

Waiting, judgment, and responsibility: lessons from divination for human oversight of irreversible decisions

Junzhe Shi

junzhe@berkeley.edu

University of California, Berkeley

Yuan Gan

Quotr AI

shengyu Tao

Chalmers University of Technology

Shumo Wang

Duke University

Yuhan Hao

University of California, San Francisco

Article

Keywords: human oversight, AI agents, divination, Yijing, responsibility gap, sequential decision-making

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11040474/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Waiting, judgment, and responsibility: lessons from divination for human oversight of irreversible decisions

Author Names & Affiliations:

Junzhe Shi^{1*}, Yuan Gan², Shengyu Tao³, Shumo Wang⁴, Yuhan Hao⁵

1. University of California, Berkeley, Systems Engineering, Berkeley, CA 94720, USA
2. Quotr AI, Castro Valley, CA, USA
3. Chalmers University of Technology, Department of Electrical Engineering, SE-412 96 Gothenburg, Sweden
4. Duke University, Department of Religious Studies, Durham, NC 27708, USA
5. University of California, San Francisco, Department of Neurological Surgery, San Francisco, CA 94143, USA

*** Corresponding Author:**

Junzhe Shi

Email: junzhe@berkeley.edu

Abstract

Human oversight of an AI agent’s irreversible actions depends on whether the user can obtain evidence and bear responsibility for the outcome. Divination offers an earlier example of support for consequential choices. Using consultation of the Yijing, the Chinese Book of Changes, we examine how a pause, an answer to interpret and reassurance about blame can affect decisions. Our model separates the value of further evidence from the effects of easing anxiety about waiting and fear of blame. In the simulations, requiring two more observations reduces an overconfident person’s actions under unfavorable conditions from 4.5% to 2.1% of decisions. Reassurance helps anxious people keep checking and fearful people act on favorable evidence: among fearful people told “go”, missed opportunities fall from 13.8% to 2.7%. Treating reassurance as evidence can instead lower payoff below the level without advice. We derive conditions under which offering both forms of relief to everyone does at least as well as random verdicts in expected material payoff. A simulation of an AI agent handing over a decision shows when requiring new evidence before action beats merely allowing checks, and when it imposes costs. We propose oversight that gives users relevant evidence and a plan they can examine, protects careful decisions, and compensates those harmed through a separate route.

Keywords: human oversight; AI agents; divination; Yijing; responsibility gap; sequential decision-making

1 Introduction

In September 1983, a Soviet early-warning system reported five incoming US missiles (Forden et al. 2000). Stanislav Petrov, the officer assessing the alert, expected a first strike to involve hundreds of missiles and noted the absence of ground-radar confirmation. After several minutes, he judged the warning false (Hoffman 1999). His task illustrates a familiar difficulty: deciding whether to trust a warning while under pressure to respond.

An AI agent, software that acts on a user’s behalf, poses a similar difficulty when it proposes to send money, delete files or release new code. An approval prompt lets the user say yes or no, but checking the proposal and accepting responsibility for a failure take more than a button. People can accept automated advice too readily, even when their own information points elsewhere (Skitka et al. 1999; Parasuraman and Manzey 2010).

Research on oversight identifies the practical support people need. Automation can erode the skills needed to judge its output (Bainbridge 1983), and oversight duties can demand more time, information and authority than overseers have (Green 2022; Sterz et al. 2024; Laux 2024). Responsibility raises a related problem. When a machine acts, it can be hard to say who should answer for harm (Matthias 2004; Sparrow 2007), and the person nearest the action may take the

blame despite having little control over the system (Elish 2019). Accepting responsibility therefore requires enough understanding and opportunity to judge the action (Tigard 2021; Kiener 2022).

People faced these problems long before AI. Decisions about war, moving home and marriage rested on uncertain information and had lasting consequences. Consulting an oracle was one common response (Park 1963; Boyer 2020). In a consultation of the Yijing, the Chinese Book of Changes, preparation delayed the answer, the answer came as a text to interpret, and believers took it to carry the authority of Heaven. We ask how these features, which a bare yes-or-no prompt lacks, could support judgment and ease the pressure to decide.

Each feature has been studied separately. The value of waiting for evidence is well established (Wald 1947; Arrow et al. 1949). Delegation can shift blame (Bartling and Fischbacher 2012), and chance can distance a divinatory verdict from the adviser’s interests (Moore 1957; Elster 1989; Boyer 2020). We combine these mechanisms in one model of decisions made as evidence arrives, linking how advice is interpreted to the costs of waiting and failure.

We show that reassurance can improve material outcomes without changing beliefs, whereas treating it as evidence can reverse the gains. We derive conditions under which offering two forms of relief to everyone, from the discomfort of waiting and from fear of blame, does at least as well in expected material payoff as allocating them through random verdicts. We then apply the model to an AI agent handing a decision to a user: whether requiring new evidence before action beats optional checking depends on how the agent weighs evidence and how often users follow it without checking. Together, these results provide a basis for oversight that supports both judgment and responsibility.

2 The oracle we use as our example

The Yijing, also called the Zhouyi, has a long record of use and interpretation (Smith et al. 1990). Its consultations show preparation, judgment and responsibility within one practice. Consider a ruler deciding whether to attack a neighbor. The scouts disagree, and a campaign would be hard to reverse. The account below draws on sources from different periods to illustrate a consultation, not to reconstruct one event.

Consultation begins with doubt. One chronicle records an officer declining a proposed divination before an attack because he had no doubt (Zuozhuan, Huan 11; Yang 1990). The Book of Documents recommends consultation “when you have a great doubt” and sets divinatory answers alongside one’s own judgment, ministers’ advice and the people’s views (Shangshu, Hongfan; Ruan 1980). The oracle is therefore one source among several.

Preparation takes time. Zhu Xi’s twelfth-century manual prescribes washing, incense and spoken words before casting begins (Adler 2020). Section 3 asks what this delay could achieve if useful

news arrived during it. The cast then produces a hexagram, a figure of six lines with accompanying texts. In the later coin method, three coins determine each line, so six throws produce the figure; the older yarrow-stalk method requires repeated dividing and counting of stalks (Wilhelm and Baynes 1967). An unmanipulated cast makes it hard to choose a preferred figure in advance.¹

For believers, this chance procedure also carried religious authority. A commentary calls the stalks spirit things produced by Heaven and says they answer “like an echo” (Zhouyi, Xici A; Ruan 1980). Yet the ruler must still interpret the answer. “It is favorable to cross the great river” might mean attacking across a border or opening negotiations. Recorded consultations show people accepting or rejecting a specialist’s reading (Smith 1989), and a warning against repeated casting discourages seeking a more agreeable answer (Zhouyi, Meng; Ruan 1980).

The text may also promise “no blame”. A commentary connects this phrase with correcting one’s faults, and one line pairs it with misfortune (Zhouyi, Xici A and Da Guo; Ruan 1980). In these passages, conduct can be judged separately from the outcome: a bad result need not mean blameworthy conduct. In the Zuo zhuan, Mu Jiang, a ruler’s widow confined after a political plot, rejects a scribe’s reassuring prediction of her release. She reads “no blame” as conditional on virtues she says she lacks, judging the reading against her own conduct (Zuo zhuan, Xiang 9; Durrant et al. 2016; Smith 1989). The speech, possibly a later composition, presents a standard for judging advice against one’s conduct (Smith 1989). This separation of conduct from outcome motivates our question of whether reassurance can ease fear of blame.

Consultation thus combines a pause, an answer requiring judgment and a way to consider responsibility. We first examine the pause: what changes when a person receives more evidence before deciding?

3 What waiting does

News arriving during preparation can change a decision. Our model asks when requiring people to keep checking improves a choice that cannot be undone. Before such a choice, a person has some evidence and can get more by waiting; a family considering emigration might hear from a relative who has already moved. Waiting costs something, but more information can improve the choice. Sequential decision theory studies this trade-off for a calibrated person, one who gives evidence its proper weight (Wald 1947; Arrow et al. 1949).

In the model, a person keeps collecting evidence while their belief lies between two thresholds, acting above the upper one and giving up below the lower one. Both thresholds depend on the time remaining. We call a rule that says what to do at each belief and time a policy. Required waiting cannot improve a calibrated policy that already maximizes expected material payoff. It can help someone who stops too soon, if the extra evidence is worth its cost.

One reason for stopping too soon is overconfidence: people often treat their evidence as more conclusive than it is (Moore and Healy 2008). In the selected case in Fig. 1, the state is bad, but two favorable observations lead an overconfident person to act at step 2. Requiring two more observations after the first would delay the decision to step 3, when the next observation breaks the favorable run. We call such a run, which looks conclusive only because evidence is overweighted, a phantom pattern.

3.1 The model

Table 1 introduces four people facing the same decision. They differ in how much weight they give evidence, how much waiting troubles them, or how much they fear blame.

Table 1. The four decision types.

Person	Evidence weighting and decision costs	Parameter values	Meaning
Calibrated	Weights evidence correctly and considers only material costs.	$\gamma = 1$, $C_{\text{anx}} = 0$, $R_{\text{blame}} = 0$	Provides the benchmark for comparison.
Overconfident	Weights each observation four times as strongly as it should be weighted.	$\gamma = 4$ (also 2 and 6); all other parameters match the calibrated person.	Becomes confident too soon.
Anxious	Experiences discomfort while waiting, making its total perceived cost four times the material cost.	$C_{\text{anx}} = 1.5$, in units where success pays 20; all other parameters match the calibrated person.	Feels pressure to end the wait.
Fearful	Fears blame after acting, making the total perceived cost of failure three times the material loss.	$R_{\text{blame}} = 200$, in the same units; all other parameters match the calibrated person.	Holds back because failure may bring blame.

The world is either good or bad, with equal prior probabilities, and stays so throughout the decision. At each step t , an observation arrives: $x_t \sim N(\mu, \sigma^2)$, a normal distribution with mean μ and spread σ . We set $\mu = +1$ in the good state, $\mu = -1$ in the bad state and $\sigma = 1.25$, so each observation points the wrong way about one time in five. Observations are independent given the state. The running evidence score L_t is measured in perceived log-odds: positive values favor the good state, negative values the bad state, and $L_0 = 0$ means even odds. Each observation adds to the score:

$$L_{t+1} = L_t + \gamma \cdot \frac{2x_{t+1}}{\sigma^2}, \quad (1)$$

The multiplier γ gives evidence its correct weight at $\gamma = 1$ and too much weight at $\gamma > 1$. After the first observation, the person can act, give up or wait; at the deadline $H = 20$, they must act or give up. Acting succeeds with probability $Q = 0.95$ in the good state and $1 - Q = 0.05$ in the bad state. Success pays $R_{\text{OK}} = 20$, failure pays $R_{\text{BAD}} = -100$ and giving up pays 0, before observation costs.

Each observation costs $C = 0.5$, including the first. Costs already paid cannot be recovered, so each choice compares acting, giving up and paying for another observation:

$$V_t(L) = \max\{A(L), 0, -(C + C_{\text{anx}}) + E[V_{t+1}(L_{t+1}) | L_t = L]\},$$

$$A(L) = s(L) R_{\text{OK}} + (1 - s(L)) (R_{\text{BAD}} - R_{\text{blame}}), \quad (2)$$

Here $V_t(L)$ is the best expected value still available. $A(L)$ is the value of acting now; giving up is worth zero. Waiting costs C plus the anxiety cost C_{anx} and leads to a new choice after the next observation, and E averages over that observation using the person's beliefs. At the deadline, $V_H(L) = \max\{A(L), 0\}$. The perceived probability of a good state is $b(L) = 1/(1 + e^{-L})$, so the perceived chance that acting succeeds is $s(L) = Q b(L) + (1 - Q)(1 - b(L))$. R_{blame} is the extra cost of being blamed for failure. Table 2 lists the parameters, which are illustrative: behavioral studies motivate the pressures, but the values are not estimated from them.

We solve equation (2) by dynamic programming: starting at the deadline, where the person compares acting with giving up, we work backward and add the option of another observation at each earlier step. This gives the best policy for each person's beliefs and costs; a simpler one-step-ahead rule is examined in Supplementary Table S2. The overconfident person treats observations as having spread $\sigma/\sqrt{\gamma}$, that is, as more precise than they are. This affects both the update in equation (1) and the forecast of future evidence; actual observations keep $\sigma = 1.25$. The Supplementary Information gives the numerical solution and precision checks (Supplementary Table S12).

Section 4 adds a random verdict before the first permitted decision: “go” with probability p and “wait” otherwise. We use $p = 1/4$ and vary it from 0 to 1 in Supplementary Fig. S2. The draw is independent of the state and evidence. In the main runs it leaves beliefs unchanged and reduces one extra cost by a share θ between 0 and 1:

$$C_{\text{anx}} \rightarrow (1 - \theta) C_{\text{anx}} \text{ after a “wait” verdict;}$$

$$R_{\text{blame}} \rightarrow (1 - \theta) R_{\text{blame}} \text{ after a “go” verdict.} \quad (3)$$

At $\theta = 0$, with no belief shift, the verdict leaves the policy unchanged and serves as a control for relief. Section 4 also examines treating the verdict as evidence: “go” adds δ to perceived log-odds and “wait” subtracts it.

For each setting, we compute each person's policy and simulate 100,000 decisions. We record actions in the bad state, giving up in the good state (missed opportunities) and average payoff. An action in the bad state can succeed and one in the good state can fail; Fig. 2a counts the latter failures separately. Unless stated otherwise, error rates are shares of all simulated decisions; rates for one verdict are shares of the decisions receiving it. Because the two states are about equally common, missing opportunities in 10% of all decisions means missing about 20% of good-state

opportunities. Payoff counts material outcomes only: success gains, failure losses and observation costs. Anxiety and blame affect choices but are excluded from reported payoff, as are the administrative costs of protection or insurance.

Shaded bands and error bars are 95% Monte Carlo intervals, showing how much results vary between simulation runs with the same parameters. Comparisons use the same simulated decisions and evidence streams, with paired intervals where reported. Figures and tables use different fixed seeds, so a repeated quantity can differ slightly. The Supplementary Information examines sensitivity to the parameters.

Table 2. Main simulation parameters. Supplementary Table S3 gives their rationale and sensitivity ranges. Section 6 also varies the share of users who accept an agent's decision without further checking.

Parameter	Value	Meaning
σ , observation spread	1.25	An observation points the wrong way about one time in five. Several moderately favorable observations usually support action.
R_{OK} , R_{BAD} , success and failure payoffs	+20, -100	A failure costs five times what a success earns.
C , observation cost	0.5	Each observation step costs one fortieth of the gain from success and one two-hundredth of the loss from failure.
Q , success probability in the good state	0.95	An action in the good state fails one time in twenty.
H , deadline	20	The person must act or give up by the final step.
γ , overconfident person's evidence weighting	4 (2 and 6 shown)	Each observation counts four times as strongly as it should.
C_{anx} , anxious person's extra waiting cost	1.5	A step of waiting feels four times as costly as its material cost.
R_{blame} , fearful person's extra failure cost	200	A failure after acting feels three times as costly as its material loss.
θ , share of pressure relieved	0.7	A credible verdict relieves seven tenths of the relevant pressure.
p , probability of "go"	1/4	One verdict in four says "go".
k , imposed waiting steps	2 steps	After the first observation, two additional observations arrive before the person can decide.
γ_a , agent's evidence weighting	4	The agent weights each observation as strongly as the overconfident person does.

Parameter	Value	Meaning
δ , belief change caused by the verdict	0 (main); 1 (shift)	In the belief-shift comparison, the verdict adds or subtracts one unit from perceived log-odds. This is less than the mean correctly weighted increase of 1.28 in the good state.

3.2 Whom waiting helps

In Fig. 2, each person must collect k more observations after the first before deciding; with $k = 2$, the first decision can come at step 3. For the overconfident person, bad-state actions fall from 4.5% of decisions without required waiting to 2.1% with two extra steps and 1.5% with three. For the very overconfident person ($\gamma = 6$), three steps reduce them from 6.1% to 1.9%. Waiting also prevents giving up too early: the overconfident person's missed opportunities fall from 7.3% to 2.1% with three steps.

Waiting cannot remove all risk: an action in the good state still fails 5% of the time. Such failures make up about 2% of all decisions under the overconfident policy (Fig. 2a, dashed line), and their frequency varies with how often people act in the good state. This remaining risk matters for fear of blame. The benefit of waiting comes from new evidence: when waiting brings none, two extra steps leave error rates unchanged and only add costs (Supplementary Table S8). Section 7 therefore proposes requiring evidence, not just time, before action.

The calibrated person loses payoff from the first extra step: two steps reduce it from 4.75 to 4.57, about 4%, with little improvement in errors. For the overconfident person, three steps raise payoff from 1.04 to 3.28, recovering about three fifths of the gap to the calibrated person. For the very overconfident person, payoff rises from -0.57 to 2.91 with three steps. Waiting thus helps those who overweight evidence and costs those who weigh it correctly.

The officer in Section 2 declined a divination because he had no doubt. In the model, overconfidence can produce certainty while a calibrated person would still wait, so felt certainty is a poor reason to skip a check. Correcting the overconfident person's weighting, in both the current evidence score and the forecast of future observations, would restore the calibrated policy; adjusting decision thresholds is another option, whose effect depends on the adjustment chosen (Supplementary Table S1). But whether a check was completed is easier to observe than how a person weighs evidence. Consultation, moreover, combined the pause with a verdict that could ease pressure.

4 How a verdict can ease pressure

Collecting more evidence also prolongs an unresolved choice. Stress can rise with uncertainty (de Berker et al. 2016), and people sometimes choose a worse outcome sooner rather than wait for a

milder one (Berns et al. 2006). Fear of blame creates a different pressure: one parent considering emigration might want to sign for the first available flat, while the other might hold back for fear of being blamed if it goes wrong.

People under uncertainty can become eager for an answer (Kruglanski and Webster 1996), so an answer that settles the question for now could bring relief even when it says “wait”. The model’s anxious person experiences waiting as four times its material cost and ends it early by acting or giving up: bad-state actions rise from the calibrated 0.2% to 0.7%, and missed opportunities from 1.8% to 10.7%. Giving up is the larger effect because failure costs five times the gain from success.

People judge harm caused by action more harshly than harm allowed by inaction and expect to regret it more (Ritov and Baron 1990; Spranca et al. 1991), although missed opportunities dominate regret over longer periods (Gilovich and Medvec 1995). The fearful person experiences failure as three times its material cost. Because even strong evidence leaves a 5% chance that acting fails, they miss opportunities in 13.8% of decisions. Anxiety and fear can thus both look like caution (Supplementary Fig. S4).

What could ease these pressures? Divination can distance advice from the diviner’s interests and help coordinate difficult decisions (Park 1963; Boyer 2020). Malinowski (1948) described greater reliance on magic in open-sea fishing than in calm lagoons, and Sosis and Handwerker (2011) found an association between reciting psalms and lower anxiety among women who remained under rocket fire. Experiments show that delegation can reduce punishment for unfair choices and the reported sense of responsibility (Bartling and Fischbacher 2012; Hamman et al. 2010). We model the discomfort of waiting and fear of blame as separate costs and calculate the effects of reducing them; how much relief a Yijing verdict provides remains an empirical question. At $\theta = 0.7$ in equation (3), “wait” reduces the anxious person’s total waiting cost from 2 to 0.95, and “go” reduces the fearful person’s total failure cost from 300 to 160. Material costs are unchanged, and each person then chooses freely using the reduced costs and the available evidence.

Fig. 3 shows relief alone: the verdict arrives after the first observation, without required waiting or a change in beliefs. Among anxious people told “wait”, missed opportunities fall from 10.7% without a verdict to 4.0%. Among fearful people told “go”, they fall from 13.8% to 2.7%, near the calibrated 1.8%. Each verdict changes only the policy of the type whose pressure it eases, and neither changes the calibrated policy. Because three in four people are told “wait” and one in four “go”, missed opportunities average 5.7% for the anxious person and 11.0% for the fearful person. The reductions grow with θ over the range examined (Supplementary Fig. S3).

The randomness itself does not improve the average. If each verdict gives the same relief whether drawn at random or given to everyone, a random draw averages the results of telling everyone “go”

and everyone “wait”, so it cannot beat the better of the two; here, “go” gives a slightly higher average payoff (Supplementary Fig. S2).

Treating the verdict as evidence can reverse these gains. If the fearful person lets the verdict shift perceived log-odds by $\delta = 1$ toward the suggested state, average payoff falls to 3.06, below the 3.71 without a verdict: relief still helps those told “go”, but losses among those told “wait” outweigh it. For the anxious person, the benefit survives $\delta = 1$ and disappears at $\delta = 2$ (Supplementary Table S4).

We now combine the two parts of consultation. In Table 3, “the procedure” requires two extra observations and then gives a verdict relieving 70% of the relevant pressure. In an equal mixture of the four types, bad-state actions fall from 1.4% to 0.7% of decisions and missed opportunities from 8.4% to 2.8%. Material payoff rises from 3.26 to 3.97, a paired gain of 0.71 ± 0.05 (95% Monte Carlo interval). The overconfident person gains 1.89, the anxious 0.60 and the fearful 0.53; the calibrated person loses 0.18. The group gains when fewer than about 85% are calibrated, with the other types in equal shares. Relief alone leaves each type’s material payoff unchanged or higher in this setting.

An institution could offer the same relief without a draw, by protecting careful decisions whether a person acts or holds back. In the model, such protection applies both reductions in equation (3) to everyone. With the same waiting requirement, no belief shift and relief equal to that of the relevant verdict, protection gives each type in Table 1 at least the expected material payoff of random verdicts (numerical gains in Supplementary Table S8). To see why, fix the evidence and available choices at the first permitted decision. The anxious and fearful types have correct beliefs and maximize $M - \lambda D$, where M is the expected remaining material payoff, λ the extra cost, and D the expected number of further observations (for anxiety) or the probability of eventually acting and failing (for fear). Subscripts 0 and 1 denote the best policies before and after relief, with $0 \leq \lambda_1 < \lambda_0$. Each policy is best at its own cost, so

$$\lambda_1(D_1 - D_0) \leq M_1 - M_0 \leq \lambda_0(D_1 - D_0). \quad (4)$$

These bounds imply $D_1 \geq D_0$ and, since $\lambda_1 \geq 0$, $M_1 \geq M_0$: relief cannot lower material payoff at any shared evidence history, and so not after averaging over histories and adding identical initial costs. Random verdicts give each anxious or fearful person either the original or the relieved policy; universal protection gives both types the relieved policy. The calibrated and overconfident types have no extra costs and are unaffected. The argument assumes correct beliefs for the types whose costs change. For someone both overconfident and fearful, easing fear with “go” raises bad-state actions from 3.2% to 3.7% and lowers payoff from 1.86 to 1.52 (Supplementary Table S6): when evidence is already overweighted, fear acts as a brake, which supports combining protection with checks.

Across the tested costs and stakes, with policies recomputed for each setting, the full procedure raises average group payoff. When a step of waiting costs 1% of the failure loss, it is nearly neutral for the anxious person (-0.03 ± 0.08), whereas waiting alone costs them 0.18 ± 0.07 . In the highest-stakes setting (success pays 50, failure costs 1,000, good-state failure probability 0.01, observation cost 1), the overconfident person gains about 14 and the group 3.5; these gains reflect the combined parameter changes (Supplementary Fig. S1).

5 How chance and Heaven gave the verdict authority

Reassurance depends on credibility. A relative, a coin toss and an oracle give different grounds for trust, even when none adds evidence about the decision. Suppose the family asks an older relative whether to emigrate. Delegation can shift some blame to the adviser (Bartling and Fischbacher 2012), giving that person a reason to hedge, delay or tell the family what it wants to hear. Comparative accounts of divination describe such concerns (Park 1963; Boyer 2020). The family's judgment then depends partly on the adviser's reasons and incentives.

A fair coin cannot be lobbied, but it offers no judgment about either option (Elster 1989; Stone 2010). In Bartling and Fischbacher's (2012) experiment, participants who left an allocation to a die were punished less after an unfair outcome than those who chose it directly: 2.98 versus 4.64 points. Delegating to a person reduced punishment more, from 4.27 to 1.31 points in that treatment. Choosing chance did not remove blame.

Religious belief gave the cast a meaning that a coin lacks. Some historical readers understood divinatory signs as expressions of an ordered world (Raphals 2013), and a Yijing commentary describes Heaven displaying images that reveal good and ill fortune (Zhouyi, Xici A; Ruan 1980). A cast could thus carry authority beyond the people present. Yet interpretation remained open to influence: Mu Jiang's scribe proposed a favorable reading, but the text gave her grounds to reject it. The oracle combined a figure hard to select in advance with an answer the reader still had to judge, and belief in divine intention could make the result acceptable to people who would resist an arbitrary choice (Elster 1989).

Advice can also echo a person's own view. To test this risk, we replace the random verdict with a "mirror": after two observations, it says "go" if the person judges the good state more likely, and "wait" otherwise, adding no information. Compared with a random verdict given at the same step and saying "go" equally often, payoff differs by at most 0.06 across the four types when the mirror only eases the relevant pressure. With the one-unit belief shift used earlier, however, bad-state actions rise to 1.1–1.7 times the rates under a random verdict read the same way; for the calibrated person, from 0.24% to 0.40% (Supplementary Table S5).

The mirror shows what happens when agreement is counted as independent confirmation. An agent that supplies new facts or checks can improve a decision; one that merely repeats the user’s belief adds nothing that should be counted. A visible record of sources and checks helps users tell the two apart. Two practical questions for AI oversight follow: what independent support does a recommendation offer, and who answers for the action and bears its losses?

6 The approval button

AI agents raise these questions when they move money, delete files or send messages. Products provide oversight through permission prompts for single actions, review of plans before work begins, and automatic checks. Section 6.1 examines how the chance to obtain further evidence affects a decision made with an agent; Sections 6.2 and 6.3 consider how users review and what responsibility approval carries.

Imagine the emigrating family using an agent to arrange their move. In this invented example, the agent reports that it has found a flat and that the landlord wants the deposit today. It asks, “Approve the transfer?” The prompt shows the amount and the receiving account but says nothing about how the flat was checked. The family must make an irreversible choice with only the information on that screen.

The first question concerns the recommendation itself. The agent may judge the flat from market information, but its answer may also reflect what the family seems to want to hear. In a study of eleven AI systems, the systems endorsed users’ actions in an average of 51% of interpersonal-conflict cases where an online community had judged the advice seeker to be in the wrong (Cheng et al. 2026). In April 2025, OpenAI withdrew an update within days after reporting that it was excessively agreeable and encouraged impulsive actions; the update had been trained partly on user feedback (OpenAI 2025a). These findings from conversations give reason to test whether agents also tailor their plans to users’ preferences. Section 5’s mirror shows the risk when such agreement is counted as fresh evidence.

6.1 Continuing to check after the agent stops

The next question is when the person has to decide. A permission rule may stop an agent because it proposes a listed kind of action, before anyone has checked whether enough evidence supports it. In two comparisons in Table 3, the person must choose at a fixed point and cannot collect further evidence, which shows what is lost without the chance to check further. Interfaces that allow another check, a revised task or a delay keep that chance.

In the first comparison, the agent weighs evidence like the overconfident person, with $\gamma_a = 4$ (the subscript a marks the agent). It collects observations until its own rule, which uses material costs only, would act or give up; the person must then act or give up on the same observations. Early on,

the agent’s threshold for acting corresponds to a calibrated probability of about 78% that the state is good. At the stated rewards and risks, acting pays off on average only above about 87%, so the agent can stop while a calibrated person would still collect evidence. A strong final observation can also overshoot the agent’s threshold and carry calibrated belief above 87%; the calibrated person acts in those cases and gives up in the others.

When evidence collection must end where the agent stops, the calibrated person misses opportunities in 19% of all decisions, about two in five good-state opportunities, and their bad-state actions rise from about 0.2% to 2.0%. The fearful person misses opportunities in 47% of all decisions, about 94% of good-state opportunities, and their average payoff falls below zero. Across the four types, average payoff falls from 3.26 to 0.95, about 71% lower. An overconfident agent can thus hand over a decision before a careful person is ready to make it.

These outcomes depend on how the agent weighs evidence. In the reported settings, a calibrated agent gives outcomes close to the calibrated benchmark for all four types, while the $\gamma_a = 4$ agent harms three of them (Supplementary Table S9). The single-observation column is harsher still: average payoff is -0.20 and missed opportunities reach 32% of all decisions. Relating either comparison to a real product requires observing real interactions, including how often users seek further evidence after a prompt.

The restriction in Table 3 can be relaxed. We compare four arrangements at the agent’s hand-off, when its own rule first chooses to act or give up. Each user can inspect all the observations the agent collected, reweigh them in their own way and decide using their own costs. The agent’s recommendation adds no evidence; it sets when the user takes over and which observations are available.

Under A, the user may act, give up or keep checking. B adds both forms of relief, cutting the extra waiting and blame costs by 70% without changing beliefs. C requires two new observations after hand-off before action; earlier observations do not count, and the user may give up at any point, including at hand-off. D combines this requirement with relief. We recompute the user’s policy for each arrangement, and every observation carries its cost. Because C and D leave users free to refuse before completing the checks, they differ from the waiting interval in Fig. 2, which delays giving up as well as acting; its reduction in missed opportunities need not carry over to this comparison. Table 4 compares payoffs for equal shares of the four types. The Supplementary Note “Agent checking and follower scenarios” specifies the timing and the treatment of the deadline.

With an overconfident agent ($\gamma_a = 4$), allowing further checks raises group payoff from 0.95 to 3.43, requiring checks raises it to 4.09, and adding relief to required checks gives 4.32. With a calibrated agent, a requirement has a cost because the agent already weighs evidence correctly: payoff is 4.68 with optional checks and 4.31 with required checks. A requirement must therefore

be judged against the evidence the agent already obtains and the cost of obtaining more. Supplementary Table S10 gives outcomes by user type.

Table 4. Mean material payoff under alternative arrangements for checking an agent’s decision. The population contains equal shares of the four types in Table 1 and no additional followers ($f = 0$). A allows further checking; B adds both forms of relief; C requires two new observations after hand-off before action; D adds relief. Giving up remains available throughout. “Decide at hand-off” requires an immediate user decision. Success pays 20, failure pays -100 and each observation costs 0.5. Results use 100,000 decisions per type and shared random streams, seed 31. The last column gives the paired mean difference $C - A$ and its 95% Monte Carlo interval, expressed as mean $\pm$ half-width. Differences are calculated before rounding, so they can differ by 0.01 from subtraction of the displayed means. Supplementary Table S10 reports outcomes by type; Supplementary Fig. S5 shows results for $\gamma_a = 4$.

Agent γ_a	Decide at hand-off	A Optional checks	B A plus relief	C Required checks	D C plus relief	Paired C – A (95% interval)
1	4.70	4.68	4.70	4.31	4.32	-0.37 ± 0.02
1.5	3.65	4.50	4.59	4.40	4.43	-0.10 ± 0.02
2	2.83	4.26	4.39	4.34	4.41	0.08 ± 0.02
4	0.95	3.43	3.72	4.09	4.32	0.65 ± 0.03
6	0.34	3.37	3.72	3.98	4.29	0.61 ± 0.03

Some users may accept the agent’s decision without further checking even when the interface allows it. We call them followers, a share f of users whose decisions always match the agent’s policy; the remaining $1 - f$ are split equally among the four types in Table 1. Followers add no checks under A or B. Under C and D, they follow the agent’s policy for continuing, subject to the same requirement of two new observations after hand-off, with early refusal allowed. By definition, relief does not change followers’ behavior. We vary f from 0 to 1 and γ_a from 1 to 6 (Supplementary Table S11).

The more users follow an overconfident agent, the more required checks matter. At $\gamma_a = 4$, raising the follower share from 0 to 0.86 increases the gain from required over optional checks from 0.65 ± 0.03 to 2.30 ± 0.10 (paired 95% Monte Carlo intervals); at the higher share, payoffs are 1.28 with optional and 3.59 with required checks. Payoff still falls as followers increase, but required checks prevent more of the loss. With a calibrated agent ($\gamma_a = 1$), required checks lower payoff by about 0.36 across the reported mixtures. At $\gamma_a = 1.5$, the difference moves from -0.10 ± 0.02 with no followers to 0.07 ± 0.05 with followers alone, and some intermediate mixtures show no clear gain or loss at this precision. Because policies are fixed, the group result is a weighted average of follower and other-user results and so changes linearly with f ; the simulation supplies each

subgroup’s result. The practical question is whether useful new evidence is obtained before approval permits an irreversible action.

6.2 What users actually review

Review also depends on what the person can see. A Yijing reading required someone to connect a text with a case and explain why it fit. An isolated approval prompt may instead show a command, transfer or file change with little about its purpose, the checks already done or the remaining uncertainty. Research on automation and AI advice gives reason to expect difficulty when people must review a finished recommendation with so little support (Skitka et al. 1999; Buçinca et al. 2021; Green 2022; Sterz et al. 2024; Laux 2024).

Being able to check further does not mean that people will. The first experiment in a published chatbot study required at least one prompt per reasoning question; 113 of 246 participants used exactly one prompt on every question, and the researchers coded 145 participants’ open responses as showing high trust without critical scrutiny (Fernandes et al. 2026). These figures measure prompt use and reported review: a person may inspect an answer without another prompt, and acceptance can vary between answers. Estimating f would require observing whether users keep checking after an agent’s hand-off.

How strongly an agent weighs evidence also needs direct measurement. A study of six language models on medical multiple-choice questions found average calibration errors, the gap between stated confidence and accuracy, of 0.060 to 0.127 (Boie et al. 2026). This motivates testing agents with unreliable confidence but does not show how strongly an agent weighs successive observations, so we keep illustrative values for γ_a . The Supplementary Information explains which measurements would be needed to estimate the model’s parameters.

Company reports give a closer view. Anthropic reported that users approved 93% of Claude Code permission prompts in March 2026 and 97% in August, while rejecting 39% of plans presented before execution (Hughes 2026; Phillippi 2026).² Frequent approval can be sensible when almost all requests are appropriate, so the approval rate alone does not show whether users review commands carefully.

A separate company-commissioned study tested whether users noticed dangerous commands. Each of 1,053 paid testers met one permission prompt whose text had been replaced with a dangerous command; approving it ran nothing harmful. Human testers blocked 13.6% of these commands, compared with 89% for the automatic checker (Phillippi 2026). Plan review comes earlier: rejecting a plan can correct the task itself, while rejecting a command interrupts work under way. Comparing the two would require measuring which harmful actions each prevents.

Review design also matters: a large study of browser warnings found that many warnings were effective and that responses varied with design (Akhawe and Felt 2013). Agents can also screen actions automatically before asking for approval; OpenAI’s agent system card examines whether the agent asks for confirmation when it should, a step before the user’s review (OpenAI 2025b).² Together, these studies suggest that many users do not check further when they could, that approval rates alone cannot show how carefully they review, and that review design and automatic screening offer distinct opportunities to catch errors. The Supplementary Information gives the product settings and usage figures behind these examples.

6.3 Who answers for an approved action

A bare approval records a choice but may reveal little about how it was reached. An agent’s answer can reflect user preferences, training choices and the interests of the organizations deploying it, so the user needs to see its reasons and check its sources. Responsibility must also be divided among users, deployers (who put an agent to use) and providers (who build it). Case studies show that the person operating a system can attract disproportionate blame, while legal liability depends on the role, jurisdiction and kind of harm (Elish 2019).

The law also distinguishes approval from effective oversight. For designated high-risk uses, the European AI Act’s duties include enabling human overseers to understand a system’s limitations, recognize overreliance and override or stop it. EU data-protection guidance distinguishes meaningful human involvement in an automated decision from a token one. Payment law separately addresses consent and authentication.³ An approval process must state what the user authorizes and what review the applicable rules require. The Supplementary Information sets out the scope, application dates and exceptions.

Delegation can also change how others judge a decision. In one experiment, participants judged a bad outcome more leniently when the task had been delegated to a machine than when the delegator had done it personally (Feier et al. 2022). An approval process therefore needs to say what protection, if any, follows from a careful review and how harm is compensated.

Institutions must decide who explains a decision, who prevents harm and who bears its costs (Santoni de Sio and Mecacci 2021). The Yijing passages in Section 2 tie “no blame” to conduct, including the correction of faults, rather than to the outcome. A modern institution can likewise judge the care taken while compensating harm under separate rules. Aviation reporting programs protect specified reports and conduct under stated conditions, and some medical compensation systems cover eligible injuries without proof of negligence.⁴ For such protection to be credible, an organization must state the care expected, appoint a reviewer authorized to uphold it and fund compensation for covered harm.

7 Three rules for the approval button

We propose three rules for actions that cannot be undone. The provider builds the agent and its controls; the deployer sets and enforces review requirements. The family’s move shows how the rules would work.

First, require relevant checks before action. The deployer should specify checks justified by the costs of error and delay, such as independent evidence or review by a qualified second person. Each check should address a specific risk and state what evidence satisfies it; confidence alone does not complete a check. The provider should block execution until the requirements are met or the deployer authorizes an exception. Enforcement at execution matters for users who routinely follow the agent, as the follower scenarios in Section 6 show. Interfaces requiring engagement before accepting AI advice have reduced reliance on incorrect recommendations, though participants found them more complex (Bućinca et al. 2021). For the family, the agent would hold the deposit until the listing and the recipient’s authority were verified. UK rules allowing delays to investigate suspected fraudulent transfers show how extra time can serve this purpose.⁵

Second, let the user examine the evidence and plan before committing. Review should connect the action to its purpose, alternatives and open questions (Green 2022; Sterz et al. 2024; Santoni de Sio and van den Hoven 2018). Approval should specify the permitted action and its conditions. A failed check, conflicting evidence or a material departure from the plan should return the decision to the user. The provider’s controls must enforce these limits during execution and record the checks for later review.

Upfront review can also be ignored. In two studies of Android permissions, 17% of participants paid attention to permissions during installation, and in the online survey 3% answered all three comprehension questions correctly (Felt et al. 2012); field research also found that permission decisions depend on context (Wijesekera et al. 2015). A plan should be short enough to review and specific about what remains unchecked. Questions that direct attention to relevant information are one way to reduce automatic acceptance (Bravo-Lillo et al. 2013). For the family, the plan could show the verified listing, mark the recipient as unchecked and ask which check to complete next. Because an agent may overlook risks or misjudge its answers (Xiong et al. 2024), evaluation should examine users’ reasons, errors detected, harmful actions accepted and beneficial actions rejected.

Third, state how careful decisions will be protected. The same standard of care should apply to acting and declining. A reviewer independent of the immediate decision should assess conduct against a stated standard of reasonable care (Dekker 2009), using recorded checks, disclosed uncertainty and reasons for choosing. The deployer should explain eligibility and arrange compensation, for example through insurance. The provider should preserve the record, though

the record alone creates no immunity.⁵ For the family, a platform could cover a deposit lost to fraud after the required checks were completed, and the family should be able to learn before paying whether this coverage exists and what it requires. UK reimbursement rules for eligible bank-transfer scams offer a partial precedent, detailed in the Supplementary Information.⁵

Users can apply these rules now: list actions that need approval, use permission controls, examine unresolved checks and inspect sources before committing. Keeping a short record of what was checked, and why one acted or held back, also helps. When approving for an organization, establish before accepting responsibility who can pause the task, who reviews a failure and who bears the loss.

8 Conclusion

The Yijing offers a starting point for understanding how advice changes decisions that cannot be undone. Our model separates the value of new evidence from relief of anxiety and fear of blame. It shows why further checks can recover losses caused by overconfidence while costing people who already weigh evidence correctly, and when offering both forms of relief to everyone does at least as well as allocating them at random. The agent simulations extend this to hand-offs: the value of required checks depends on how the agent uses evidence and whether users keep examining its recommendation.

These results support approval processes that require relevant evidence, open the plan to review and protect both careful action and careful refusal. Such protection should be backed by separate arrangements for compensating covered harm. The same design questions arise when advice comes from a person or a committee. Testing these arrangements with users will require measuring the evidence they obtain, the harmful actions and missed opportunities that follow, and their anxiety and fear of blame. Such evidence would show whether the proposed protections are credible and which users benefit from them.

References

- Adler JA (trans and ed) (2020) The original meaning of the Yijing: commentary on the scripture of change, by Zhu Xi. Columbia University Press, New York. <https://doi.org/10.7312/zhu-19124>
- Akhawe D, Felt AP (2013) Alice in Warningland: a large-scale field study of browser security warning effectiveness. In: Proceedings of the 22nd USENIX Security Symposium. USENIX Association, Berkeley, p 257–272. <https://www.usenix.org/conference/usenixsecurity13/technical-sessions/presentation/akhawe>

- Arrow KJ, Blackwell D, Girshick MA (1949) Bayes and minimax solutions of sequential decision problems. *Econometrica* 17:213–244. <https://doi.org/10.2307/1905525>
- Bainbridge L (1983) Ironies of automation. *Automatica* 19:775–779. [https://doi.org/10.1016/0005-1098\(83\)90046-8](https://doi.org/10.1016/0005-1098(83)90046-8)
- Bartling B, Fischbacher U (2012) Shifting the blame: on delegation and responsibility. *The Review of Economic Studies* 79:67–87. <https://doi.org/10.1093/restud/rdr023>
- Berns GS, Chappelow J, Cekic M, Zink CF, Pagnoni G, Martin-Skurski ME (2006) Neurobiological substrates of dread. *Science* 312:754–758. <https://doi.org/10.1126/science.1123721>
- Boie SD, Reis F, Frey N, Grünewald E, Balzer F (2026) Calibration of self-reported confidence and accuracy of large language models in medical question answering. *Journal of Medical Systems* 50:103. <https://doi.org/10.1007/s10916-026-02430-0>
- Boyer P (2020) Why divination? Evolved psychology and strategic interaction in the production of truth. *Current Anthropology* 61:100–123. <https://doi.org/10.1086/706879>
- Bravo-Lillo C, Komanduri S, Cranor LF, Reeder RW, Sleeper M, Downs J, Schechter S (2013) Your attention please: designing security-decision UIs to make genuine risks harder to ignore. In: *Proceedings of the Ninth Symposium on Usable Privacy and Security (SOUPS '13)*. ACM, New York, article 6, p 1–12. <https://doi.org/10.1145/2501604.2501610>
- Buçinca Z, Malaya MB, Gajos KZ (2021) To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW1):188. <https://doi.org/10.1145/3449287>
- Cheng M, Lee C, Khadpe P, Yu S, Han D, Jurafsky D (2026) Sycophantic AI decreases prosocial intentions and promotes dependence. *Science* 391:eac8352. <https://doi.org/10.1126/science.aec8352>
- de Berker AO, Rutledge RB, Mathys C, Marshall L, Cross GF, Dolan RJ, Bestmann S (2016) Computations of uncertainty mediate acute stress responses in humans. *Nature Communications* 7:10996. <https://doi.org/10.1038/ncomms10996>
- Dekker SWA (2009) Just culture: who gets to draw the line? *Cognition, Technology & Work* 11:177–185. <https://doi.org/10.1007/s10111-008-0110-7>
- Durrant S, Li W-y, Schaberg D (trans) (2016) *Zuo tradition / Zuozhuan: commentary on the "Spring and Autumn Annals"*, 3 vols. University of Washington Press, Seattle. <https://doi.org/10.1515/9780295806730>
- Elish MC (2019) Moral crumple zones: cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society* 5:40–60. <https://doi.org/10.17351/ests2019.260>
- Elster J (1989) *Solomonic judgements: studies in the limitation of rationality*. Cambridge University Press, Cambridge

- Feier T, Gogoll J, Uhl M (2022) Hiding behind machines: artificial agents may help to evade punishment. *Science and Engineering Ethics* 28:19. <https://doi.org/10.1007/s11948-022-00372-7>
- Felt AP, Ha E, Egelman S, Haney A, Chin E, Wagner D (2012) Android permissions: user attention, comprehension, and behavior. In: *Proceedings of the Eighth Symposium on Usable Privacy and Security (SOUPS '12)*. ACM, New York, article 3, p 1–14. <https://doi.org/10.1145/2335356.2335360>
- Fernandes D, Villa S, Nicholls S, Haavisto O, Buschek D, Schmidt A, Kosch T, Shen C, Welsch R (2026) AI makes you smarter but none the wiser: the disconnect between performance and metacognition. *Computers in Human Behavior* 175:108779. <https://doi.org/10.1016/j.chb.2025.108779>
- Forден G, Podvig P, Postol TA (2000) False alarm, nuclear danger. *IEEE Spectrum* 37:31–39. <https://doi.org/10.1109/6.825657>
- Gilovich T, Medvec VH (1995) The experience of regret: what, when, and why. *Psychological Review* 102:379–395. <https://doi.org/10.1037/0033-295X.102.2.379>
- Green B (2022) The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review* 45:105681. <https://doi.org/10.1016/j.clsr.2022.105681>
- Hamman JR, Loewenstein G, Weber RA (2010) Self-interest through delegation: an additional rationale for the principal-agent relationship. *American Economic Review* 100:1826–1846. <https://doi.org/10.1257/aer.100.4.1826>
- Hoffman D (1999) "I had a funny feeling in my gut". *The Washington Post*, 10 Feb 1999, p A19. <https://www.washingtonpost.com/wp-srv/inatl/longterm/coldwar/shatter021099b.htm>. Accessed 6 Sep 2026
- Hughes J (2026) How we built Claude Code auto mode: a safer way to skip permissions. Anthropic, 25 Mar 2026. <https://anthropic.com/engineering/claude-code-auto-mode>. Accessed 7 Sep 2026
- Kiener M (2022) Can we bridge AI's responsibility gap at will? *Ethical Theory and Moral Practice* 25:575–593. <https://doi.org/10.1007/s10677-022-10313-9>
- Kruglanski AW, Webster DM (1996) Motivated closing of the mind: "seizing" and "freezing". *Psychological Review* 103:263–283. <https://doi.org/10.1037/0033-295X.103.2.263>
- Laux J (2024) Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act. *AI & Society* 39:2853–2866. <https://doi.org/10.1007/s00146-023-01777-z>
- Malinowski B (1948) *Magic, science and religion and other essays*. Free Press, Glencoe

- Matthias A (2004) The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6:175–183. <https://doi.org/10.1007/s10676-004-3422-1>
- Moore DA, Healy PJ (2008) The trouble with overconfidence. *Psychological Review* 115:502–517. <https://doi.org/10.1037/0033-295X.115.2.502>
- Moore OK (1957) Divination—a new perspective. *American Anthropologist* 59:69–74. <https://doi.org/10.1525/aa.1957.59.1.02a00060>
- OpenAI (2025a) Expanding on what we missed with sycophancy. OpenAI, San Francisco, 2 May 2025. <https://openai.com/index/expanding-on-sycophancy/>. Accessed 7 Sep 2026
- OpenAI (2025b) ChatGPT agent system card. OpenAI, San Francisco, 17 Jul 2025. https://cdn.openai.com/pdf/839e66fc-602c-48bf-81d3-b21eacc3459d/chatgpt_agent_system_card.pdf. Accessed 7 Sep 2026
- Parasuraman R, Manzey DH (2010) Complacency and bias in human use of automation: an attentional integration. *Human Factors* 52:381–410. <https://doi.org/10.1177/0018720810376055>
- Park GK (1963) Divination and its social contexts. *Journal of the Royal Anthropological Institute of Great Britain and Ireland* 93:195–209. <https://doi.org/10.2307/2844242>
- Phillippi C (2026) Auto mode is now the default in Claude Code for Pro, Max, and Team plans. Anthropic, 7 Aug 2026. <https://claude.com/blog/auto-mode-default-in-claude-code>. Accessed 10 Sep 2026
- Raphals L (2013) Divination and prediction in early China and ancient Greece. Cambridge University Press, Cambridge. <https://doi.org/10.1017/CBO9780511863233>
- Ritov I, Baron J (1990) Reluctance to vaccinate: omission bias and ambiguity. *Journal of Behavioral Decision Making* 3:263–277. <https://doi.org/10.1002/bdm.3960030404>
- Ruan Y (阮元) (ed) (1980) Shisanjing zhushu (十三經注疏) [The Thirteen Classics with commentaries and subcommentaries]. Zhonghua shuju, Beijing [in Chinese]. Cited for the Zhouyi (周易), including the Xici (繫辭) A and B and the Shuogua (說卦), and for the Shangshu (尚書), Hongfan (洪範); English translation of the Zhouyi in Wilhelm and Baynes (1967)
- Santoni de Sio F, Mecacci G (2021) Four responsibility gaps with artificial intelligence: why they matter and how to address them. *Philosophy & Technology* 34:1057–1084. <https://doi.org/10.1007/s13347-021-00450-x>
- Santoni de Sio F, van den Hoven J (2018) Meaningful human control over autonomous systems: a philosophical account. *Frontiers in Robotics and AI* 5:15. <https://doi.org/10.3389/frobt.2018.00015>

- Skitka LJ, Mosier KL, Burdick M (1999) Does automation bias decision-making? *International Journal of Human-Computer Studies* 51:991–1006. <https://doi.org/10.1006/ijhc.1999.0252>
- Smith K Jr (1989) Zhouyi interpretation from accounts in the Zuozhuan. *Harvard Journal of Asiatic Studies* 49:421–463. <https://doi.org/10.2307/2719259>
- Smith K Jr, Bol PK, Adler JA, Wyatt DJ (1990) *Sung dynasty uses of the I Ching*. Princeton University Press, Princeton
- Sosis R, Handwerker WP (2011) Psalms and coping with uncertainty: religious Israeli women's responses to the 2006 Lebanon War. *American Anthropologist* 113:40–55. <https://doi.org/10.1111/j.1548-1433.2010.01305.x>
- Sparrow R (2007) Killer robots. *Journal of Applied Philosophy* 24:62–77. <https://doi.org/10.1111/j.1468-5930.2007.00346.x>
- Spranca M, Minsk E, Baron J (1991) Omission and commission in judgment and choice. *Journal of Experimental Social Psychology* 27:76–105. [https://doi.org/10.1016/0022-1031\(91\)90011-T](https://doi.org/10.1016/0022-1031(91)90011-T)
- Sterz S, Baum K, Biewer S, Hermanns H, Lauber-Rönsberg A, Meinel P, Langer M (2024) On the quest for effectiveness in human oversight: interdisciplinary perspectives. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. ACM, New York, p 2495–2507. <https://doi.org/10.1145/3630106.3659051>
- Stone P (2010) Three arguments for lotteries. *Social Science Information* 49:147–163. <https://doi.org/10.1177/0539018409359362>
- Tigard DW (2021) There is no techno-responsibility gap. *Philosophy & Technology* 34:589–607. <https://doi.org/10.1007/s13347-020-00414-7>
- Wald A (1947) *Sequential analysis*. Wiley, New York
- Wijesekera P, Baokar A, Hosseini A, Egelman S, Wagner D, Beznosov K (2015) Android permissions remystified: a field study on contextual integrity. In: *Proceedings of the 24th USENIX Security Symposium*. USENIX Association, Berkeley, p 499–514. <https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/wijesekera>
- Wilhelm R, Baynes CF (trans) (1967) *The I Ching or Book of Changes*, 3rd edn. Princeton University Press, Princeton
- Xiong M, Hu Z, Lu X, Li Y, Fu J, He J, Hooi B (2024) Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In: *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*. <https://iclr.cc/virtual/2024/poster/18135>

Yang B (楊伯峻) (ed) (1990) Chunqiu Zuozhuan zhu (春秋左傳注) [The Zuozhuan annotated], 2nd edn. Zhonghua shuju, Beijing [in Chinese]. Cited as Zuozhuan, Huan 11 and Xiang 9; English translation in Durrant et al. (2016)

Notes

1. Casting probabilities and their relation to the model verdict are given in the Supplementary Information under “Casting probabilities and the model verdict”.
2. The approval and plan-rejection rates come from successive company reports and describe different stages of work. The dangerous-command study used paid testers and disabled harmful execution. These rates describe those reported settings. Further product figures, survey details and the distinction between requesting confirmation and reviewing a request appear in the Supplementary Information under “Product reports and review settings”.
3. The Supplementary Information, under “Legal scope and application dates”, sets out the product-liability, human-oversight, data-protection and payment provisions discussed in Section 6, including their scope and phased application.
4. The Supplementary Information, under “Protection for reporting and compensation for harm”, explains the aviation and medical examples in Section 6. Eligibility, exceptions and the route to compensation differ across these arrangements.
5. The Supplementary Information, under “Expert reliance and payment decisions”, gives further background on decision records, reliance on experts and the payment examples in Section 7, including reimbursement limits.

Figure legends

Fig. 1. How a short run of favorable observations can prompt an early decision. This selected simulation illustrates the “phantom pattern”: the state is bad, but the first two observations favor the good state. The black line shows accumulated evidence in log-odds calculated with correct weighting, with one observation at each step. The dashed lines mark the early action thresholds for the calibrated person (blue) and the overconfident person (orange), expressed in these same evidence units. The overconfident person gives each observation four times its correct weight and acts at step 2 (circled), making a final commitment. By step 3, the evidence has fallen below that person’s threshold, and by step 4 it strongly favors the bad state. The shaded area marks two required steps of additional evidence collection after the first observation. Under this requirement, the earliest decision is at step 3. The example was selected using the following criteria: the overconfident person acts at step 2 while the calibrated person has yet to act, and the evidence falls below the overconfident threshold at step 3 before later turning against the good state. Fig. 2 reports how often errors occur across the simulations.

Fig. 2. The effects of required waiting depend on how a person weights evidence. The four settings differ only in the multiplier applied to each observation: $\gamma = 1$ gives correct weighting, while $\gamma = 2, 4$ and 6 give increasing overweighting. Each person must collect k additional observations after the first before deciding. a, Actions in the bad state, labeled “avoidable catastrophes” in the plot. For $\gamma = 4$, two additional steps reduce the rate from 4.5% to 2.1%, and three reduce it to 1.5%. The corresponding rates for $\gamma = 6$ are 6.1%, 2.7% and 1.9%. For the calibrated person, two steps reduce the rate from 0.22% to 0.19%. The dashed line counts failures after action in the good state, which remain between 2.1% and 2.5% of all decisions for $\gamma = 4$. The risk of failure conditional on such an action stays at one in twenty; waiting changes how often the person acts in the good state. b, Missed opportunities, defined as giving up in the good state. For $\gamma = 4$, two additional steps reduce the rate from 7.3% to 3.1%, because required waiting also prevents early abandonment. c, Average payoff. Two additional steps lower the calibrated person’s payoff from 4.75 to 4.57. For $\gamma = 4$, payoff rises from 1.04 to 3.28 with three additional steps and is approximately unchanged at four. For $\gamma = 6$, it rises from -0.57 to 3.08 with four additional steps. All error rates use all decisions, including both states. Shaded bands show 95% Monte Carlo intervals from 100,000 decisions per point.

Fig. 3. Verdicts can ease different pressures by changing the perceived cost of waiting or acting. The plots show missed opportunities for the anxious person (a), whose total waiting cost is four times the material cost, and the fearful person (b), whose total failure cost is three times the material loss. Each plot compares no device, a random verdict with no relief ($\theta = 0$), and a verdict that reduces the relevant pressure by 70% ($\theta = 0.7$). Verdicts arrive after the first observation, with no required wait and no belief shift; the person can continue collecting evidence. A “go” verdict occurs with probability one fourth. Verdict-specific percentages use all decisions receiving that verdict, including both states; the no-device percentage uses all decisions for that type. The dashed line shows the calibrated person’s rate of 1.8%. With relief, anxious people receiving “wait” miss opportunities in 4.0% of decisions, compared with 10.7% without the device. Those receiving “go” keep the same policy and have a rate of 10.9%, within sampling error. For fearful people receiving “go”, the missed-opportunity rate falls from 13.8% to 2.7%; those receiving “wait” keep the same policy and a rate of 13.8%. Even after “go”, the fearful person acts according to the evidence. Their rate of action in the bad state rises from 0.05% to 0.11%, remaining below the calibrated rate of about 0.2%. Error bars show 95% Monte Carlo intervals. Each type has 100,000 simulated decisions, of which approximately 25,000 receive “go”. Sampling variation can produce small differences between verdict groups even when their policies are unchanged.

Tables

Table 3. Outcomes for equal shares of the four types in Table 1. Error rates use all decisions, including both states. “Own judgment” uses each person’s policy. “The procedure” requires two additional observations, then a random verdict (“go” with probability $\frac{1}{4}$) with 70% relief ($\theta = 0.7$). “Interval only” omits the verdict; “relief only” omits required waiting. In “agent-timed decision”, an agent with $\gamma_a = 4$ collects evidence until its policy stops; the user then acts or gives up without further checking. “Single observation” requires a decision after the first observation.

Success pays 20 (Table 2). “Act at once” and “single observation” incur one observation cost; “never act” incurs none. The upper block doubles the material observation cost from 0.5 to 1. The anxious person’s extra cost stays fixed, giving total waiting costs of four and two and a half times the material cost, respectively.

	act at once	never act	own judgment	the procedure	interval only	relief only	agent-timed decision	single observation
a step of waiting = 1% of the failure loss								
actions in the bad state	50%	0	1.60%	0.87%	0.89%	1.57%	2.17%	2.30%
missed opportunities	0	50%	15.9%	5.4%	6.6%	12.7%	25.1%	32.4%
average payoff	-41.0	0	1.41	1.90	1.80	1.68	-0.02	-0.70
a step of waiting = 0.5% of the failure loss (main text)								
actions in the bad state	50%	0	1.39%	0.74%	0.78%	1.33%	2.12%	2.30%
missed opportunities	0	50%	8.4%	2.8%	3.5%	6.5%	23.1%	32.4%
average payoff	-40.5	0	3.26	3.97	3.89	3.48	0.95	-0.20

Each type has 100,000 simulated decisions with shared random streams. The 95% Monte Carlo intervals use the per-decision average across types: half-widths are at most 0.13 payoff units, 0.07 percentage points for bad-state actions and 0.25 percentage points for missed opportunities.

Supplementary Table S7 gives paired comparisons with own judgment. In the main setting, payoff changes are $+0.71 \pm 0.05$ (procedure), $+0.62 \pm 0.04$ (interval), $+0.22 \pm 0.02$ (relief) and -2.32 ± 0.06 (agent-timed decision). Procedure gains by type are -0.18 ± 0.02 (calibrated), $+1.89 \pm 0.11$ (overconfident), $+0.60 \pm 0.07$ (anxious) and $+0.53 \pm 0.06$ (fearful).

Data availability

This study used no human-participant, survey or personal data. All data underlying the figures and tables are outputs of the simulations described in Section 3 and the Supplementary Information. The results files behind every figure and table of the article and the Supplementary Information

(CSV files, one per figure or table), the simulation code, the fixed random seeds, the automated tests and a script that regenerates every number and figure are openly available at <https://github.com/shijunzhe/oracle-interval> and archived at Zenodo (<https://doi.org/10.5281/zenodo.22967587>). The README file maps each figure and table to its results file and script. The historical sources are the published texts and translations cited in the reference list; the product figures quoted in Section 6 are taken from the cited public reports, with the settings and dates given in the Supplementary Information.

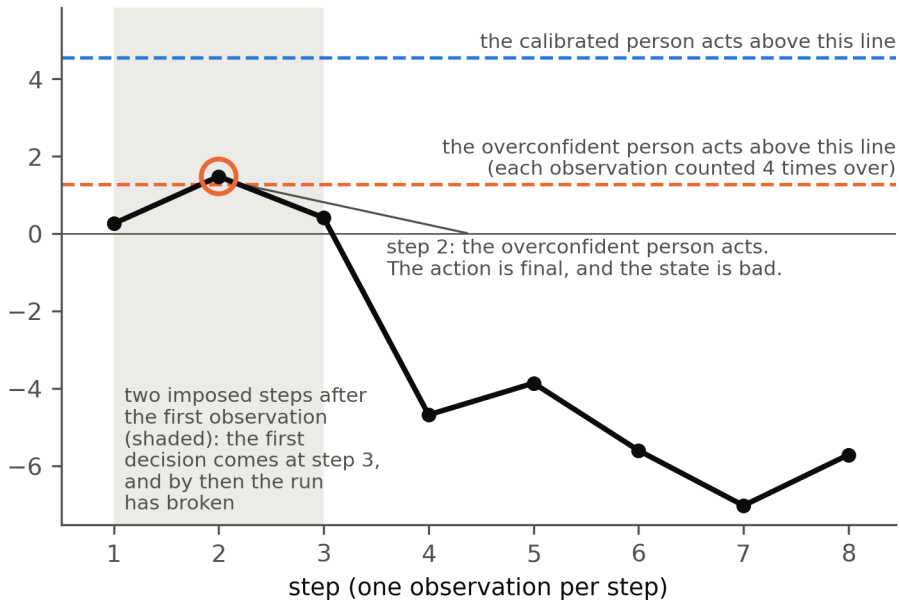
Ethical approval

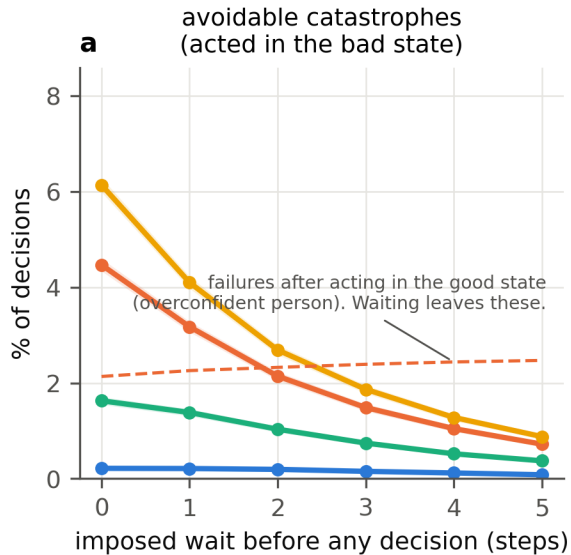
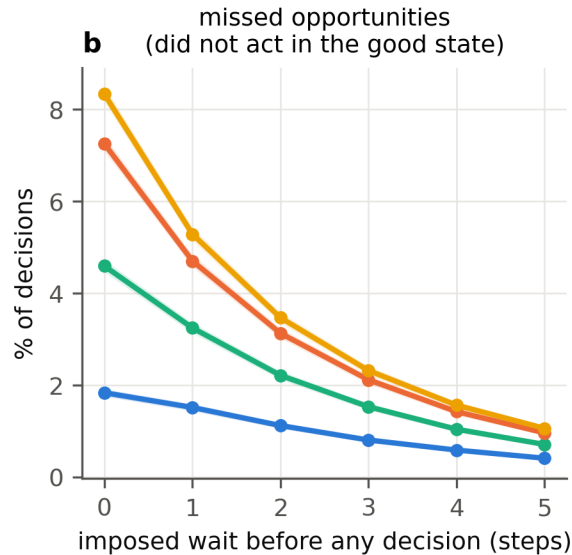
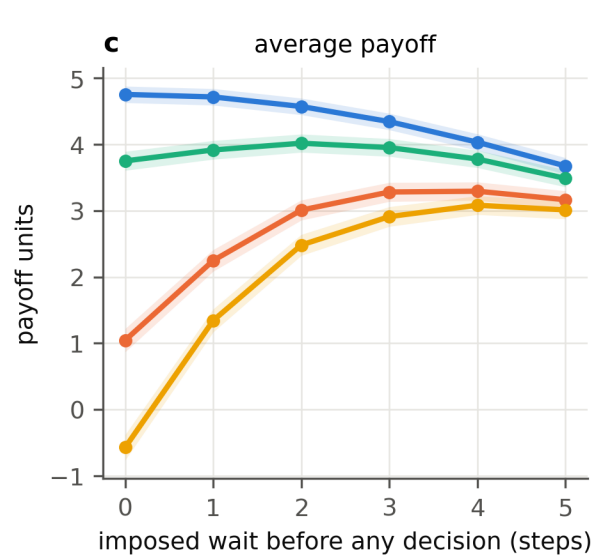
Not applicable. This study did not involve human participants, human material, animals or personal data. It consists of a historical analysis of published texts and computational simulations of a decision model, together with publicly reported aggregate statistics. Under the policies of the University of California, Berkeley Committee for Protection of Human Subjects, research that does not involve human subjects is not subject to review; no approval or exemption number therefore exists. The research was performed in accordance with the editorial policies of Springer Nature and the research-integrity policies of the authors' institutions. The Declaration of Helsinki does not apply, as no human participants were involved.

Informed consent

Not applicable. No human participants were involved, so informed consent was not required. No participant under 16 years of age was involved.

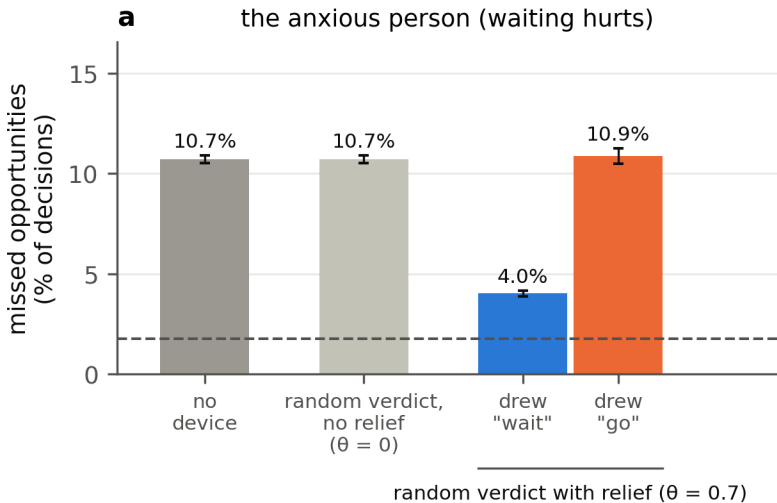
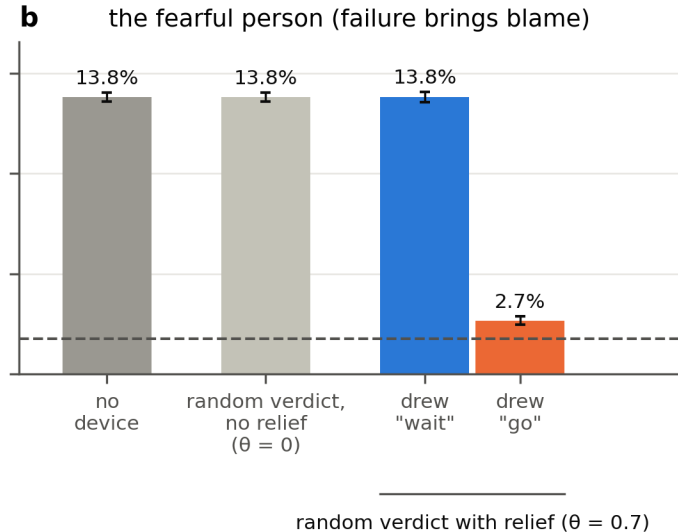
evidence that the state is good
(log-odds; 0 means even)





— calibrated ($\gamma = 1$) — mildly overconfident ($\gamma = 2$) — overconfident ($\gamma = 4$) — very overconfident ($\gamma = 6$) - - - failures after acting in the good state, $\gamma = 4$ (panel a)

--- dashed line: the calibrated person, 1.8%



Supplementary Files

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

- [supplementaryinformation2026.09.25.docx](#)