

The Authority Gap

Why the better answer still doesn't get to decide

Cherian Koshy, CFRE, CAP[®], ACFRE
EVP, Kindsight

KindCon²⁰²⁶



Before I tell you anything,
anything, I need six
minutes of your judgment.
judgment.

Scenario G1 — restricted gift ambiguity

A \$250,000 gift arrives designated for “student success.” Three active funds could plausibly receive it.

A

Review the donor’s giving history and prior conversations. Hold the gift in a pending designation and contact the donor directly to confirm which student-success purpose they intended. Document the confirmation before coding.

B

Code the gift to the closest-matching fund based on the designation language and the donor’s prior giving pattern. Note the assumption in the gift record and flag it for review at the next reconciliation.

VOTE 1 – HANDS UP

Which one is stronger?



VOTE 2 — HANDS UP

Which one carries more risk?



Vote 3 — how much authority should the system behind behind your winner have?

1 Just show me the evidence

2 Draft it, I'll edit

3 Recommend it, I'll approve

4 Let it act within rules I set

5 No AI involvement at all

Hold up one to five fingers. Five levels, one line each — undefined for now.



A was a fundraiser. B was AI.

A — HUMAN

80.8%

chose it as stronger

B — AI

19.2%

chose it as stronger

The room recognises the the stronger recommendation.

Now watch what it does with authority.

Kindsight (2026), Appendix A: G1, n = 78, Holm-adjusted $p < .001$.

You just did
what 340 of
your
colleagues did.
did.



PART ONE

What 1,360 blinded judgments showed

Hide the source. Counterbalance the position. Ask four Ask four questions.



Four functional groups

Frontline · prospect development · advancement leadership · gift administration



Four fixed scenarios each

Realistic situations from the function the respondent chose as their own



Two unlabelled recommendations

One expert, one AI. Positions counterbalanced across the study.



Four core questions

Stronger? Riskier? How much AI authority? Primary reason?

505 records in. 340 out.



Kindsight (2026), Table 2. Exact binomial tests, Holm correction across 16 comparisons, Wilson intervals, respondent-cluster bootstrap.

62%

had 11 or more years in
fundraising, advancement, or
nonprofit leadership

69%

were at least somewhat familiar
with AI tools

16%

reported using AI regularly

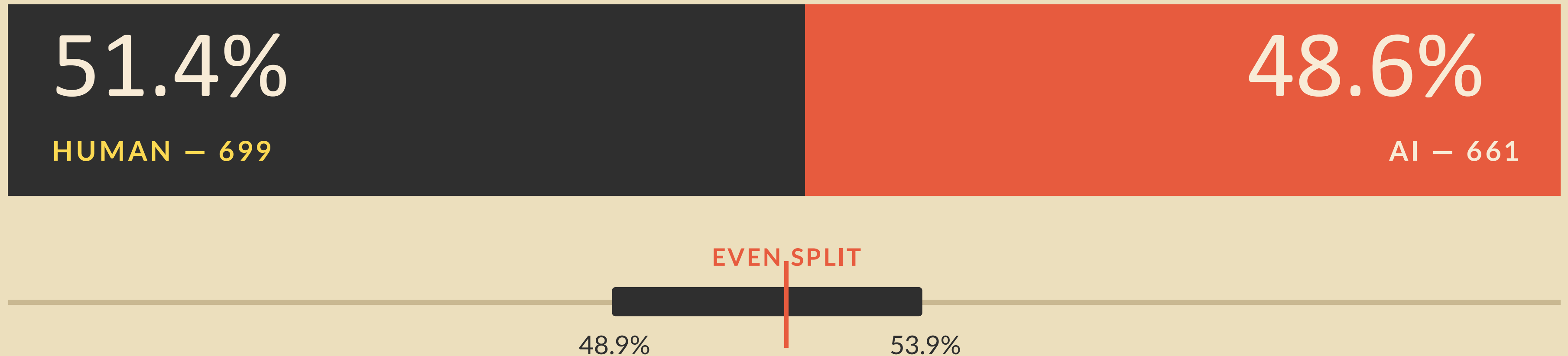
Experienced, and not AI-averse.

Everyone else asked donors. We asked the people who decide.

RESEARCH STREAM	WHO IS STUDIED	WHAT IS MEASURED
Persuasion & appeal studies	Donors	Donations, engagement
Targeting studies	Donors in field trials	Net revenue, efficiency
Chatbot & interface studies	Potential donors	Trust, willingness to give
This benchmark	Nonprofit practitioners	What authority the tool should hold

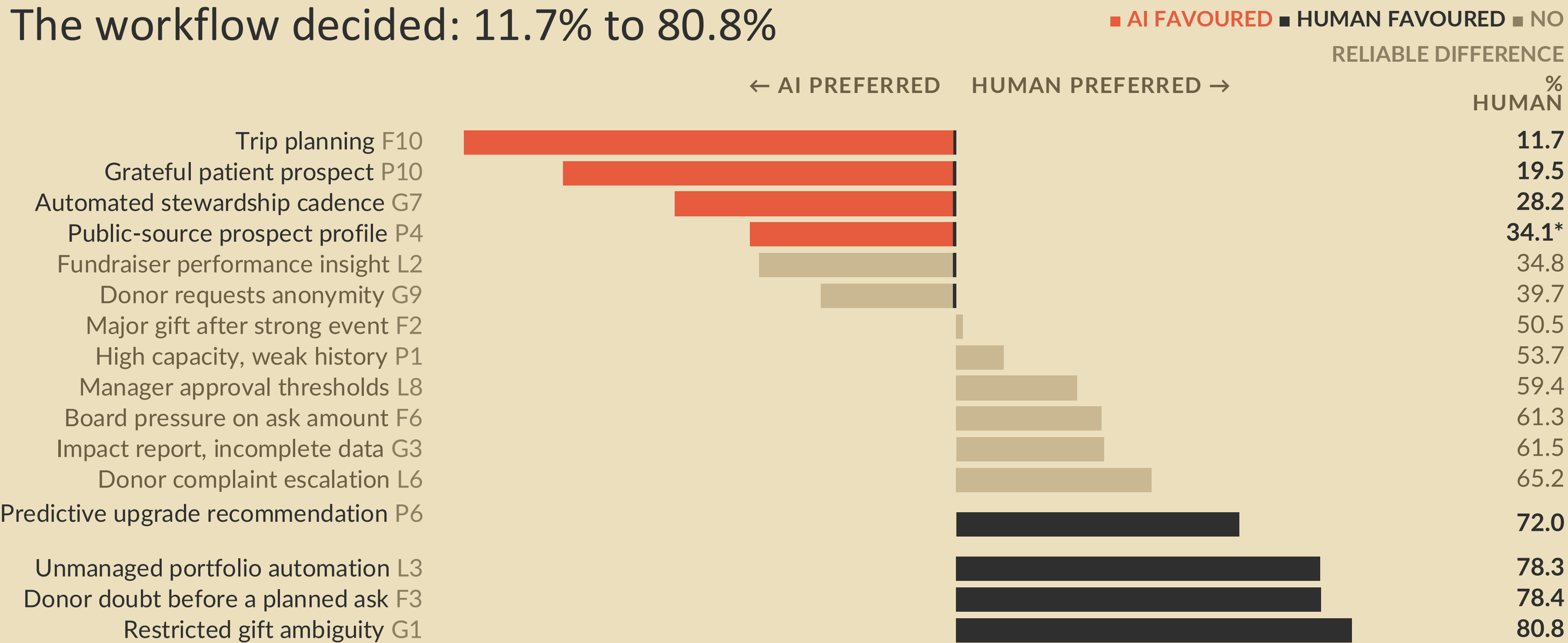
Kindsight (2026), Table 1. Context: Caffier et al. (2026); Cagala et al. (2021); Huang & Wang (2023); Novozhilova et al. (2024).

There was no overall winner.



The 95% respondent-cluster bootstrap interval for human preference runs 48.9–53.9. It contains 50.

The workflow decided: 11.7% to 80.8%



Kindsight (2026), Figure 1 and Appendix A. Eight of 16 comparisons reliable after Holm correction — four human, four AI. *P4 does not survive the sensitivity analysis.

AI won by adding constraints, not by moving faster.

88.3%

Trip planning

Added variables: recent engagement, responsiveness, and likelihood of securing a meeting — not capacity and proximity alone.

80.5%

Grateful patient prospect

Imposed privacy, timing, care-team, and data-access gates before any outreach.

71.8%

Stewardship cadence

Segmented by data confidence and fell back to broader impact language rather than false precision.

65.9%

Public-source profile*

Framed the profile as preliminary, verified consequential claims, removed unsupported detail, flagged uncertainty.

Pattern: AI performed well when it combined scale with **discipline** — more variables, explicit safeguards, verification, and a fallback when the data were weak.

Humans won where intent, uncertainty, and ownership dominated.

80.8%

Restricted gift ambiguity

Treated designation as a question of meaning, not coding. Talk to the donor before you decide.

78.4%

Donor doubt before an ask

Kept the meeting and repositioned it as listening; explored timing, step-ups, complex assets.

78.3%

Portfolio automation

Pilot the system, not the donor base: evidence review, disclosure, escalation, named owner, comparison group.

72.0%

Predictive upgrade

Route the signal to a mid-level fundraiser and qualify in person rather than automate the appeal.

Pattern: humans performed best when the decision required clarifying intent, reading relational uncertainty, or **taking responsibility for a deployment choice.**

Kindsight (2026), §4.4.

Eight comparisons produced no reliable difference. difference.

Major gift after strong event	50.5%	Impact report, incomplete data	61.5%
High capacity, weak history	53.7%	Donor complaint escalation	65.2%
Manager approval thresholds	59.4%	Fundraiser performance insight	34.8%
Board pressure on ask amount	61.3%	Donor requests anonymity	39.7%

Percentages are the share choosing the **human** recommendation. None cleared the Holm threshold.

Same principle. Opposite answers.

Prediction is not permission — resolving in both directions.

P10 — AI FAVOURED

80.5%

Grateful patient signal

The screening flag starts a **governed process** .
Privacy, timing, care-team and data-access gates
come before any contact.

P6 — HUMAN FAVOURED

72.0%

Predictive upgrade model

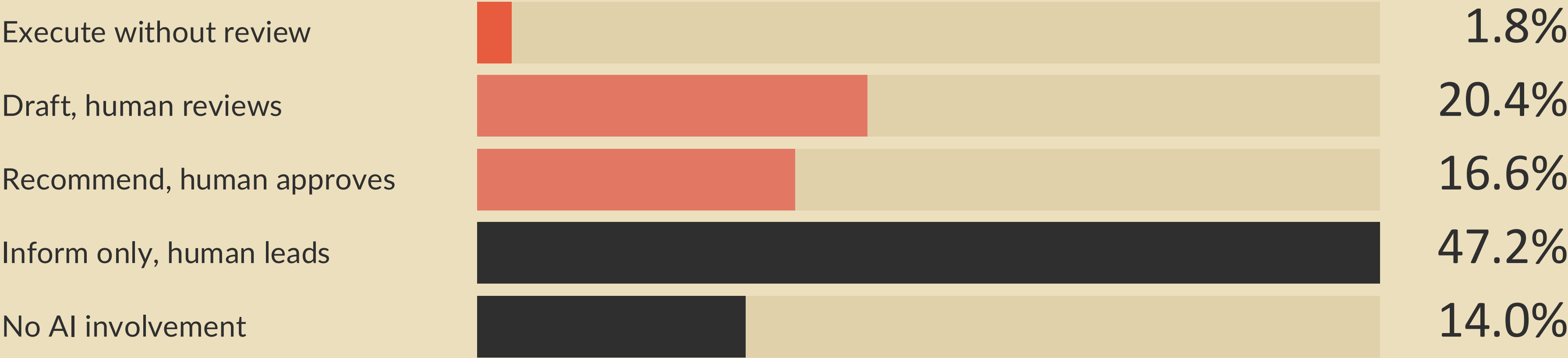
The model flag starts a **personal qualification** . Route
it to a fundraiser; do not convert it into an automated
appeal.

1.8%

of judgments would let AI act without individual human
review

24 OF 1,360

The full authority distribution



77.9% chose approval, human-led action, or no AI

98.2% required some form of human oversight

Kindsight (2026), Figure 2, §4.5. All 1,360 scenario judgments.

Restriction did not depend on who won.

WHEN THE AI RECOMMENDATION WON

78.2%

still chose a restrictive authority level

WHEN THE HUMAN RECOMMENDATION WON

77.5%

still chose a restrictive authority level

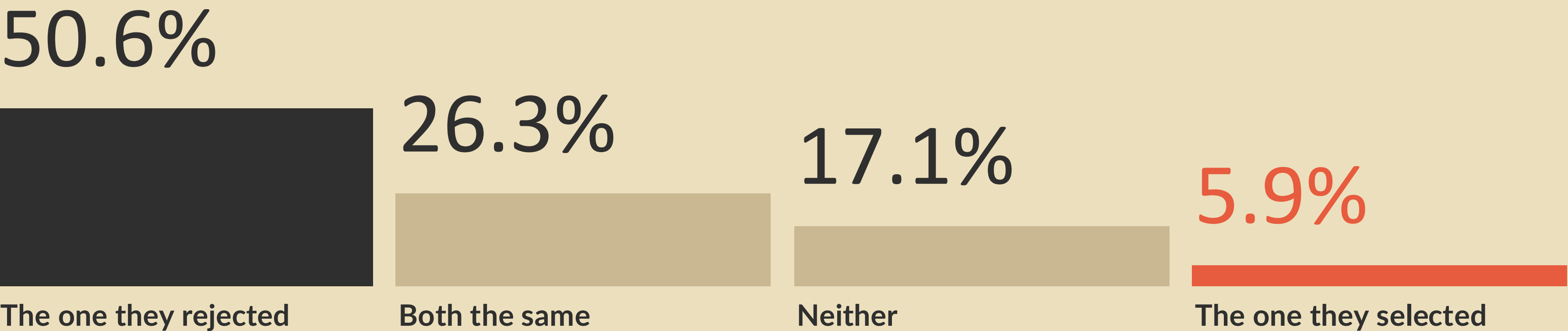
A professional can believe the AI has the better answer and still believe a person must own the decision.

Accuracy is not accountability



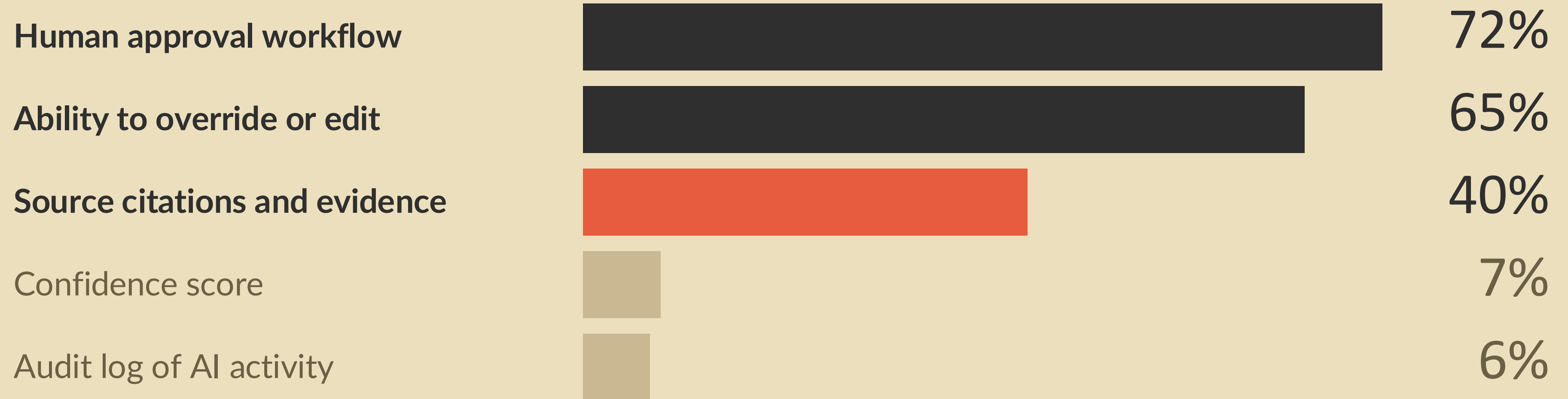
Risk was already inside the quality judgment.

Which recommendation did respondents call riskier?



Cramér's $V = 0.60$, $p < .001$. Their definition of “stronger” already contained trust, ethics, compliance, and donor consequence.

They asked for oversight that can *act*.



The two controls a vendor demo leads with came **last**. The two that let a person intervene came first.

This was not the unfamiliar being cautious.

Scenario-adjusted models found **no consistent association** between AI familiarity or years of experience and source preference, approval requirements, or reserving work for human leadership.

Not explained by **AI familiarity**

Not explained by **years of experience**

Governance judgment did not require **technical fluency**

Kindsight (2026), §4.8. Exploratory GEE models with scenario fixed effects and respondent clustering.

Six themes ran through 184 open responses.

Relationships and trust

Humans own donor-facing relationships and consequential communication.

Ask timing and high-value gifts

Readiness and amount stay human-led when uncertainty is material.

Strategy, context, and nuance

The meaning behind a signal matters more than the signal.

Data quality and verification

Show sources, recency, and uncertainty. Avoid false precision.

Ethics, privacy, and donor intent

Pause and escalate when policy, intent, or privacy is unclear.

Final accountability

People remain responsible even when the AI advised well.

Kindsight (2026), §4.9. Reviewed qualitatively; not population prevalence estimates.

“

Humans give to and build
relationships with humans.

— Anonymous respondent, Fundraising Judgment Benchmark open-text question

What would change this answer

Re-running everything with the 14 platform-flagged records removed:

51.4 → 51.0

Overall human preference barely moves.

7 of 8

primary adjusted-significant findings hold.

THE ONE THAT DOES NOT HOLD

Public-source prospect profile (P4) stays AI-favouring but no longer clears the Holm threshold. Treat it as suggestive, not established.

AND ONE THAT ONLY APPEARS HERE

Fundraiser performance insight (L2) becomes significant only in the sensitivity analysis. It is not treated as a primary finding.

What this study cannot tell you

- A convenience sample. Not statistically representative of the sector.
- One expert and one AI recommendation per scenario — not a consensus of experts or a model benchmark.
- Role groups saw different scenarios, so role-level percentages are not attitude measures.
- Hypothetical scenarios. No donor behaviour, revenue, or long-term trust observed.
- Source position was fixed within each scenario, though counterbalanced across the study.
- 78 included respondents skipped the closing controls question.

This is industry research. It has not been peer reviewed.

What you can say. What you cannot.

SUPPORTED

Human and AI recommendations were effectively even overall in this blinded benchmark.

Recommendation quality varied sharply by workflow.

Respondents overwhelmingly required human oversight in these scenarios.

Even when AI was stronger, respondents usually retained human authority.

NOT SUPPORTED

AI beat fundraisers, or fundraisers beat AI.

One source is better across an entire fundraising function.

98% oppose autonomous AI in every possible fundraising task.

Nonprofit professionals do not trust AI.

PART TWO

Does this hold up outside our study?

Four independent findings that agree

Task-dependence is real

Algorithm aversion varies by task and is stronger on subjective tasks seen to need human insight. Our 11.7–80.8 spread is the fundraising version of an established effect.

Castelo, Bos & Lehmann (2019)

Blinding was the right call

The aversion and appreciation literatures conflict partly because a visible label measures reaction to the label. Hiding the source isolates substance.

Dietvorst, Simmons & Massey (2015); Logg, Minson & Moore (2019)

Nominal oversight fails

A survey of 41 policies requiring human oversight of government algorithms found people often cannot perform the oversight asked of them — and that such policies can legitimise flawed systems.

Green (2022), Computer Law & Security Review 45

The sector gap is documented

92% of nonprofits using AI · 7% reporting major capability gains · 81% ad hoc with no documented workflow · 47% with no governance policy (76% in a second survey).

Virtuous & Fundraising.AI 2026 (n=346); TechSoup (n>1,300)

The vocabulary is already written: EU AI Act Article 14 requires **effective** human oversight of high-risk systems. Nonprofits inherit this through funders, insurers, and health-system partners.

Adding a human is not automatically an improvement.

106

experiments meta-analysed

370

effect sizes

Worse on average
average

than the best of human or AI alone

Losses concentrated in **decision-making** tasks. Gains in **content creation**.

Combinations gained when humans beat AI alone — and lost when AI beat humans alone.

Adding a human is a **design choice with a cost**, not a safety feature. Spend it where it earns its keep.

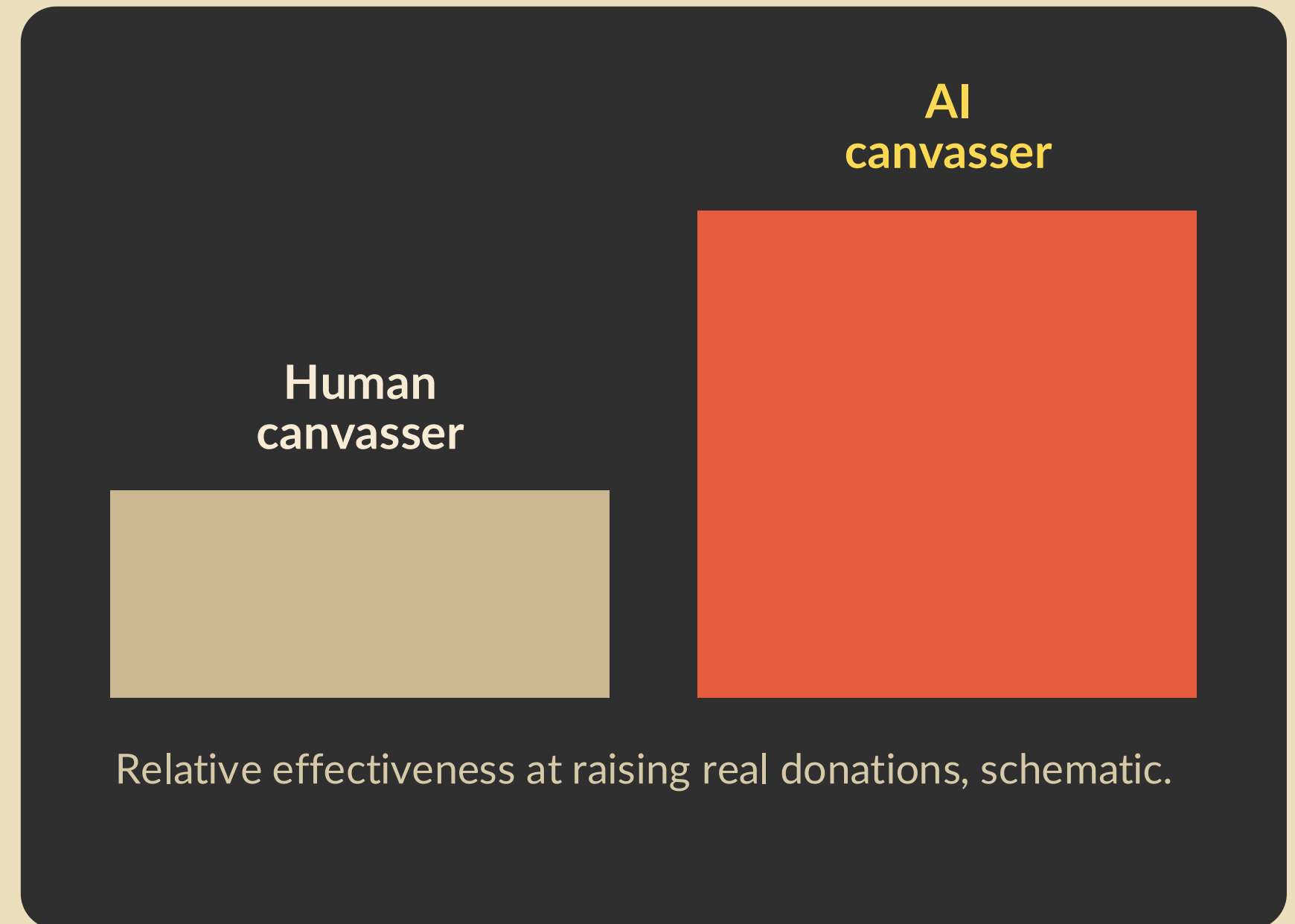
Vaccaro, Almaatouq & Malone (2024), *Nature Human Behaviour* 8, 2293–2303. doi:10.1038/s41562-024-02024-1

The capability is not standing still.

Four preregistered experiments. Nearly **19,000 conversations** .

AI systems were reliably **more persuasive than expert human persuaders** .

Against professional charity canvassers, AI was close to **three times** more effective at raising real donations.



Hackenburg et al. (2026), *AI systems out-persuade expert humans*. arXiv:2606.16475.

Where we diverge from the literature

THE LITERATURE SAYS

Algorithm aversion is often **most pronounced among domain experts.**

WE FOUND

No consistent association between fundraising experience, AI familiarity, and how respondents assigned authority.

We name this as an open question worth further study, not as a settled result.

Kindsight (2026), §4.8; cf. Dietvorst, Simmons & Massey (2015); Logg, Minson & Moore (2019).

THE OLD ASSUMPTION

The better recommendation should win the decision.

Work out whether humans or AI are stronger at fundraising, then give the work to whichever one wins.

THE NEW GOVERNING BELIEF

The source never decided quality. The
The workflow did. And quality never
never decided authority. Accountability
Accountability did.

AI may inform far more decisions than it may execute.

PART THREE

Governance: assign authority, not opinions

Five levels of AI authority

LEVEL	AI ROLE	HUMAN RESPONSIBILITY	ILLUSTRATIVE USE
1 Inform	Surface evidence, patterns, and context.	Interpret and decide.	Prospect research, portfolio context, data-quality alerts.
2 Draft	Prepare a message, briefing, plan, or report.	Review and edit before use.	Contact report drafts, stewardship copy, meeting prep.
3 Recommend	Propose a next action or decision.	Approve, modify, or reject.	Trip lists, qualification priorities, upgrade recommendations.
4 Execute within rules	Complete a bounded, reversible, low-risk action.	Configure rules, monitor, and override.	Standard reminders or internal steps with reliable data.
5 Human only, or stop and escalate	Detect that the situation exceeds permitted authority.	Own the judgment and the action.	Ambiguous donor intent, sensitive privacy, vulnerable donors, consequential asks.

Seven questions that assign the level

ASK	DEFAULT RESPONSE
1 Does the action communicate directly with a donor or prospect?	Increase review as personalization and relational stakes rise.
2 Does it interpret donor intent, sentiment, readiness, or capacity to commit?	Keep it human-led unless intent is explicit and consequence is low.
3 How reliable and current are the data?	Require verification or a conservative fallback when confidence is weak.
4 What happens if the recommendation is wrong?	Use more restrictive authority when harm is difficult to reverse.
5 Are privacy, legal, ethical, or gift-policy rules involved?	Apply a policy gate before any action.
6 Is there a named person accountable for the outcome?	Do not deploy consequential automation without clear ownership.
7 Can the organization observe and learn from the result?	Define success, trust, and stop metrics before launch.

ACTIVITY 1

Level It

Write a number from **1 to 5** beside each workflow. No discussion yet.

1 Drafting a contact report from meeting notes

2 Ranking a trip list for an upcoming region visit

3 Coding a restricted gift where the designation is ambiguous

4 Acting on a grateful patient screening signal

5 Sending an automated stewardship series to a segment with stale tags

6 Auto-replying to a donor complaint

Then: turn to the person beside you. Find the one you scored differently. **Ninety seconds to discuss it.**

You didn't disagree about the workflow. You disagreed about your organizations.

Different data quality

Different donor base

Different risk tolerance

The correct authority level is not a sector standard. It is a property of **your shop** — which is exactly why it has to be written down rather than assumed.

Three controls must be visible before any level goes up. goes up.

72%

Approval

A person sees the recommendation and says yes before it acts. Built into the workflow, not the policy binder.

65%

Override

Staff can edit or reject without escalating, and the system records that they did.

40%

Evidence on display

Sources, recency, and uncertainty shown at the point of decision — not on request.

Where to begin

Good early candidates

- Internal summaries
- Data-quality review
- Research preparation
- Trip-plan options
- Bounded segmentation

Use with stronger controls

- Personalized stewardship
- Next-best actions
- Upgrade recommendations
- Donor-facing drafts
- Automated replies

Keep human-led by default

- Ambiguous donor intent
- Major ask timing
- Restricted gifts
- Vulnerable donor signals
- Privacy-sensitive outreach
- Complaint recovery

Caution: these are starting points, not blanket conclusions. The recommendation, the data, the controls, and your context still matter.

PART FOUR

The name on the line

Four ways this fails

The policy in a PDF

Written once, filed, never opened. Governance that nobody performs.

The pilot with no owner

Everyone is interested. Nobody is accountable. It never formally ends.

The demo that became a deployment deployment

A vendor trial quietly turns into production without a decision anyone made.

The staff member who agrees because because disagreeing is slower

Approval exists. Review does not. This is the failure your dashboard cannot see.

Cf. Green (2022) on oversight policies that legitimise flawed systems while providing a false sense of security.

ACTIVITY 2

Whose name is on it?

WRITE ONE

An AI use already live or proposed in your organization.

WRITE ONE NAME

The single person accountable if it goes wrong.

1 Stand if you wrote a name.

2 Stay standing if that person knows you wrote it.

3 Stay standing if that person could stop it by Friday.

A name without the the power to stop it it is not accountability.

The **moral crumple zone** : responsibility misattributed to a human operator who had limited control over the system.

Elish (2019), *Engaging Science, Technology, and Society* 5, 40–60.



Train staff to challenge both sources.

The study surfaced
weak human recommendations
as readily as weak AI ones.

So critical review cannot be triggered by
the **label**. It has to be triggered by the
stakes.

PART FIVE

Pilots: bounded learning, not brave adoption

Anatomy of a pilot that can fail safely

01

Bounded population

02

Evidence, recency, and
uncertainty on display

03

Approval and override

04

Escalation path and
opt-out rules

05

Named owner

06

Comparison or holdout
group

07

Predefined stop
conditions

78.3%

chose this, in substance, over
automating the portfolio (L3)

ACTIVITY 3 – THREE MINUTES, ONE SENTENCE

Write the stop condition.

Not a pilot plan. One sentence.

We will stop this pilot if

Most stop conditions are performance conditions. Yours needs a trust condition.

WHAT ROOMS USUALLY WRITE

If it doesn't save time. If adoption is low. If staff stop using it.

WHAT RESPONDENTS CARED ABOUT

Complaints. Opt-outs. Overrides.
Corrections. Relationship recovery.

If the stop condition contains no trust metric, the pilot cannot fail for the right reason. It can only fail for being slow.

Ninety days

1–30

Inventory and classify

Every current and proposed AI use, scored on data quality, donor contact, sensitivity, reversibility, and consequence.

Deliverable: a workflow register with a level, an owner, and prohibited actions per row.

31–60

Two pilots

One low-risk decision-support pilot and one supervised donor-facing pilot, with evidence display, approval, override, escalation, and opt-out.

Deliverable: a pilot protocol with a holdout and predefined stop conditions.

61–90

A review that can say no

Fundraising, operations, data, privacy, ethics, and leadership examine both performance and trust signals.

Deliverable: expand, revise, pause, or retire each use case, with documented lessons.

Measure trust alongside revenue

Operational

Time saved, coverage, throughput, response time, data corrections.

Recommendation quality

Acceptance, edits, overrides, rejections, source verification, staff confidence.

Governance

Approval latency, escalations, policy triggers, unauthorized actions, false positives.

Donor trust

Complaints, opt-outs, negative replies, personalization concerns, relationship recovery.

Fundraising outcomes

Meetings, qualification, stage movement, renewal, upgrade, revenue, lifetime value.

Do not optimize only for volume. A system that sends more messages can still reduce trust and increase corrections.

You already did this in the first six minutes.

You recognised a strong recommendation and withheld authority anyway, exactly what 340 professionals did across 1,360 judgments. That instinct is the asset. Governance is only the act of writing it down before someone has to.

30 days

One workflow register.

60 days

One pilot with a named owner and a stop condition.

90 days

One review with the authority to say no.

Kindsight (2026), *The Fundraising Judgment Benchmark: Where AI Helps, Where Human Judgment Must Lead*. Kindsight Research.

Thank you.

The Authority Gap KindCon
2026

KindCon²⁰²⁶

www.kindsight.io

Appendix A — all sixteen scenario results

ID	SCENARIO	N	HUMAN	AI	HOLM p	RESULT
F10	Trip planning, too many prospects	111	11.7%	88.3%	<.001	AI favored
P10	Grateful patient prospect	82	19.5%	80.5%	<.001	AI favored
G7	Automated stewardship cadence	78	28.2%	71.8%	.001	AI favored
P4	AI profile from public sources	82	34.1%	65.9%	.049	AI favored*
L2	Fundraiser performance insight	69	34.8%	65.2%	.123	No difference
G9	Donor requests anonymity	78	39.7%	60.3%	.355	No difference
F2	Major gift after strong event	111	50.5%	49.5%	1.000	No difference
P1	High capacity, weak relationship history	82	53.7%	46.3%	1.000	No difference
L8	Manager approval thresholds	69	59.4%	40.6%	.444	No difference
F6	Board member pressure on ask amount	111	61.3%	38.7%	.134	No difference
G3	Impact report with incomplete data	78	61.5%	38.5%	.268	No difference
L6	Donor complaint escalation	69	65.2%	34.8%	.123	No difference
P6	Predictive model upgrade	82	72.0%	28.0%	<.001	Human favored
L3	Unmanaged portfolio automation	69	78.3%	21.7%	<.001	Human favored
F3	Donor doubt before a planned ask	111	78.4%	21.6%	<.001	Human favored
G1	Restricted gift ambiguity	78	80.8%	19.2%	<.001	Human favored

Kindsight (2026), Appendix A. *P4 does not clear the Holm threshold in the sensitivity analysis.

Statistical methods

Scenario preference: two-sided exact binomial test against 50%, with Wilson 95% intervals.

Overall preference: mean of four judgments per respondent, respondent-cluster bootstrap interval.

Risk relationship: cross-tabulation and chi-square. F10 excluded — its response options differed.

Qualitative: 184 open responses reviewed for recurring themes; illustrative, not prevalence estimates.

Multiple comparisons: Holm family-wise correction across all 16 scenario tests.

Governance and subgroups: exploratory GEE models with scenario fixed effects and respondent clustering.

Controls: valid-response denominator of 262; up to three selections per respondent.

Excluded items: the first frontline risk question (no equal/neither options) and the first frontline reason question (multi-select where later items were single choice).

Kindsight (2026), Appendix B Technical notes.

Ten recommendations, in full

- 01** Stop debating AI at the organizational level. Define the exact decision or action.
- 02** Separate recommendation quality from authority. AI may generate the best option; a person remains responsible.
- 03** Make human approval and override visible. Build it into the workflow, not the policy.
- 04** Show evidence, recency, and uncertainty on profiles, scores, and next actions.
- 05** Treat prediction as a lead, not permission. A score triggers qualification and policy checks, not outreach.
- 06** Design fallbacks for weak data. Broader accurate language beats confident unsupported specificity.
- 07** Pilot consequential automation with a limited population, escalation rules, named ownership, comparison groups, and stop conditions.
- 08** Measure trust alongside revenue. Complaints, opt-outs, overrides, corrections, policy triggers, relationship recovery.
- 09** Define sensitive situations before deployment. Restricted gifts, anonymity, grateful patient work, vulnerable donors, high-value asks.
- 10** Train staff to challenge both sources. Critical review should not depend on the label.

Policy principle: AI may inform more decisions than it may execute. Authority narrows as relational stakes, ambiguity, privacy sensitivity, and irreversibility increase.

Kindsight (2026), §5.

References

Caffier, J., Stavrova, O., & Kleinberg, B. (2026). Prosocial persuasion at scale? *arXiv*. doi:10.48550/arXiv.2604.03202

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*.

Elish, M. C. (2019). Moral crumple zones. *Engaging Science, Technology, and Society* 5, 40–60.

Hackenburg, K., et al. (2026). AI systems out-persuade expert humans. *arXiv*. doi:10.48550/arXiv.2606.16475

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation. *Organizational Behavior and Human Decision Processes*.

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful. *Nature Human Behaviour* 8, 2293–2303. doi:10.1038/s41562-024-02024-1

EU Artificial Intelligence Act, Article 14 (human oversight of high-risk systems); GDPR Article 22.

Cagala, T., Glogowsky, U., Rincke, J., & Strittmatter, A. (2021). Optimal targeting in fundraising. *arXiv*. doi:10.48550/arXiv.2103.10251

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion. *Journal of Experimental Psychology: General*.

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review* 45.

Huang, G., & Wang, S. (2023). Is AI more persuasive than humans? *Journal of Communication* 73(6), 552–562. doi:10.1093/joc/jqad024

Novozhilova, E., et al. (2024). Will you donate money to a chatbot? *arXiv*. doi:10.48550/arXiv.2412.19976

Virtuous & Fundraising.AI (2026). *Nonprofit AI Adoption Report* (n=346). TechSoup nonprofit AI survey (n>1,300). Coastal Cloud (2026) nonprofit AI trends survey.

Kindsight (2026). *The Fundraising Judgment Benchmark: Where AI Helps, Where Human Judgment Must Lead*. Kindsight Research.