in regime II. Cutting that saving means making the correct procedure the cheaper one: grounding sets assembled before the outage, permits issued at the work point instead of the office, and tooling that removes the reason to skip a step. Fines do not substitute. A fine is discounted by the detection probability, and that probability is a decision taken by the other two layers.

For the supervisor, the model returns an uncomfortable answer. Neither the fine for lenient duty nor the joint-liability coefficient moves the equilibrium, while the supervisor's reputational loss LS and the firm's detection rate carry large rank correlations. Supervisory performance should therefore be made visible rather than expensive. Publication of supervision records, rotation across crews, and a career track in which a clean record is an asset act on LS. A penalty schedule that is rarely applied, because the firm does not want to lose the supervisor, acts on nothing.

For the firm, the binding constraint is the cost of regulation relative to its accountability, and the diagonal of Figure 8(c) represents this trade-off. A utility facing a stronger sanction from the parent company or regulator can afford a more expensive safety apparatus and remain in the central state. One facing a weak sanction drifts into the region where regulation is abandoned. Schedule pressure, g, works in the same direction and is the parameter most directly under managerial control. A threshold of 2.43 times the baseline is not a wide margin in an industry where outage windows are negotiated under political pressure.

For the regulator, the model argues against the two instruments that dominate current practice. Raising the fine for supervisory failure and raising the joint-liability share both leave the attracting state where it is, over ranges wider than anything a regulator would set. What moves the state is the perceived probability that a departure produces an accident. Near-miss reporting raises that probability, immediate feedback on violations raises it, and so does publication of investigation findings. β is also the only parameter that carries the system into the compliant regime for every $C>1.20$, the range in which strong regulation is not self-financing at zero risk.

One qualification comes from the hazard-backlog extension. Closing defects faster lowers the backlog and dampens the cycle. This does not remove the cycle. Substitution between supervision and regulation generates the oscillation, not the accumulation of defects; therefore, a program that tracks only the closure rate will report progress while the oscillation continues.

8.3. Relation to the Literature

Three bodies of work bear on this one. The closest are the evolutionary analyses of safety supervision in coal mining[21–24], construction[25,26], and food safety[27]: same three-population structure, same use of pure-strategy stability, and in several cases a transition interval that is reported but never analyzed. The differences are substantive. Accident probability is endogenous and interactive, which is where the additional thresholds and the transition interval come from. Penalties are proportional to fault; therefore, the joint punishment that overstates deterrence in the earlier models never arises. The supervisory population is given an explicit shirking incentive, which makes the second link of the chain strategic rather than mechanical.

Conservative replicator systems are the second body of work, and the connection here is exact rather than analogical[19]. In that literature, conservation is the classical source of first integrals and neutral cycles, and the question that matters is which invariant sets organize the flow. Conservation is not generic in the present model. It follows from the structure of (9), in which each payoff difference is independent of the population's own strategy, and the same feature makes every face invariant. Proposition 4 excludes interior attractors, Lemma 1 excludes states in the relative interior of a face, Proposition 5 supplies the first integral on the compliance face, and Proposition 6 shows what a slow variable does to the conservation. The center becomes a relaxation cycle, which the bifurcation literature would read as the loss of a conservative center under a singular perturbation[35].

The economics of enforcement is the third. Peltzman's risk-compensation argument[32] and the Becker–Stigler tradition behind it[36,37] supply the mechanism. Proposition 3 is the local version of this argument. The equilibrium analysis in Section 3 shows that the local effect does not reverse the global comparative statics, a distinction that tends to disappear when the two are discussed together. That fines leave the equilibrium where it is fits the crowding-out evidence[38–40] and the multitask result that a penalty on one dimension reallocates effort rather than raising it[41]. What the model adds is a mechanism specific to the supervision chain: the fine is multiplied by a detection probability that the sanctioned agent does not control.

Behavioral and accident-causation research supplies the interpretation of β and the supervisory layer[9–11]. Safety-climate research supplies the empirical warrant for treating supervision as a predictor of crew behavior[12–16]. The contribution to that literature is not another empirical regularity. It is a formal account of why a supervisory arrangement that works at one level of risk perception fails at another.

8.4. Limitations

Two assumptions bound the behavioral side. The model is deterministic and treats populations as well-mixed within a task. A large maintenance unit has a small crew and

imitation is local; therefore, a network version with a Fermi rule, in the tradition of spatial and statistical physics models of imitation[42,43], is the natural extension. The supervisor's social cost is also collapsed into one parameter, eS, although in practice, it depends on crew composition, rotation policy, and seniority. A richer specification would make eS endogenous.

Three further limitations bound what the numbers can carry. Risk enters through its objective probability. A prospect-theoretic weighting[44,45] would raise the effective β for small probabilities, which would move the thresholds but not the classification, since that is monotone in β. The calibration is a protocol rather than a data set. Six of the 25 parameters rest on expert elicitation or published estimates, and the global analysis quantifies their influence without turning the thresholds into estimates. The hazard backlog is a single stock with a linear closure rate, whereas real rectification is prioritized by severity. This adds a second dimension that the model does not carry.

None of this touches the method. The classification of Proposition 2 and the invariance results of Section 7.5 follow from the payoff structure; therefore, they do not depend on the calibration. The numerical thresholds do, and they should be read as the values implied by the stated evidence rather than as estimates for a particular utility.

9. Conclusions

This study examines a chain that runs from the crew's compliance, through the supervisor's diligence, to the firm's willingness to pay for supervision. Treating it as a three-population evolutionary game converts a governance question into a question about the shape of a replicator system, and the system can be solved exactly. A vertex criterion classifies all eight pure-strategy states in closed form. At the calibrated values, the risk sensitivity separates four regimes by three thresholds. In the first, crews take shortcuts whatever the firm does. In the second, supervision is the only barrier. The third is a transition interval in which no pure strategy is stable. In the fourth, compliance survives the withdrawal of regulation.

The transition interval is not a residual case. The volume is preserved in logit coordinates; therefore, no interior state is asymptotically stable. On the compliance face, the reduction admits an explicit first integral, and its interior equilibrium is a nonlinear center whose coordinates and period follow from the parameters. A hazard backlog breaks the conservation and turns the center into a relaxation cycle. The amplitude is then set by the parameters, and it falls as the defects are closed faster, although the cycle does not disappear.

The policy conclusion runs against the instrument mix that regulators rely on. The fine for supervisory failure and the joint-liability share leave the attracting state unchanged over ranges far wider than any regulator would set; the preventive efficacy of regulation does so only above 0.085, a little over a quarter of its baseline value. What moves it is the effort a shortcut saves, the cost of regulation relative to accountability, the schedule pressure that makes regulation expensive, and above all, the perceived probability that a departure produces an accident. Only the last carries the system into the compliant regime whenever that regime is reachable ($C>1.20$), and it is also the one that a near-miss reporting system and a credible investigation process can change.

Acknowledgements

The authors thank the project management staff for general discussions of supervision and incentive practice; these consultations informed contextual understanding and were not conducted or analyzed as research interviews or empirical data. During the preparation of this study, the authors used MATLAB R2023b and Python 3.12 for the purposes of code development, numerical computation, and figure generation. The authors have

reviewed and edited the output and take full responsibility for the content of this publication. This work was supported by the National Natural Science Foundation of China (grant number 71573086) and the Philosophy and Social Science Applied Research Major Project of Higher Education Institutions of Henan Province (grant number 2018-YYZD-10).

Author contributions

Conceptualization, X.Y. and K.M.; Methodology, X.Y.; Software, K.M.; Validation, K.M.; Formal Analysis, X.Y. and K.M.; Investigation, X.Y.; Data Curation, K.M.; Writing—Original Draft Preparation, X.Y.; Writing—Review and Editing, X.Y. and K.M.; Visualization, K.M.; Supervision, X.Y.; Project Administration, X.Y.; Funding Acquisition, X.Y. All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Data availability

No new empirical data were generated or analyzed in this study. All results reported in this Article are derived from the model specified in Sections 2–6 and the parameter values listed in Tables 1 and 7, and are fully reproducible from the code we provide. The MATLAB R2023b and Python 3.12 source code that generates every figure, table and numerical value reported here, together with a README describing how to run it, is provided as Supplementary File S1. The same code regenerates the trajectory data and the figure files shown in the Article. Further inquiries can be directed to the corresponding author.

References

1. State Administration for Market Regulation of China. Safety code of electric power industry—Part of electric lines (GB 26859-2011). Standards Press of China, Beijing (2011).
2. Standing Committee of the National People's Congress. Work Safety Law of the People's Republic of China (2021 Revision). Beijing (2021).
3. de Souza, D. F., Martins, W. A., Martinho, E. & Tatizawa, H. Accidents leading to electrical shocks in Brazilian electric power distribution: An analysis. *IEEE Ind. Appl. Mag.* **30**, 61–67 (2024). https://doi.org/10.1109/mias.2023.3328507
4. Chen, S., Wang, D., Zhang, X., Shao, B., Cao, K. & Li, Z. A spatiotemporal analysis of personal casualty accidents in China's electric power industry. *Heliyon* **10**, e33855 (2024). https://doi.org/10.1016/j.heliyon.2024.e33855
5. Lu, D., Xu, C., Mi, C., Wang, Y., Xu, X. & Zhao, C. Establishment of a key hidden danger factor system for electric power personal casualty accidents based on text mining. *Information* **12**, 243 (2021). https://doi.org/10.3390/info12060243
6. Jiang, Y., McCalley, J. D. & Van Voorhis, T. Risk-based resource optimization for transmission system maintenance. *IEEE Trans. Power Syst.* **21**, 1191–1200 (2006). https://doi.org/10.1109/tpwrs.2006.879257
7. Li, W. & Korczynski, J. A reliability-based approach to transmission maintenance planning and its application in BC Hydro system. *IEEE Trans. Power Deliv.* **19**, 303–308 (2004). https://doi.org/10.1109/tpwrd.2003.820183
8. Zhang, C., Fang, Z. & Dong, W. Preventive maintenance strategy for multi-component systems in dynamic risk assessment. *Reliab. Eng. Syst. Saf.* **254**, 110611 (2025). https://doi.org/10.1016/j.ress.2024.110611
9. Reason, J. Human Error (Cambridge University Press, 1990). https://doi.org/10.1017/cbo9781139062367
10. Rasmussen, J. Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Trans. Syst. Man Cybern.* **SMC-13**, 257–266 (1983). https://doi.org/10.1109/tsmc.1983.6313160
11. Leveson, N. A new accident model for engineering safer systems. *Saf. Sci.* **42**, 237–270 (2004). https://doi.org/10.1016/s0925-7535(03)00047-x
12. Zohar, D. Safety climate in industrial organizations: Theoretical and applied implications. *J. Appl. Psychol.* **65**, 96–102 (1980). https://doi.org/10.1037/0021-9010.65.1.96
13. Neal, A. & Griffin, M. A. A study of the lagged relationships among safety climate, safety motivation, safety behavior, and accidents at the individual and group levels. *J. Appl. Psychol.* **91**, 946–953 (2006). https://doi.org/10.1037/0021-9010.91.4.946

14. Christian, M. S., Bradley, J. C., Wallace, J. C. & Burke, M. J. Workplace safety: A meta-analysis of the roles of person and situation factors. *J. Appl. Psychol.* **94**, 1103–1127 (2009). https://doi.org/10.1037/a0016172
15. Guldenmund, F. W. The nature of safety culture: A review of theory and research. *Saf. Sci.* **34**, 215–257 (2000). https://doi.org/10.1016/s0925-7535(00)00014-x
16. Cooper, M. D. Towards a model of safety culture. *Saf. Sci.* **36**, 111–136 (2000). https://doi.org/10.1016/s0925-7535(00)00035-7
17. Smith, J. M. & Price, G. R. The logic of animal conflict. *Nature* **246**, 15–18 (1973). https://doi.org/10.1038/246015a0
18. Taylor, P. D. & Jonker, L. B. Evolutionary stable strategies and game dynamics. *Math. Biosci.* **40**, 145–156 (1978). https://doi.org/10.1016/0025-5564(78)90077-9
19. Hofbauer, J. & Sigmund, K. Evolutionary Games and Population Dynamics (Cambridge University Press, 1998). https://doi.org/10.1017/cbo9781139173179
20. Friedman, D. Evolutionary games in economics. *Econometrica* **59**, 637–666 (1991). https://doi.org/10.2307/2938222
21. Liu, D., Xiao, X., Li, H. & Wang, W. Historical evolution and benefit–cost explanation of periodical fluctuation in coal mine safety supervision: An evolutionary game analysis framework. *Eur. J. Oper. Res.* **243**, 974–984 (2015). https://doi.org/10.1016/j.ejor.2014.12.046
22. Liu, Q., Li, X. & Hassall, M. Evolutionary game analysis and stability control scenarios of coal mine safety inspection system in China based on system dynamics. *Saf. Sci.* **80**, 13–22 (2015). https://doi.org/10.1016/j.ssci.2015.07.005
23. Liu, Q., Li, X. & Meng, X. Effectiveness research on the multi-player evolutionary game of coal-mine safety regulation in China based on system dynamics. *Saf. Sci.* **111**, 224–233 (2019). https://doi.org/10.1016/j.ssci.2018.07.014
24. Tu, J., Shi, Q., Li, M. & Li, J. Evolutionary game analysis on safety supervision for coal mine considering speculative behavior. *IEEE Access* **12**, 20907–20918 (2024). https://doi.org/10.1109/access.2024.3362826
25. Chen, Y., Zeng, Q., Zheng, X., Shao, B. & Jin, L. Safety supervision of tower crane operation on construction sites: An evolutionary game analysis. *Saf. Sci.* **152**, 105578 (2022). https://doi.org/10.1016/j.ssci.2021.105578
26. Su, W., Gao, X., Jiang, Y. & Li, J. Developing a construction safety standard system to enhance safety supervision efficiency in China: A theoretical simulation of the evolutionary game process. *Sustainability* **13**, 13364 (2021). https://doi.org/10.3390/su132313364
27. Sun, J., Fan, R. & Yang, Z. An evolutionary game analysis of periodical fluctuation in food safety supervision. *Mathematics* **10**, 1326 (2022). https://doi.org/10.3390/math10081326
28. Xin, X. Research on safety regulation of chemical enterprise under third-party mechanism: An evolutionary approach. *Symmetry* **11**, 1057 (2019). https://doi.org/10.3390/sym11081057
29. You, M., Li, S., Li, D., Cao, Q. & Xu, F. Evolutionary game analysis of coal-mine enterprise internal safety inspection system in China based on system dynamics. *Resour. Policy* **67**, 101673 (2020). https://doi.org/10.1016/j.resourpol.2020.101673
30. Feng, F., Liu, C. & Zhang, J. China's railway transportation safety regulation system based on evolutionary game theory and system dynamics. *Risk Anal.* **40**, 1944–1966 (2020). https://doi.org/10.1111/risa.13528
31. Tirole, J. Hierarchies and bureaucracies: On the role of collusion in organizations. *J. Law Econ. Organ.* **2**, 181–198 (1986). https://doi.org/10.1093/oxfordjournals.jleo.a036907
32. Peltzman, S. The effects of automobile safety regulation. *J. Political Econ.* **83**, 677–725 (1975). https://doi.org/10.1086/260352
33. McKay, M. D., Beckman, R. J. & Conover, W. J. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **21**, 239–245 (1979). https://doi.org/10.1080/00401706.1979.10489755
34. Marino, S., Hogue, I. B., Ray, C. J. & Kirschner, D. E. A methodology for performing global uncertainty and sensitivity analysis in systems biology. *J. Theor. Biol.* **254**, 178–196 (2008). https://doi.org/10.1016/j.jtbi.2008.04.011
35. Kuznetsov, Y. A. Elements of Applied Bifurcation Theory (Springer, 2004). https://doi.org/10.1007/978-1-4757-3978-7
36. Becker, G. S. Crime and punishment: An economic approach. *J. Political Econ.* **76**, 169–217 (1968). https://doi.org/10.1086/259394
37. Stigler, G. J. The optimum enforcement of laws. *J. Political Econ.* **78**, 526–536 (1970). https://doi.org/10.1086/259646
38. Gneezy, U. & Rustichini, A. A fine is a price. *J. Legal Stud.* **29**, 1–17 (2000). https://doi.org/10.1086/468061
39. Frey, B. S. & Jegen, R. Motivation crowding theory. *J. Econ. Surv.* **15**, 589–611 (2001). https://doi.org/10.1111/1467-6419.00150
40. Bénabou, R. & Tirole, J. Intrinsic and extrinsic motivation. *Rev. Econ. Stud.* **70**, 489–520 (2003). https://doi.org/10.1111/1467-937x.00253
41. Holmström, B. & Milgrom, P. Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design. *J. Law Econ. Organ.* **7**, 24–52 (1991). https://doi.org/10.1093/jleo/7.special_issue.24
42. Nowak, M. A. & May, R. M. Evolutionary games and spatial chaos. *Nature* **359**, 826–829 (1992). https://doi.org/10.1038/359826a0

907 43. Perc, M., Jordan, J. J., Rand, D. G., Wang, Z., Boccaletti, S. & Szolnoki, A. Statistical physics of human cooperation. *Phys. Rep.*
908 **687**, 1–51 (2017). https://doi.org/10.1016/j.physrep.2017.05.004
909 44. Kahneman, D. & Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica* **47**, 263–291 (1979).
910 https://doi.org/10.2307/1914185
911 45. Tversky, A. & Kahneman, D. Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertain.* **5**, 297–
912 323 (1992). https://doi.org/10.1007/bf00122574

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FileS1code.zip
- FileS2figuresdata.zip