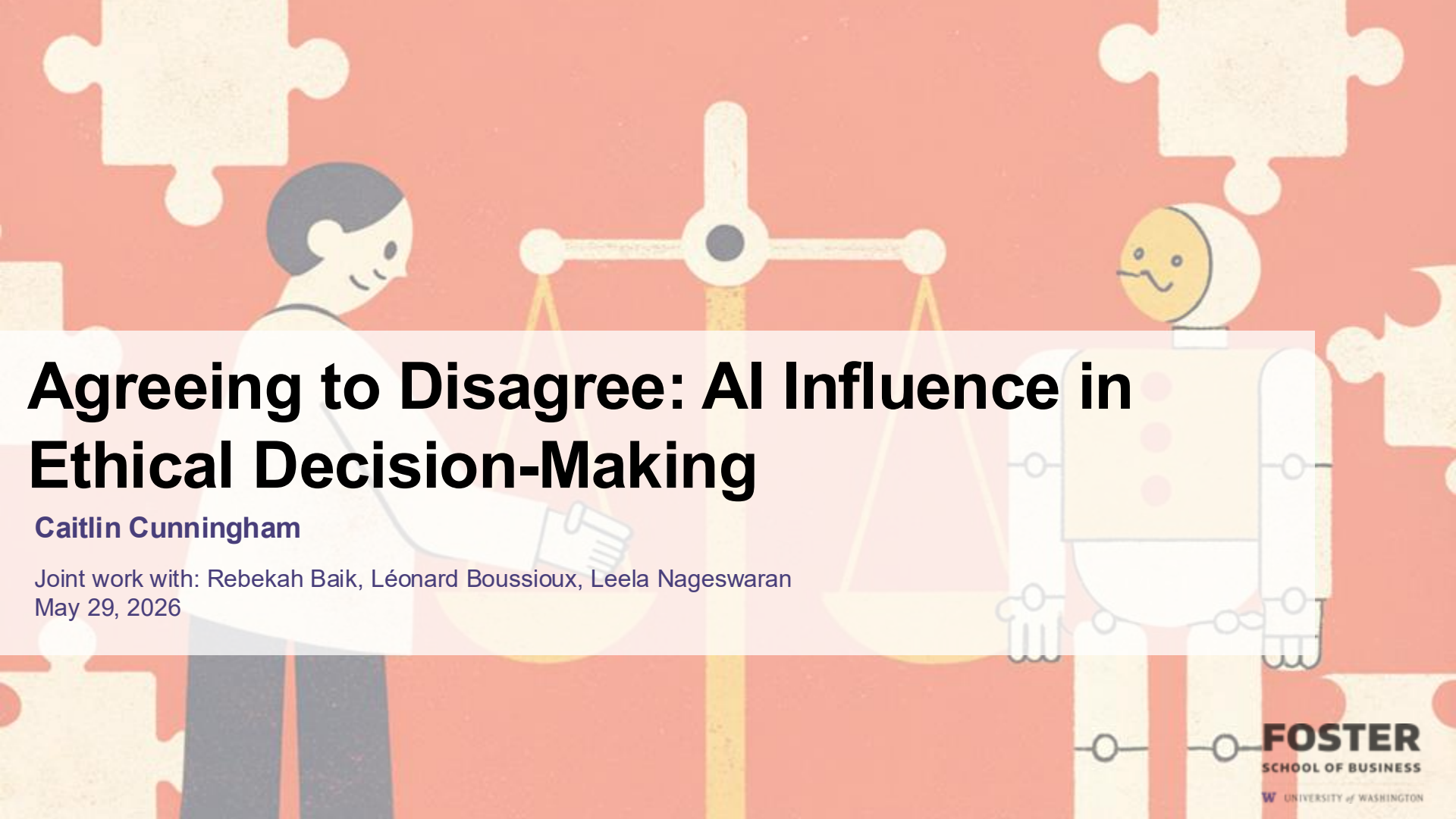


Agreeing to Disagree: AI Influence in Ethical Decision-Making

Caitlin Cunningham

Joint work with: Rebekah Baik, Léonard Boussioux, Leela Nageswaran

May 29, 2026

FOSTER
SCHOOL OF BUSINESS
UNIVERSITY of WASHINGTON

Ethical Contexts: Highly Subjective

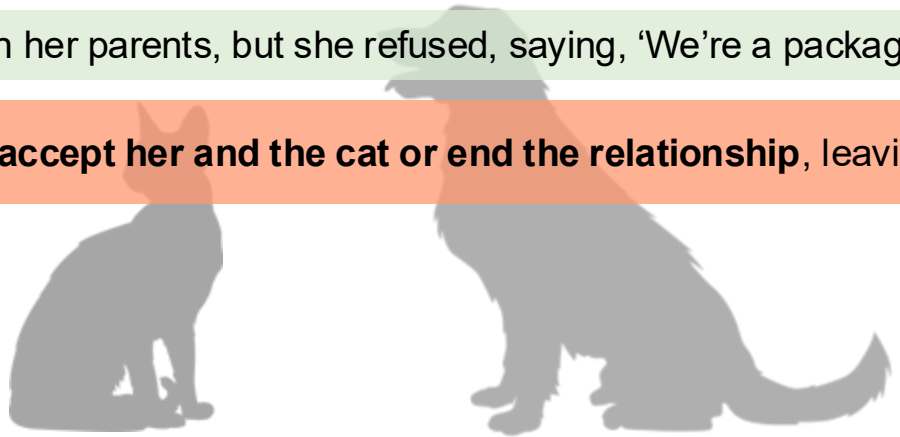
Am I in the wrong…

if I don't want my girlfriend to bring her cat when she moves in with me?

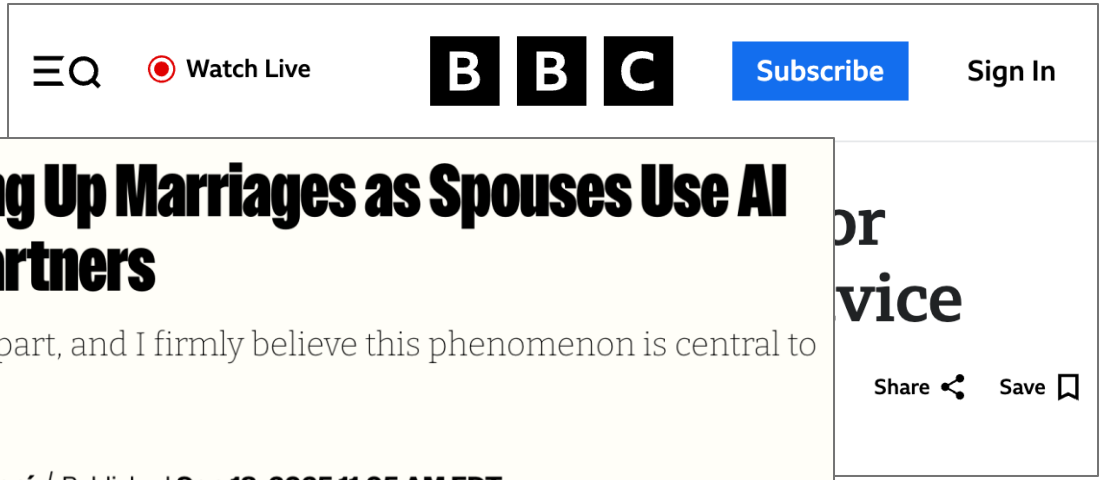
I love my girlfriend and want to marry her, but I hate her cat. It jumps everywhere, steps on me, and cries loudly all night.

I suggested the cat live with her parents, but she refused, saying, ‘We’re a package deal.’

She demanded I choose—**accept her and the cat or end the relationship**, leaving me stuck.



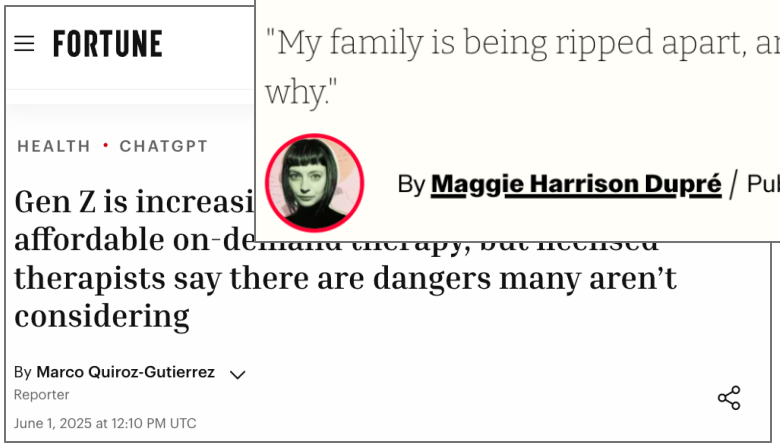
Getting Personal with AI



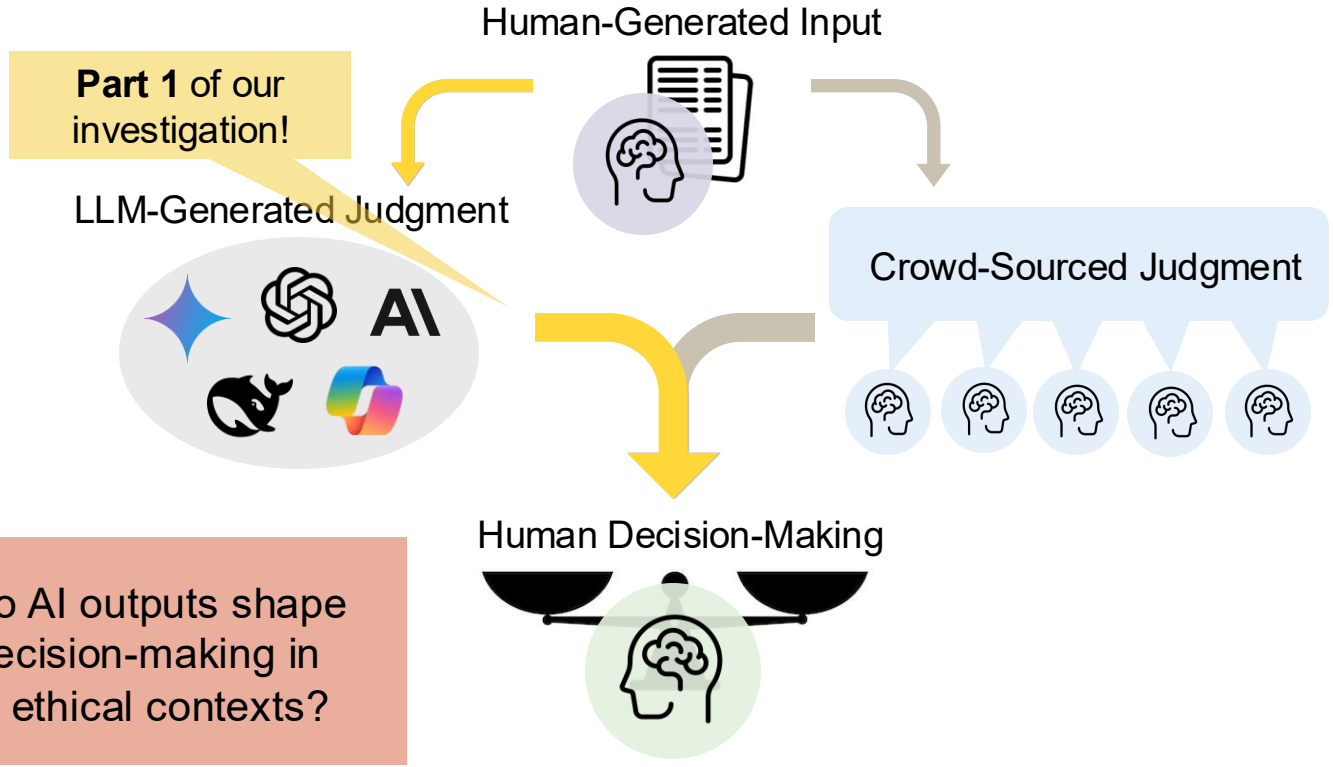
ChatGPT Is Blowing Up Marriages as Spouses Use AI to Attack Their Partners

"My family is being ripped apart, and I firmly believe this phenomenon is central to why."

By **Maggie Harrison Dupré** / Published **Sep 18, 2025 11:05 AM EDT**



Human-AI Interaction in Ethical Contexts



… how do AI outputs shape human decision-making in everyday ethical contexts?

Krügel, S., Ostermaier, A. & Uhl, M. ChatGPT's inconsistent moral advice influences users' judgment. Sci Rep 13, 4569 (2023).
Dillion, D., Mondal, D., Tandon, N. et al. AI language model rivals expert ethicist in perceived moral expertise. Sci Rep 15, 4084 (2025).

Dataset: Reddit Anecdotes from Scruples

Over 32,000 real-life, ethical anecdotes shared on Reddit (Lourie et al. 2021).

Community members vote on which party is in the wrong:

- The Author (the one telling the story)
- The Other (the other person in the story)
- Everybody
- or Nobody

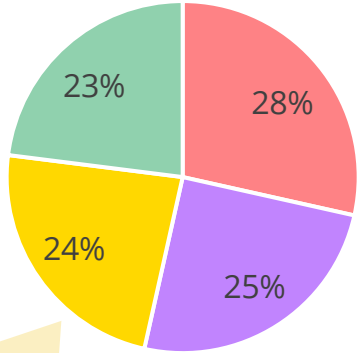
Votes are transformed into "*portions of blame*," giving us a probability distribution that sums to 1.

Lourie, Nicholas, Ronan Le Bras, and Yejin Choi. "Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 15. 2021.

Normalized Entropy as a Measure of Controversy

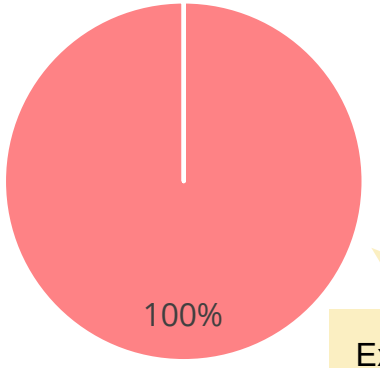
The entropy between "*portions of blame*" assigned by Reddit community members captures the **controversy** level of the anecdote.

Am I wrong for expecting my mom to want to support me through my first delivery and post partum?



Example of **high entropy (0.99)** → **highly controversial** (no clear consensus)

Am I wrong for punching someone who made me mad?

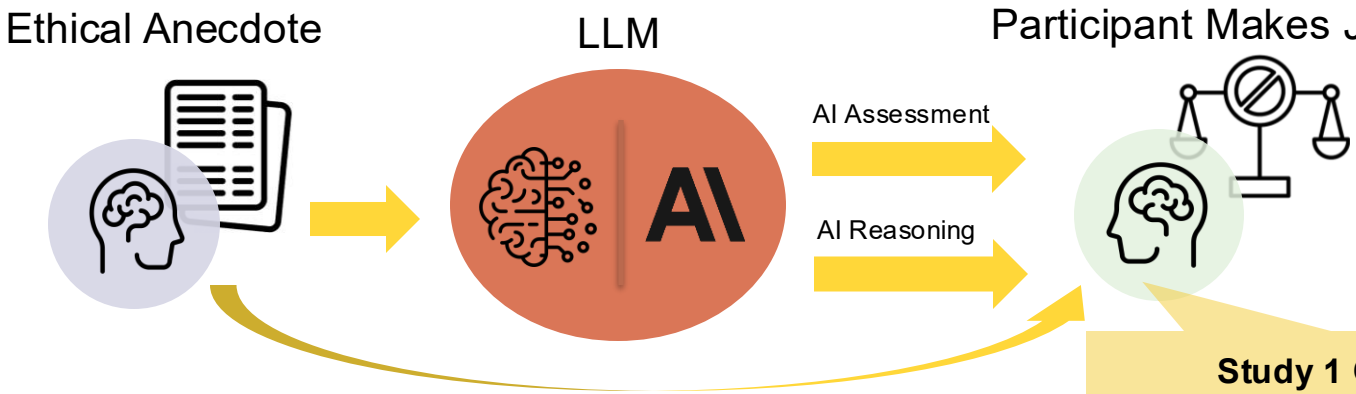


Example of **low entropy (0.0)** → **not controversial** (clear consensus amongst community members)

Author Other Everybody Nobody

Lab Experiment: Study 1

How do AI judgments (and written reasoning behind those judgments) shape human decision-making in everyday ethical contexts?



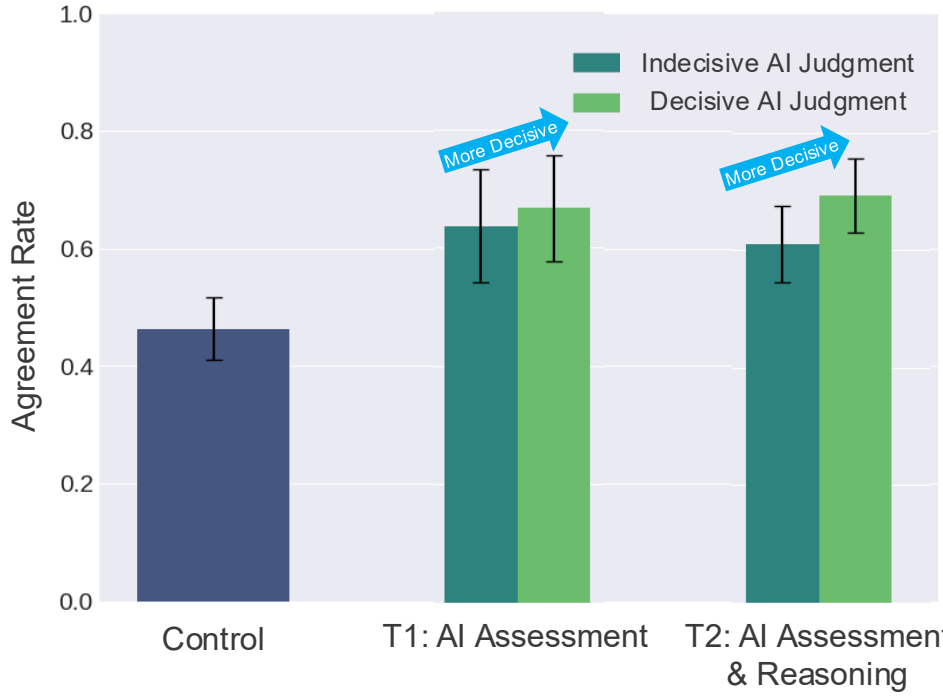
Study 1 Conditions:

- Control (no AI content)
- T1: AI Assessment
- T2: AI Assessment & Reasoning

95 student participants judge 10 anecdotes: 950 total judgments in Study 1

Exposure to AI Influences Agreement

AI Decisiveness is calculated as 1 minus the entropy between the LLM assigned portions of blame.



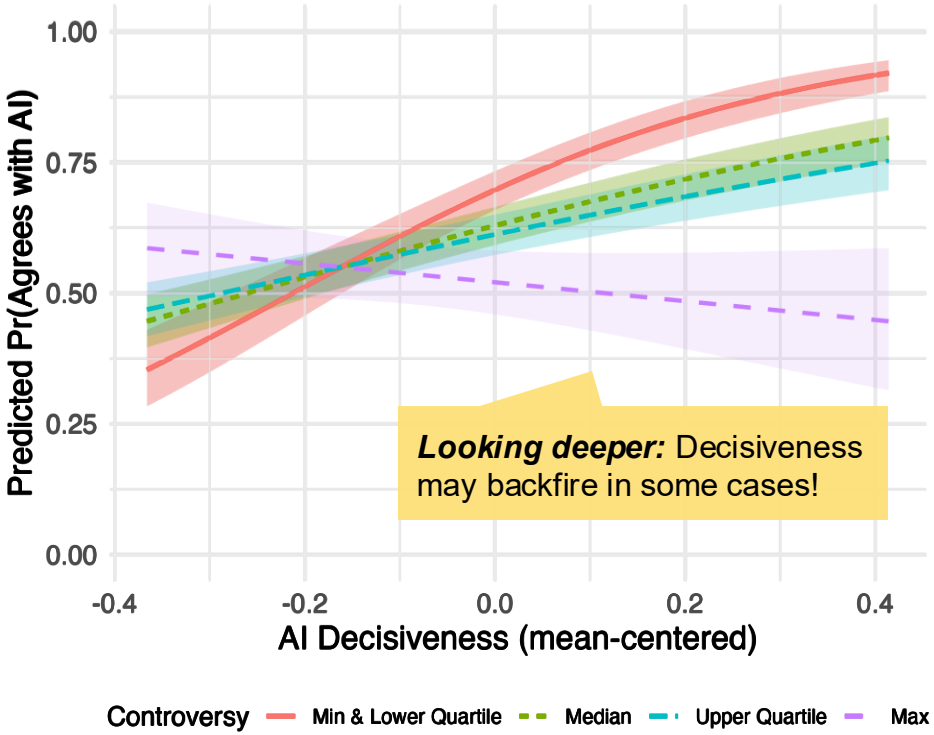
	DV: Agrees with AI Decision
Baseline	*Control*
T1: AI Assessment	0.827***
T2: AI Assessment & Reasoning	
Anecdote Controversy	
AI Decisiveness	2.306*** (0.471)
AI Decisiveness × Controversy	−4.688** (1.630)
Constant	−0.251· (0.149)
Anecdote Controls	✓
Observations	950

Note: Standard errors in parentheses.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$.

At first glance: If the AI is **decisive** (gives polarized judgments), humans agree more.

Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: evidence for a confidence heuristic. Journal of Behavioral Decision Making, 17, 39–57.
Tenney ER, MacCoun RJ, Spellman BA, Hastie R (2007) Calibration trumps confidence as a basis for witness credibility. Psychological Science 18(1):46–50.

Exposure to AI Influences Agreement

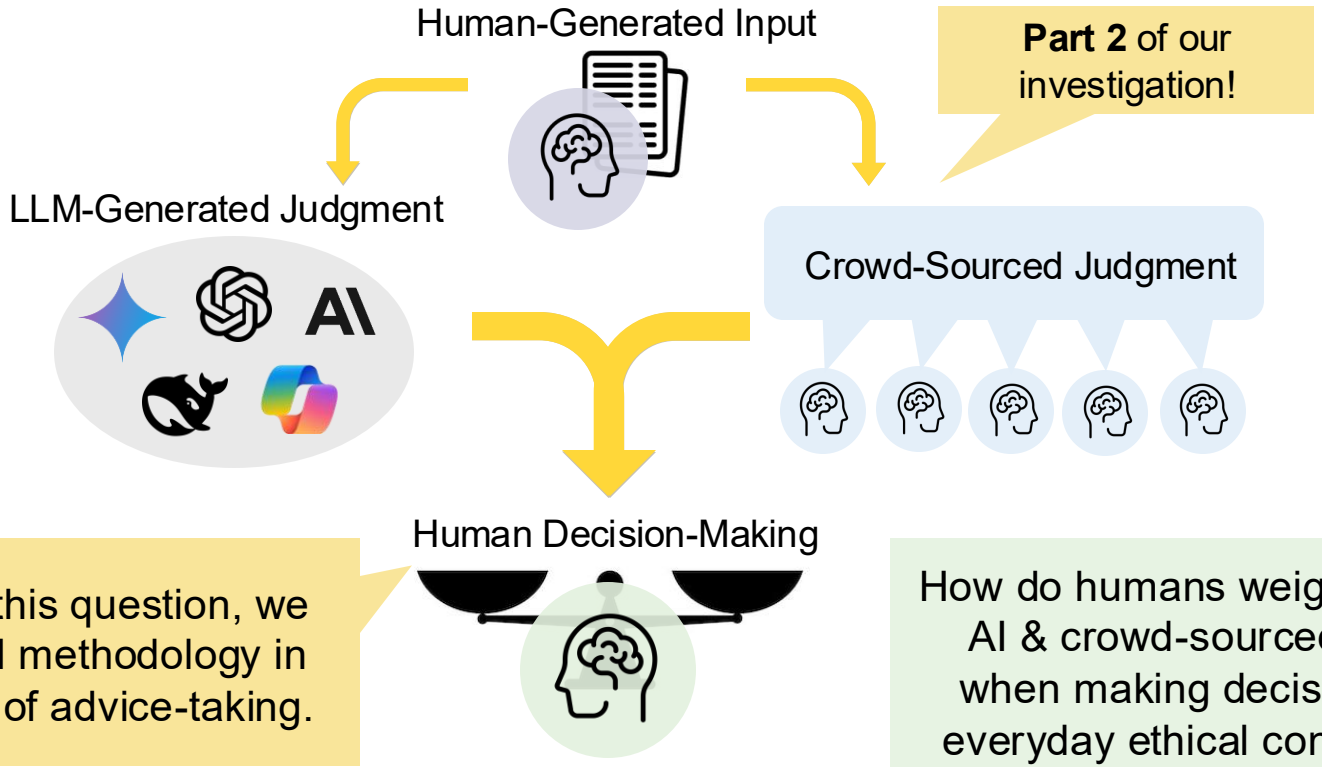


	DV: Agrees with AI Decision
Baseline	*Control*
T1: AI Assessment	0.827^{***}
	(0.177)
T2: AI Assessment & Reasoning	0.993^{***}
	(0.180)
Anecdote Controversy	-0.757^{**}
	(0.283)
AI Decisiveness	2.306^{***}
	(0.471)
AI Decisiveness × Controversy	-4.688^{**}
	(1.630)
Constant	$-0.251^{\cdot}$
	(0.149)
Anecdote Controls	✓
Observations	950

Note: Standard errors in parentheses.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$.

Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: evidence for a confidence heuristic. Journal of Behavioral Decision Making, 17, 39–57.
Tenney ER, MacCoun RJ, Spellman BA, Hastie R (2007) Calibration trumps confidence as a basis for witness credibility. Psychological Science 18(1):46–50.

Human-AI Interaction in Ethical Contexts



To answer this question, we use a novel methodology in the context of advice-taking.

How do humans weigh both AI & crowd-sourced input when making decisions in everyday ethical contexts?

Revealed Preference Model of Advice Weighting

We model participants' weights of each advice source as a preference that is revealed through their judgment regarding the party (*Author* or *Other*) that is most at fault.

Participants select the alternative with higher utility.

Crowd-Sourced Judgment

AI Judgment

Latent utility: $U_{ij}^{(k)} = W_{Crowd,i} \cdot p_{Crowd,j}^{(k)} + W_{AI,i} \cdot p_{AI,j}^{(k)} + \varepsilon_{ij}^{(k)}$

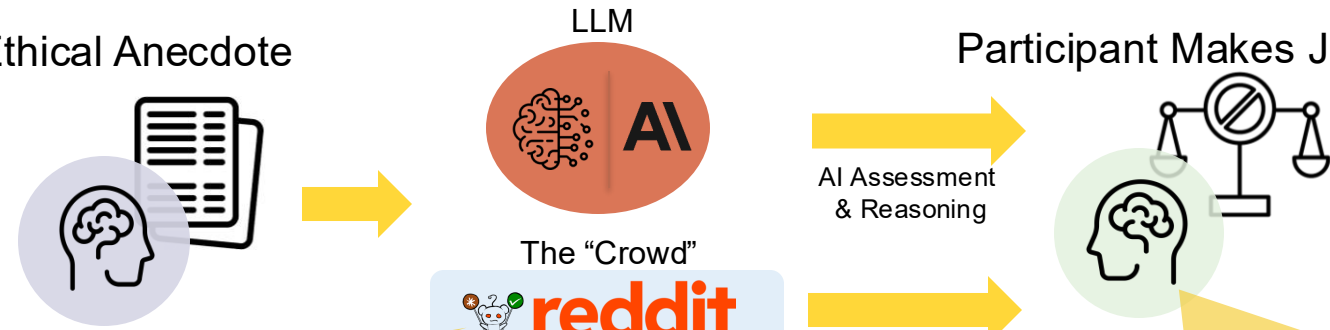
$k \in \{ Author, Other \}$

Weights vary randomly across participants.

$$Pr_{ij}(choose\ Author) = \frac{\exp(U_{ij}^{Author})}{\exp(U_{ij}^{Author}) + \exp(U_{ij}^{Other})}$$

Lab Experiment: Study 2

How do humans weigh both AI & crowd-sourced input when making decisions in ethical contexts? To answer this question, we use the following lab experiment.

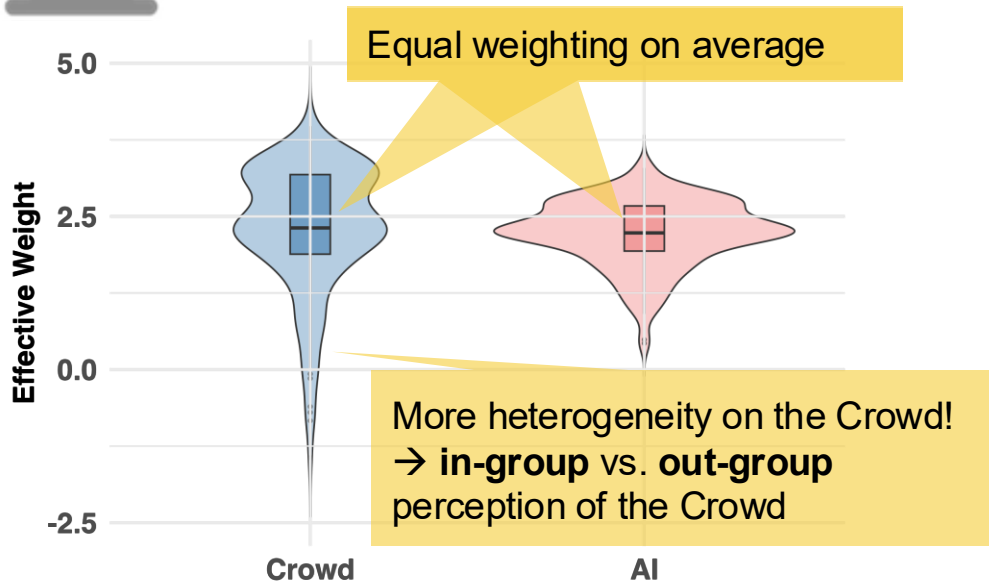


147 Prolific participants judge 10 anecdotes: 1,470 total judgments in Study 2

Revealed Advisor Preferences: Advice Weights

	All Treatment
Weight on Crowd ($\hat{\beta}_{\text{Crowd}}$)	2.310*** (0.267)
Weight on AI ($\hat{\beta}_{\text{AI}}$)	2.225*** (0.240)
SD of Random Slope (Crowd)	1.684*** (0.293)
SD of Random Slope (AI)	1.148*** (0.313)
Log Likelihood	−471.34
Observations	1,044

Note: Standard errors in parentheses.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$.



Terry, D. J., & Hogg, M. A. (1996). Group norms and the attitude-behavior relationship: A role for group identification. Personality and Social Psychology Bulletin, 22, 776–793.

Mechanism Analysis: Distinguishing Similarity in Reasonings

But what leads to differences in advisor preference across contexts?

$$U_{ij}^{(k)} = W_{Crowd,i} \cdot p_{Crowd,j}^{(k)} + W_{AI,i} \cdot p_{AI,j}^{(k)} + \ldots + \varepsilon_{ij}^{(k)}$$

$$W_{Crowd \times CS,i}\left(p_{Crowd,j}^{(k)} \cdot CS_j\right) + W_{AI \times CS,i}\left(p_{AI,j}^{(k)} \cdot CS_j\right)$$

AI Assessment of Fault Probabilities:

Author: 67% | **Other:** 36%

AI Reasoning:

The author bears 67% of the responsibility in this scenario…

… [content cropped for brevity] …

Crowd's Assessment of Fault Probabilities:

Author: 78% | **Other:** 22%

Crowd's Reasoning:

The majority view holds that the author was wrong for revealing …

… [content cropped for brevity] …

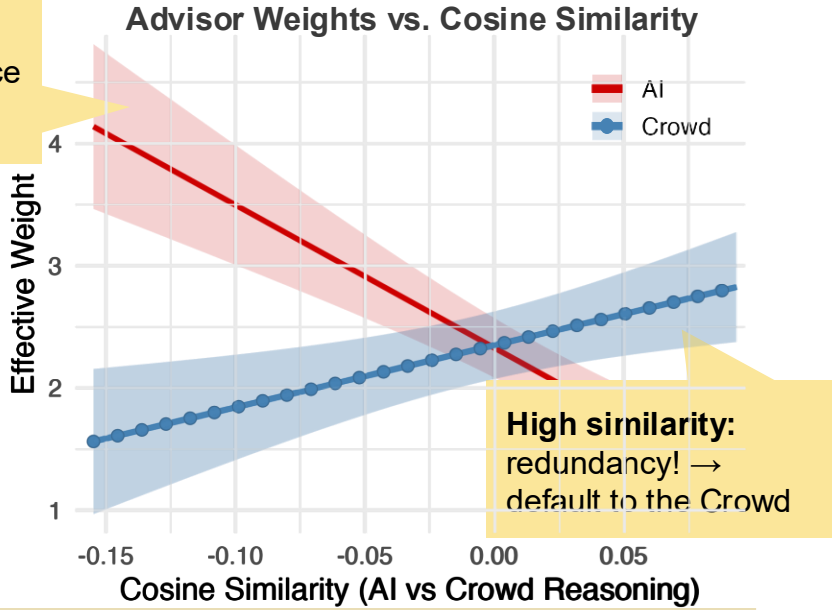
High divergence:
AI offers unique advice → higher AI weight

…]

Cosine Similarity

[0.0249015465, 0.02637814916, …]

text embedding of Crowd Reasoning

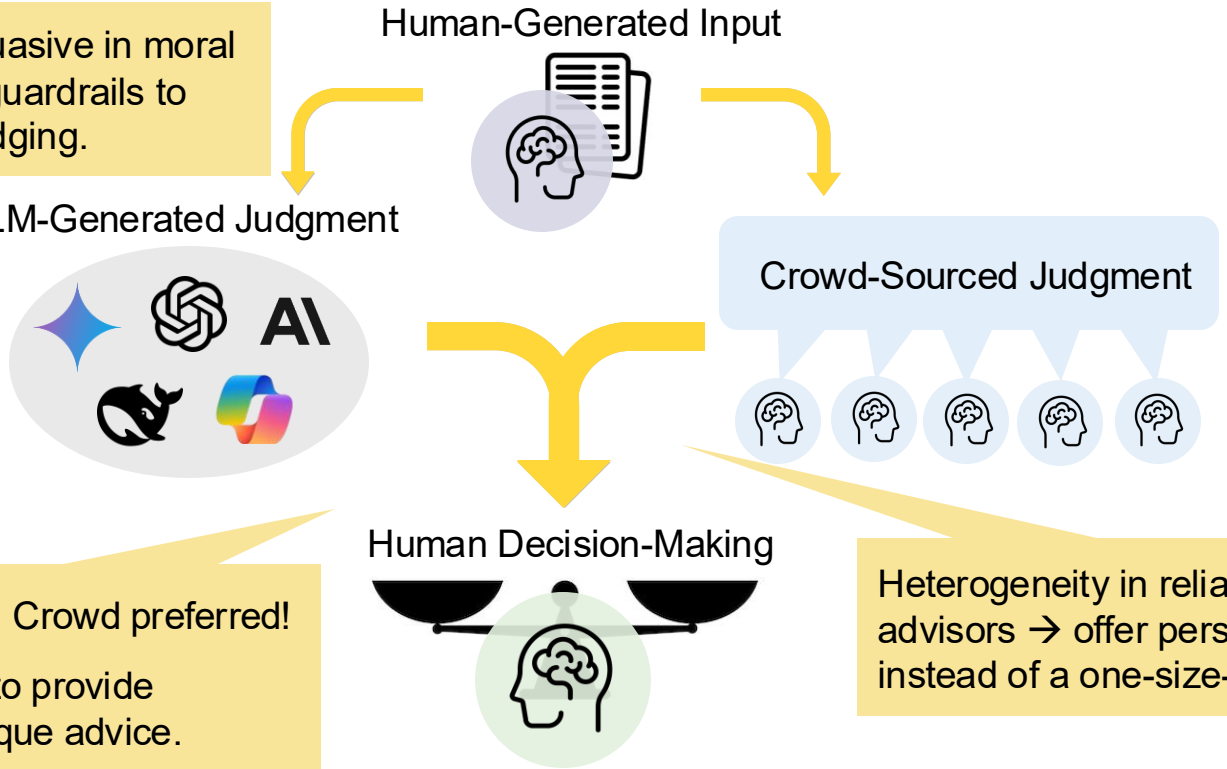


Note: Cosine Similarity is mean-centered for interpretability.

Surowiecki, James. The Wisdom of Crowds. Anchor Books, 2005.
Yaniv I, Milyavsky M (2007) Using advice from multiple sources to revise and improve judgments. Organizational Behavior and Human Decision Processes 103(1):104–120.

Managerial Implications

AI outputs are persuasive in moral scenarios → need guardrails to avoid one-sided nudging.



Similar advice → Crowd preferred!

Use scalable AI to provide meaningfully unique advice.

Heterogeneity in reliance on the advisors → offer personalization instead of a one-size-fits-all.

You be the judge! Who is in the wrong?

Is the boyfriend in the wrong

if he doesn't want his girlfriend to bring her cat when she moves in?

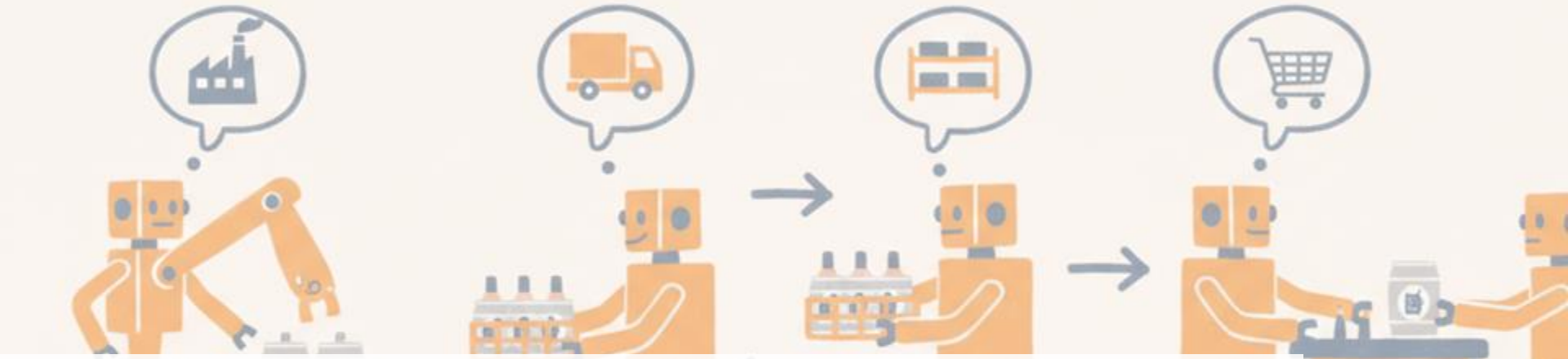
"The boyfriend is in the wrong."

"The girlfriend is in the wrong."

Thank you!

csc27@uw.edu

FOSTER
SCHOOL OF BUSINESS
UNIVERSITY of WASHINGTON



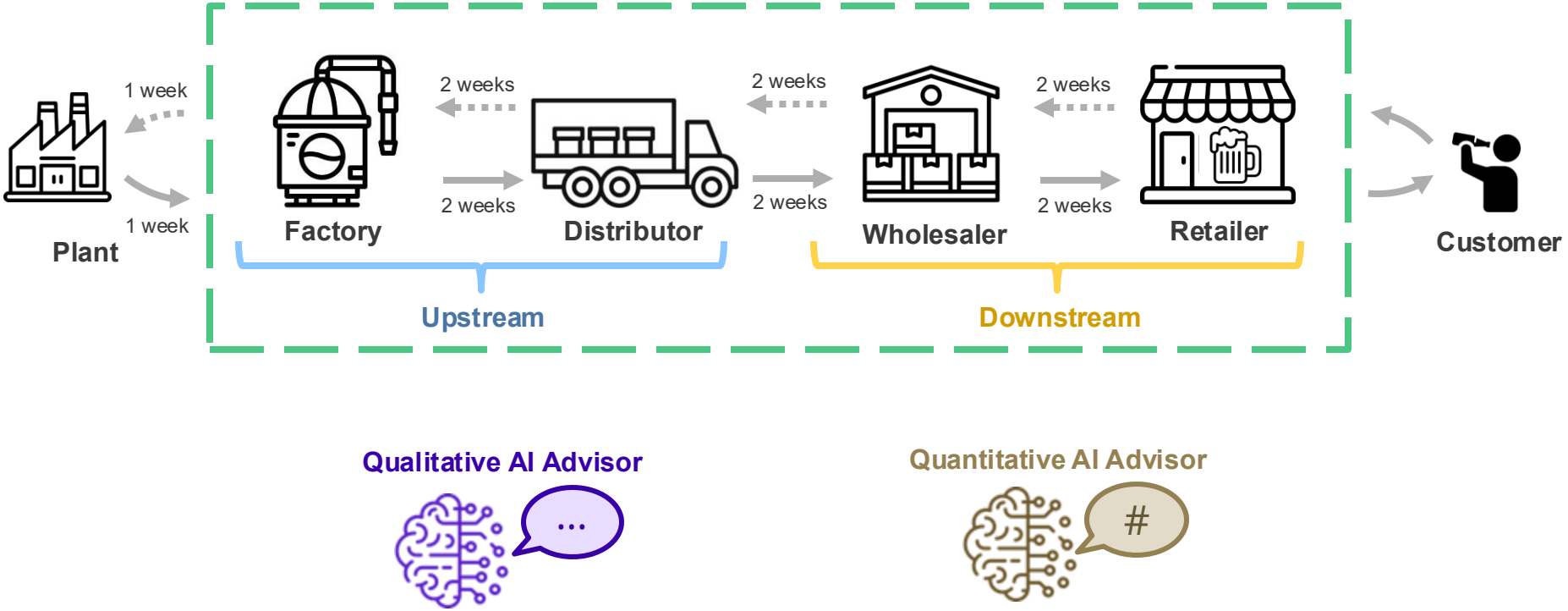
Decision Making on Draft: Human-AI Collaboration in the Beer Game

Caitlin Cunningham

Joint work with: Rebekah Baik, Leela Nageswaran, Léonard Boussioux

FOSTER
SCHOOL OF BUSINESS
UNIVERSITY of WASHINGTON

Human-AI Collaboration in Supply Chains



Information delays denoted by dotted left arrows. Shipping delays denoted by solid right arrows

AI & Ethics in Business

The Guardian

‘Sliding into an abyss’: experts warn over rising use of AI for mental health support

Rachel Hall

Sat 30 Aug 2025 06.00 EDT

Therapists say they are seeing negative impacts of people increasingly turning to AI chatbots for help

Newsrooms Are
Considerations

By Meg Little Reilly,
Community News at

Published Apr 22, 2024,

WHO, IT
on AI us

11 July 2025 | Ne

WIR

CAROLINE HASKINS

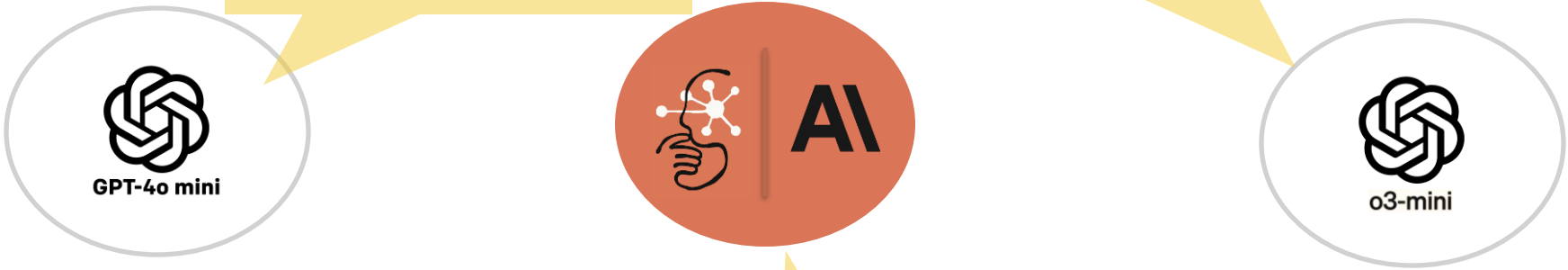
People Who Say They’re Experiencing AI Psychosis Beg the FTC for Help

The Federal Trade Commission received 200 complaints mentioning ChatGPT between November 2022 and August 2025. Several attributed delusions, paranoia, and spiritual crises to the chatbot.

Common Models on a Scale of ‘Decisiveness’

Non-reasoning models like 4o-mini show the most “diplomacy” or “indecisiveness” in judgments

Contrarily, o3-mini is very decisive.



We choose Claude Sonnet 3.7 for our lab experiments because it balances diplomacy and decisiveness (and offers access to the internal chain-of-thought!).

Study 1: Appendix Slides

Study 1 Conditions

Treatment 1: participant weighs the AI's assessment when judging

AI Decision: Author

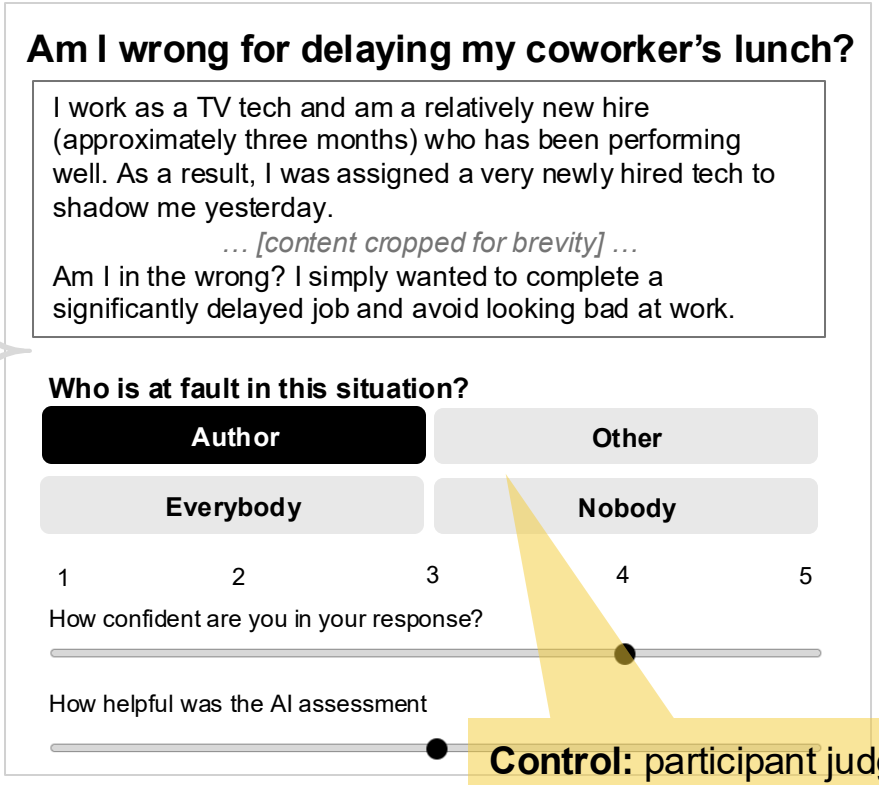
AI Assessment of Fault Probabilities:

Author: 70%	**Other:** 30%
Everybody: 0%	**Nobody:** 0%

AI Reasoning:

The author bears the majority of blame in this situation. While having productivity concerns is valid, denying a reasonable lunch break around noon (a standard time for meals) and asking someone to wait an additional 2 hours (until approximately 2PM) is inconsiderate of basic human needs. The author was prioritizing job performance over the shadow's wellbeing. The previous offer at 9AM isn't equivalent, as that's breakfast time, not lunch.

Treatment 2: participant weighs both the AI's assessment and reasoning



Control: participant judges without any advice

Revealed Preference Model of Advice Weighting

We model participants' weights of each advice source as a preference that is revealed through their judgment regarding the party that is most at fault.

Participants select the alternative with higher utility.

Human Judgment

AI Judgment

Latent utility: $U_{ij}^{(k)} = W_{Human,i} \cdot p_{Human,j}^{(k)} + W_{AI,i} \cdot p_{AI,j}^{(k)} + \varepsilon_{ij}^{(k)}$

Study 1: $k \in \{Author, Other, Everybody, Nobody\}$

Study 2: $k \in \{Author, Other\}$

$$Pr_{ij}(choose\ Author) = \frac{\exp(U_{ij}^{Author})}{\sum_k \exp(U_{ij}^{Author})}$$

Weights vary randomly across participants.

Study 1: Aggregate Results

AI receives higher weight on average!

	(1)	(2)
Weight on Human ($\hat{\beta}_{\text{control}}$)	1.563***	1.599***
	(0.236)	(0.238)
Weight on AI ($\hat{\beta}_{\text{AI}}$)	2.352***	—
	(0.204)	
Weight on AI with No Reasoning ($\hat{\beta}_{\text{AI,T1}}$)	—	2.202***
		(0.253)
Weight on AI with Reasoning ($\hat{\beta}_{\text{AI,T2}}$)	—	2.503***
		(0.298)
SD of Random Slope (Human)	0.005	0.002
	(0.580)	(0.594)
SD of Random Slope (AI)	1.068***	—
	(0.259)	
SD (AI with No Reasoning)	—	0.831*
		(0.403)
SD (AI with Reasoning)	—	1.354***
		(0.366)
Log Likelihood	−567.50	−567.14
Observations	605	605

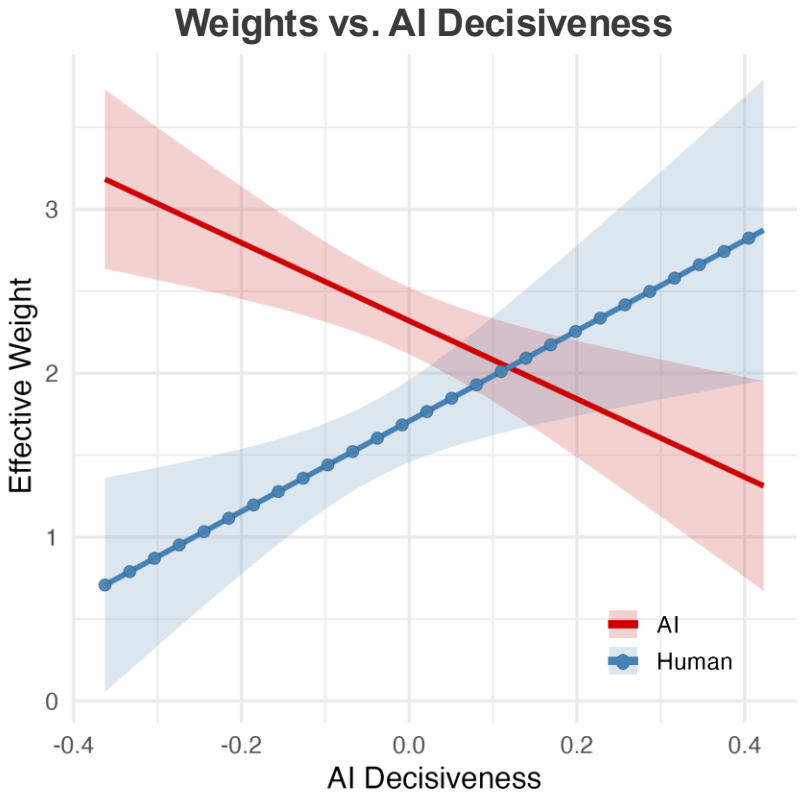
Note: Standard errors in parentheses. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Explanations increase weighting on average.

Study 1: Decisiveness Subsets

	Indecisive AI	Decisive AI
Weight on Human ($\hat{\beta}_{\text{human}}$)	1.370***	2.498***
	(0.272)	(0.517)
Weight on AI ($\hat{\beta}_{\text{AI}}$)	2.511***	1.902***
	(0.266)	(0.296)
SD of Random Slope (Human)	0.006	0.005
	(0.849)	(1.334)
SD of Random Slope (AI)	1.214**	0.930*
	(0.402)	(0.432)
Observations	350	255

Note: Standard errors in parentheses.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$.

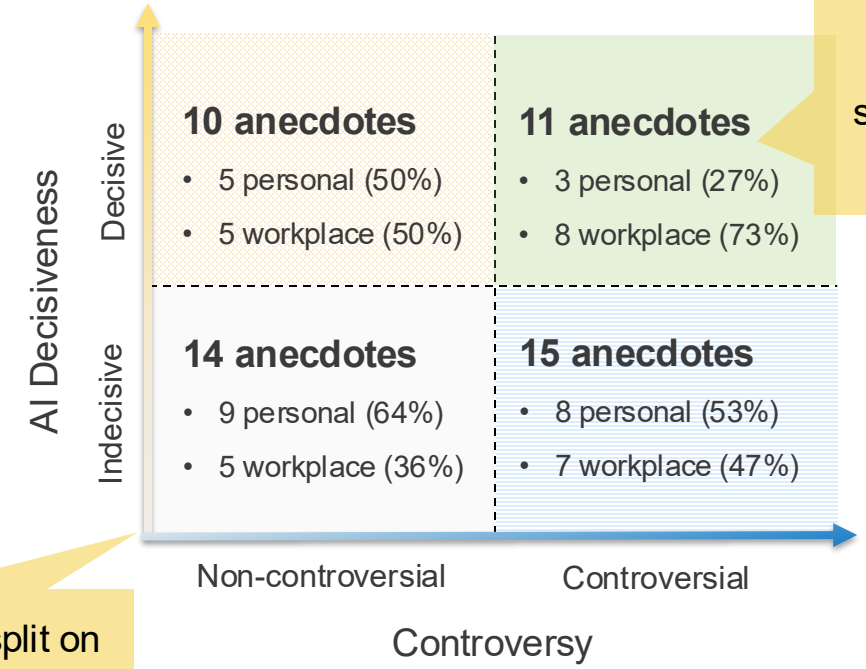


Study 1: Controversy Subsets

	Non-Controversial	Controversial
Weight on Human ($\hat{\beta}_{\text{human}}$)	1.479^{***}	1.596^{***}
	(0.345)	(0.322)
Weight on AI ($\hat{\beta}_{\text{AI}}$)	2.398^{***}	2.216^{***}
	(0.257)	(0.260)
SD of Random Slope (Human)	0.021	0.005
	(2.311)	(0.809)
SD of Random Slope (AI)	0.856.	0.819.
	(0.483)	(0.451)
Log Likelihood	−273.29	−297.29
Observations	296	309

Note: Standard errors in parentheses. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$.

Study 1: Anecdote Categories by Decisiveness & Controversy



The AI is weighted **least** at the cross-section of Decisiveness and Controversy. This cross shows the highest proportion of **workplace** anecdotes!

Both cross-sections split on median Decisiveness and median Controversy.

Study 1 Additional Outcomes: Perceived Confidence & Helpfulness

AI output does increase confidence…

… but the inclusion of reasoning does not add significant additional benefit beyond the AI assessment.

However, reasoning does significantly contribute to perceived helpfulness of advice.

	(1) **DV: Confidence**	(2) **DV: Confidence**	(3) **DV: Helpfulness**
Baseline	*Control*	*T1: AI Assessment*	*T1: AI Assessment*
T: All Treatment (T1 + T2)	0.096*		
	(0.048)		
T2: AI Assessment & Reasoning		−0.098	0.734***
		(0.067)	(0.105)
Anecdote Controversy	0.017	0.028	0.001
	(0.080)	(0.097)	(0.152)
AI Decisiveness	0.444***	0.593***	0.384
	(0.132)	(0.163)	(0.255)
Constant	3.779***	3.863***	2.473***
	(0.111)	(0.126)	(0.195)
Controls:			
Anecdote Text Length	✓	✓	✓
Anecdote Category	✓	✓	✓
Anecdote Sentiment	✓	✓	✓
Model	Mixed-effects OLS	Mixed-effects OLS	Mixed-effects OLS
Observations	950	605	605
Log Likelihood	−1,036.386	−650.152	−916.385
AIC	2,090.77	1,318.31	1,850.77
BIC	2,134.48	1,357.95	1,890.42

Note: Standard errors in parentheses. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$.

Study 2: Appendix Slides

Study 2 Conditions: Dual Advisor Presentation

Am I wrong for telling my manager that a colleague is going to hand in their notice?

Okay so me and my colleague are pretty close. One late shift, she tells me in confidence that she plans to leave. Cool.

… [content cropped for brevity] …

It got me thinking though. Should I have lied to my manager? Am I in the wrong?

Who is at fault in this situation?

AI Assessment of Fault Probabilities:

Author: 67% | **Other:** 36%

AI Reasoning:
The author bears 67% of the responsibility in this scenario…

… [content cropped for brevity] …

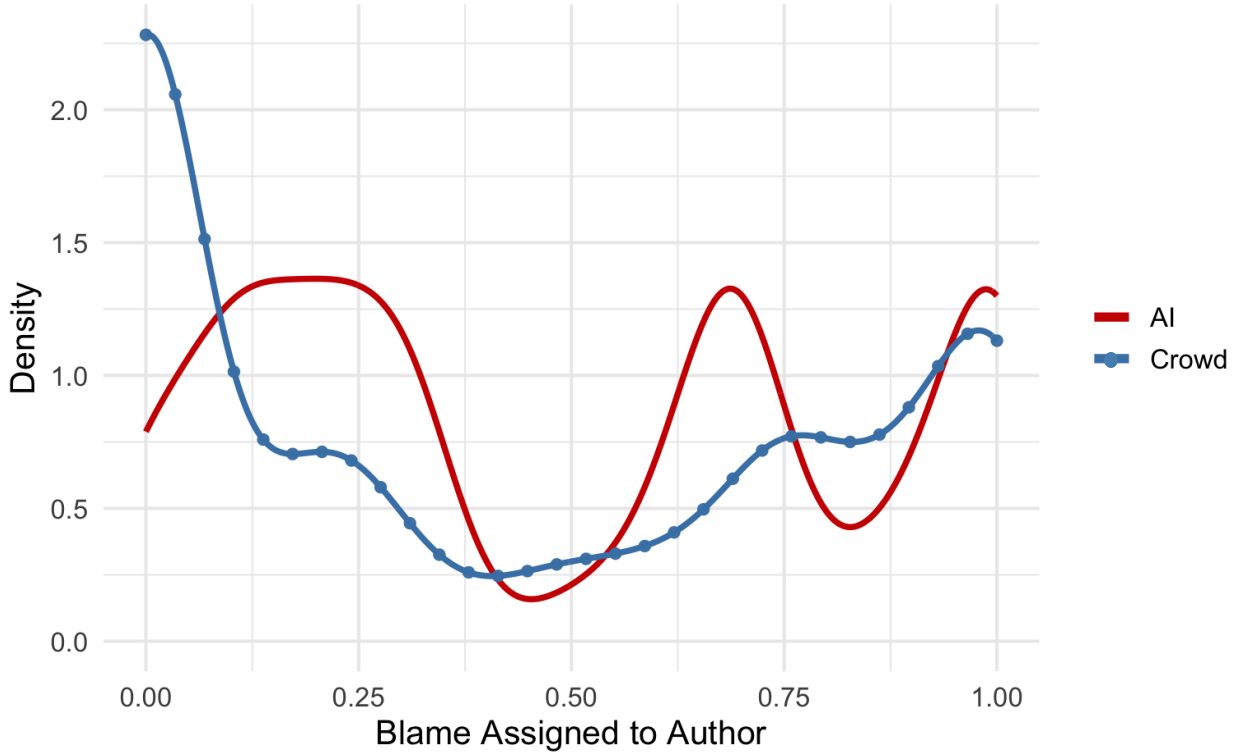
Crowd's Assessment of Fault Probabilities:

Author: 78% | **Other:** 22%

Crowd's Reasoning:
The majority view holds that the author was wrong for revealing a colleague's confidential plans…

… [content cropped for brevity] …

Study 2: Author Blame Density Plot



Study 2: Influential AI? The Double-edged Sword of Decisiveness

$$U_{ij}^{(k)} = W_{Human,i} \cdot p_{Human,j}^{(k)} + W_{AI,i} \cdot p_{AI,j}^{(k)} + \cdots + \varepsilon_{ij}^{(k)}$$

$$W_{Human \times AI\_D,i}\left(p_{Human,j}^{(k)} \cdot AI\_Decisiveness_j\right) + W_{AI \times AI\_D,i}\left(p_{AI,j}^{(k)} \cdot AI\_Decisiveness_j\right)$$

	All Treatment
Weight on Human ($\hat{\beta}_{human}$)	1.563***
	(0.236)
Weight on AI ($\hat{\beta}_{AI}$)	2.352***
	(0.204)
SD of Random Slope (Human)	0.005
	(0.580)
SD of Random Slope (AI)	1.068***
	(0.259)
Observations	605

Note: Standard errors in parentheses.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$.

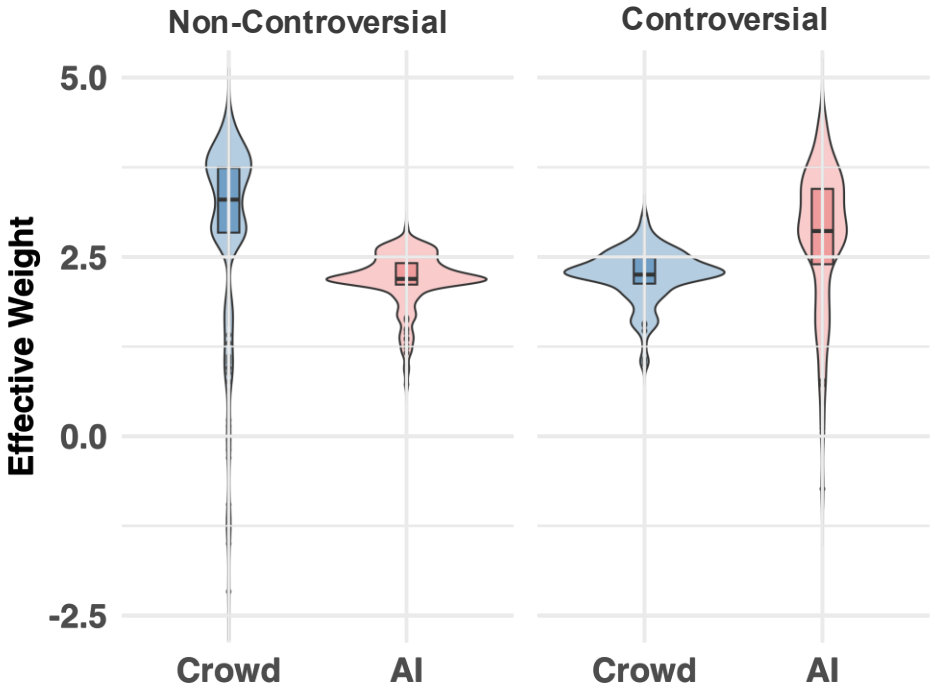
Indecisive AI	Decisive AI
1.370***	2.498***
(0.272)	(0.517)
2.511***	1.902***
(0.266)	(0.296)
0.006	0.005
(0.849)	(1.334)
1.214**	0.930*
(0.402)	(0.432)
350	255

+ weight

Study 2: Advisor Preferences Across Controversy

	Non-Controversial	Controversial
Weight on Crowd ($\hat{\beta}_{\text{crowd}}$)	2.845***	2.245***
	(0.510)	(0.367)
Weight on AI ($\hat{\beta}_{\text{AI}}$)	2.173***	2.758***
	(0.407)	(0.432)
SD of Random Slope (Crowd)	2.390***	1.089·
	(0.551)	(0.556)
SD of Random Slope (AI)	1.176·	1.981***
	(0.668)	(0.494)
Observations	502	542

Note: Standard errors in parentheses. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$.

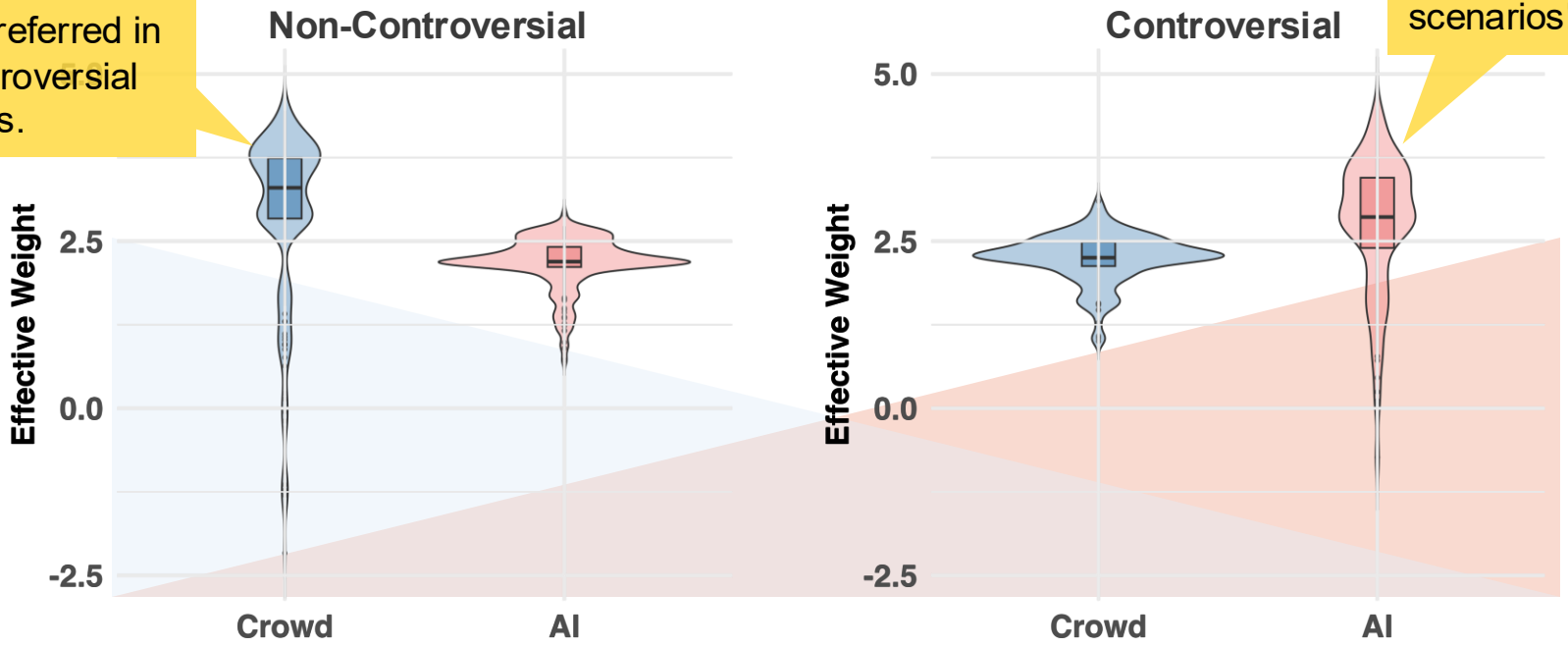


Advisor Preferences in the Face of Controversy

But how does advisor preference change across anecdote dimensions, like **controversy**?

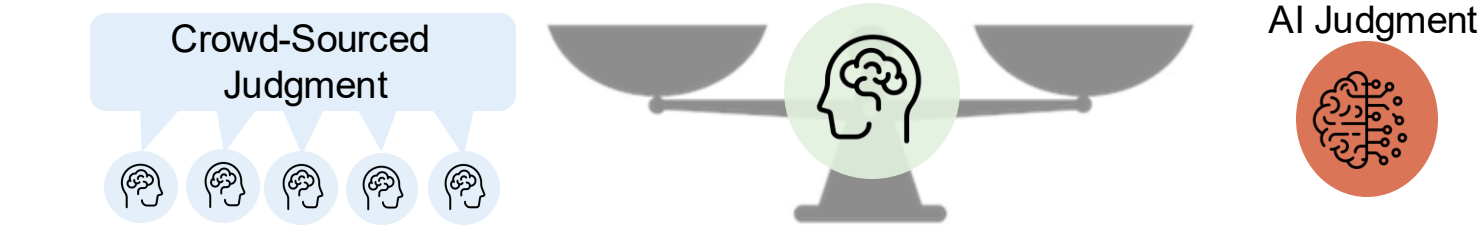
Crowd preferred in non-controversial scenarios.

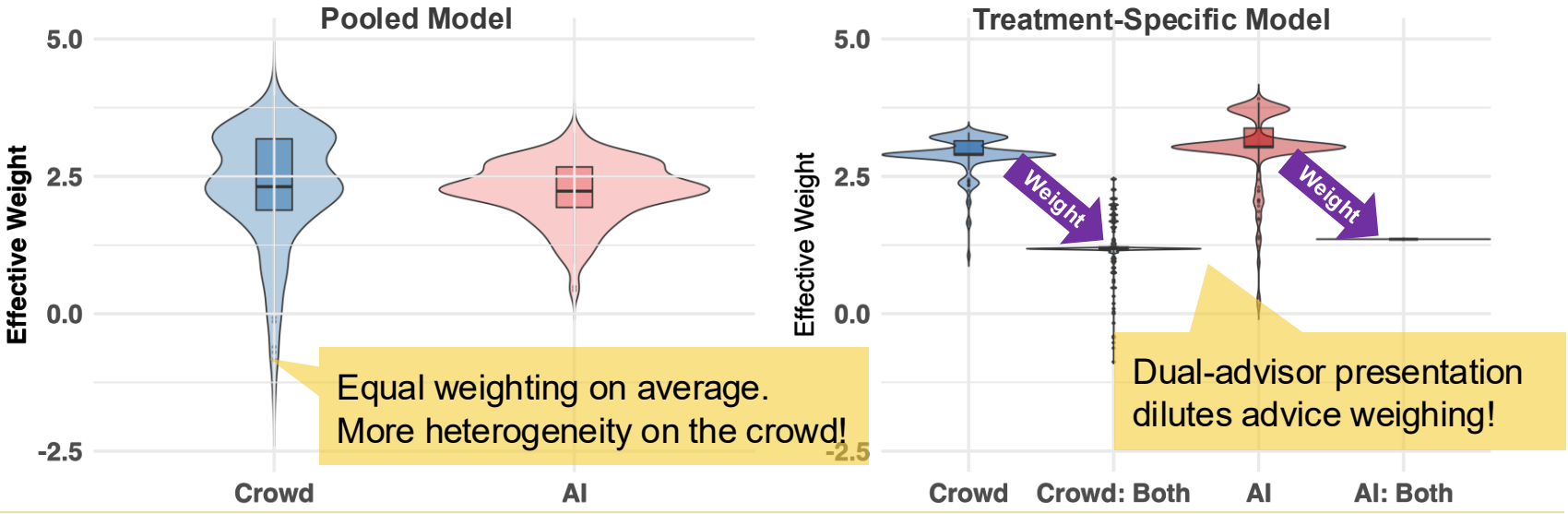
AI preferred in controversial scenarios!



Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes 151:90–103
Glikson E, Woolley AW (2020) Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals 14(2):627–660.

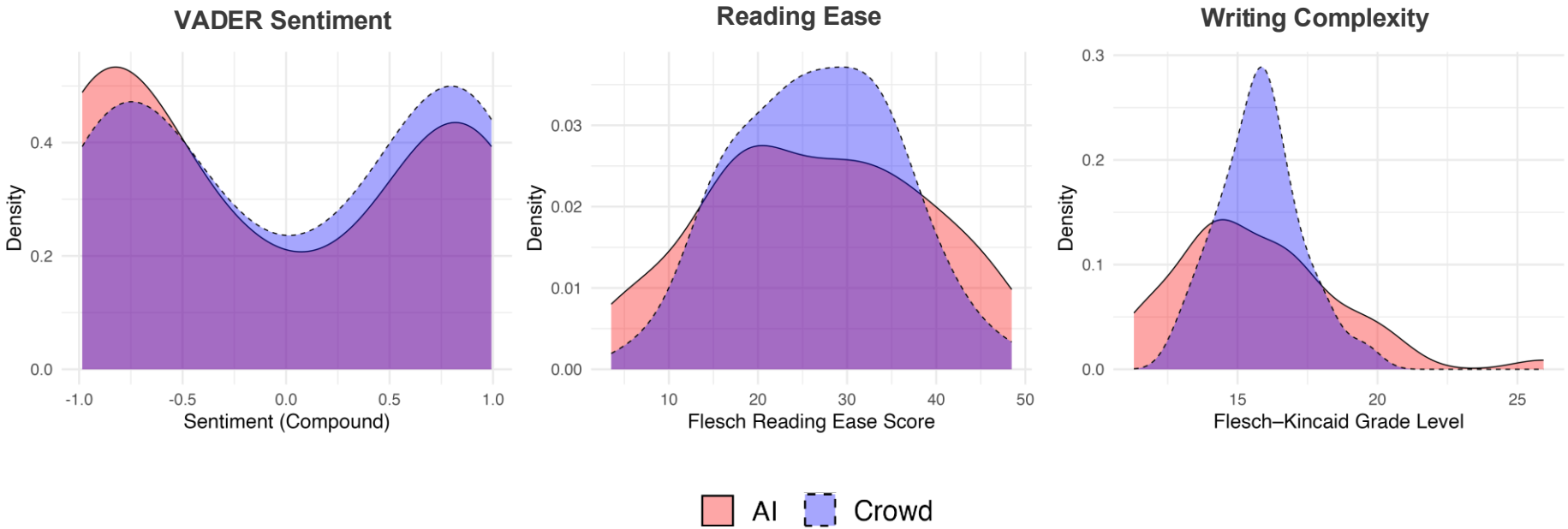
Revealed Advisor Preferences: Advice Weights



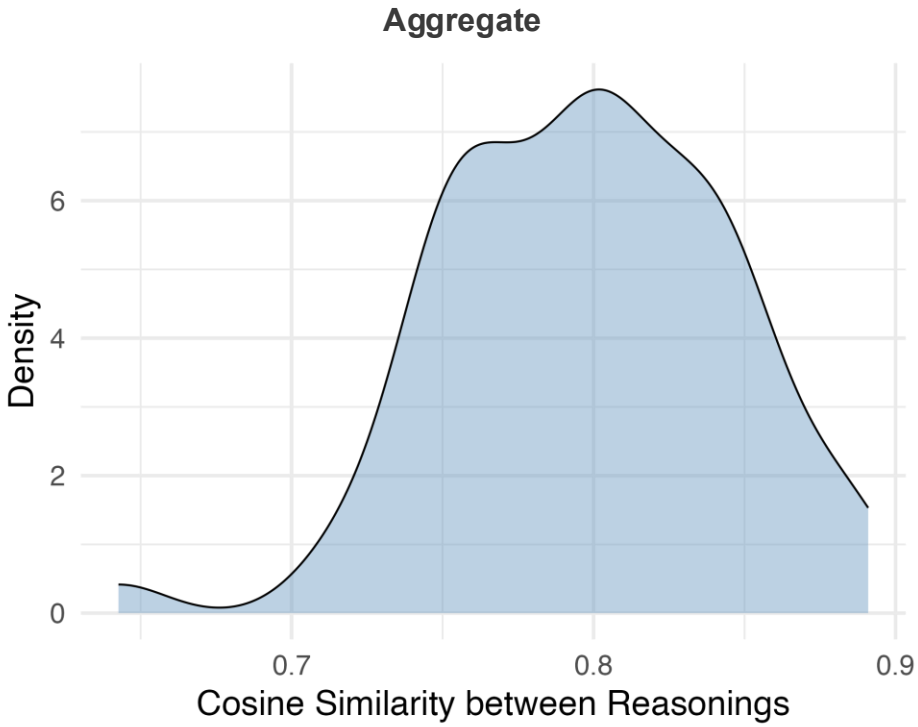


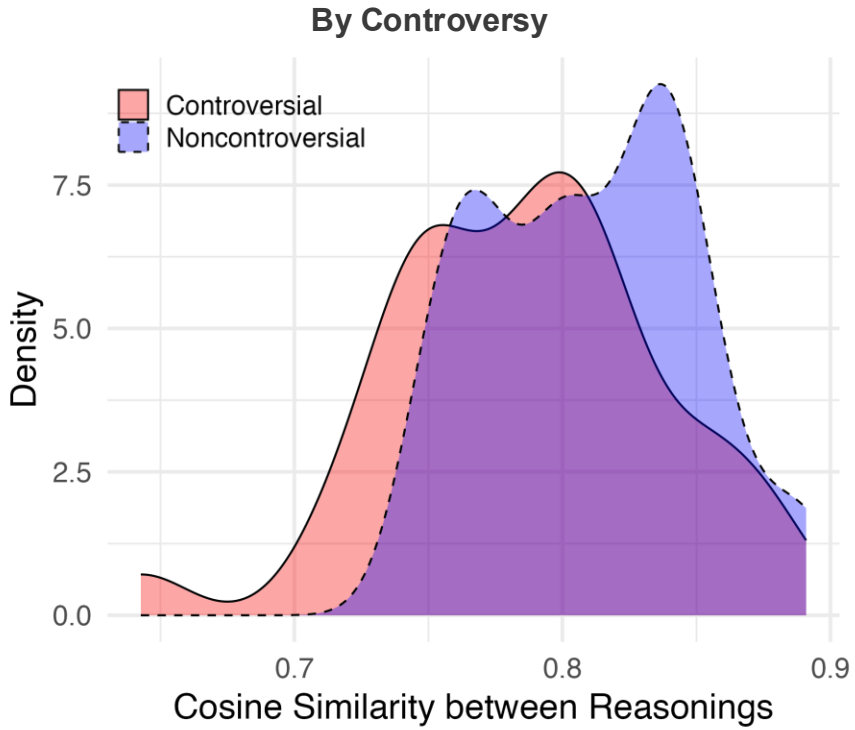
Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes 151:90–103
Terry, D. J., & Hogg, M. A. (1996). Group norms and the attitude-behavior relationship: A role for group identification. Personality and Social Psychology Bulletin, 22, 776–793.

Study 2: Density of NLP Metrics by Advisor

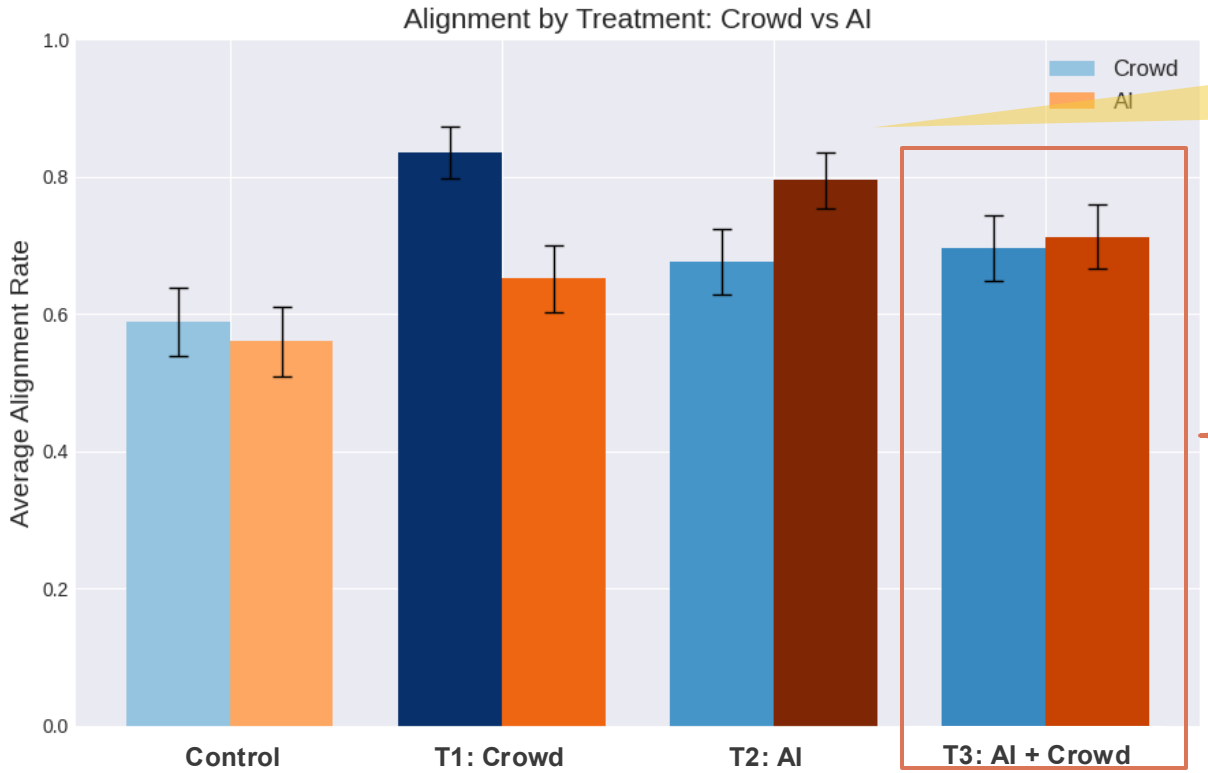


Study 2: Density of Cosine Similarity of Reasonings





Study 2 Logit Results: Alignment with Crowd & AI

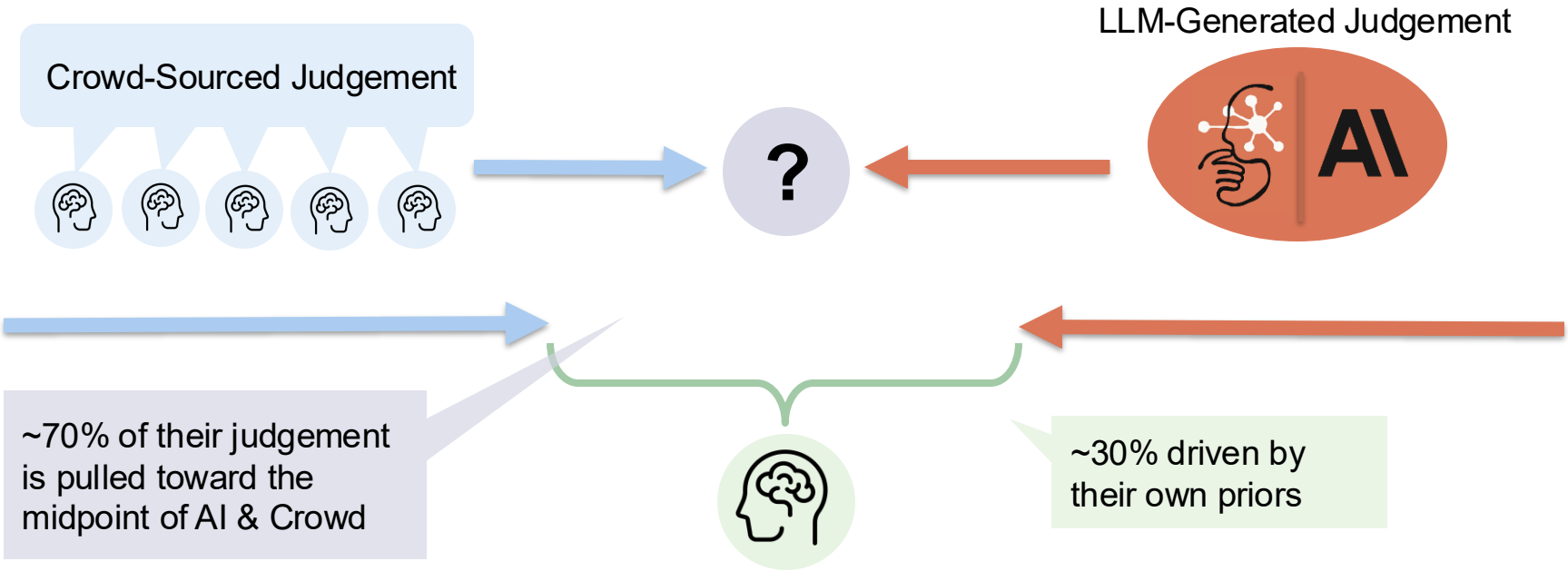


Treatments are strong predictors of alignment with the Crowd (T1 or T3) or the AI (T2 or T3).

Alignment is close when exposed to both sources!

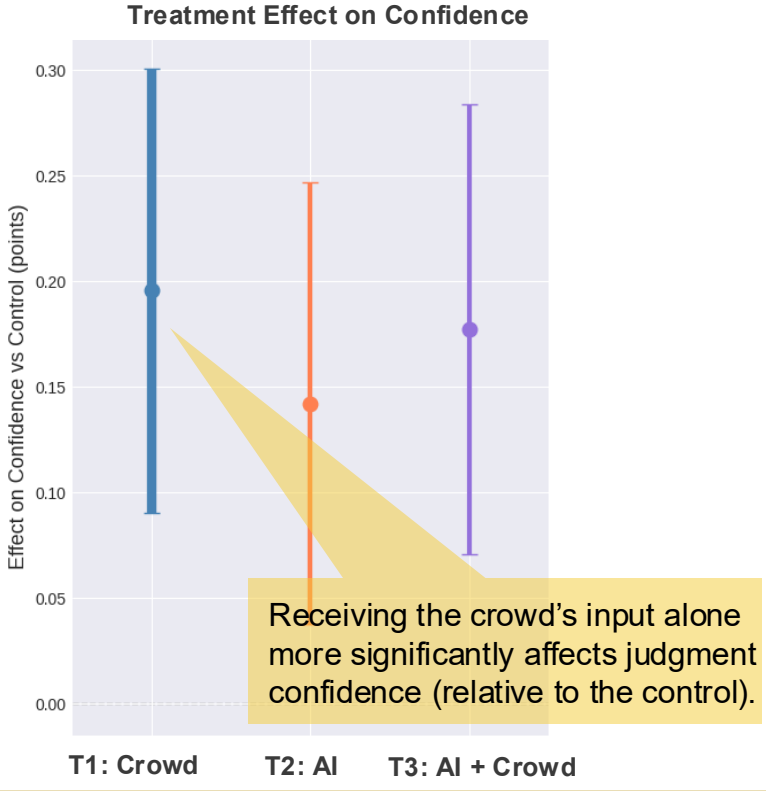
But how do participants incorporate the different sources?

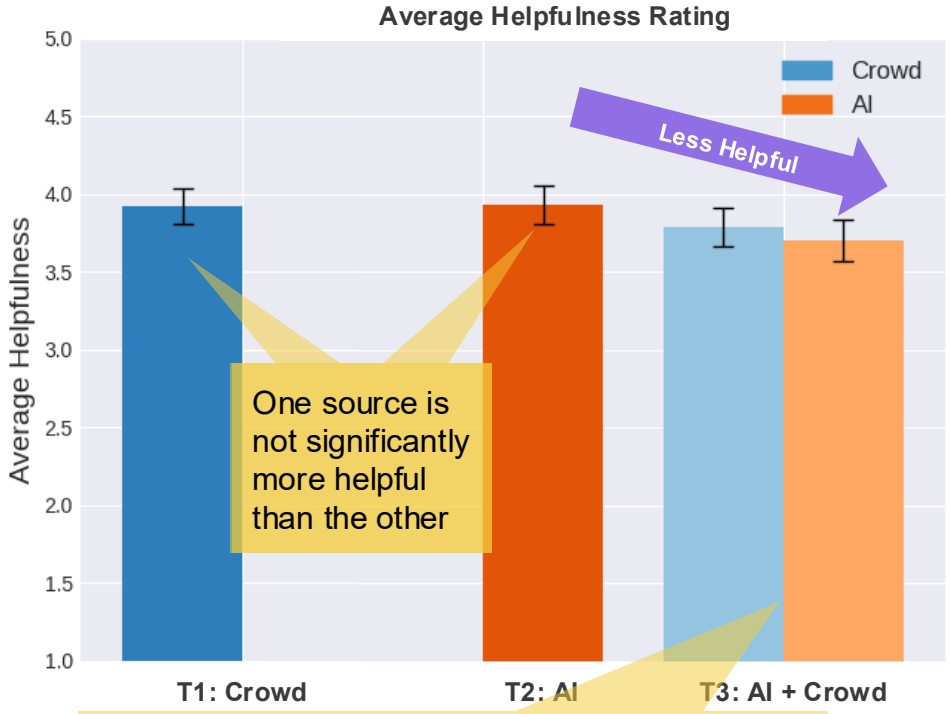
Study 2: “Splitting the Difference”



Participants blend both sources, but they also incorporate their own idiosyncratic judgement. This suggests advice integration is structured but not mechanical averaging.

Study 2: Perceived Confidence and Helpfulness





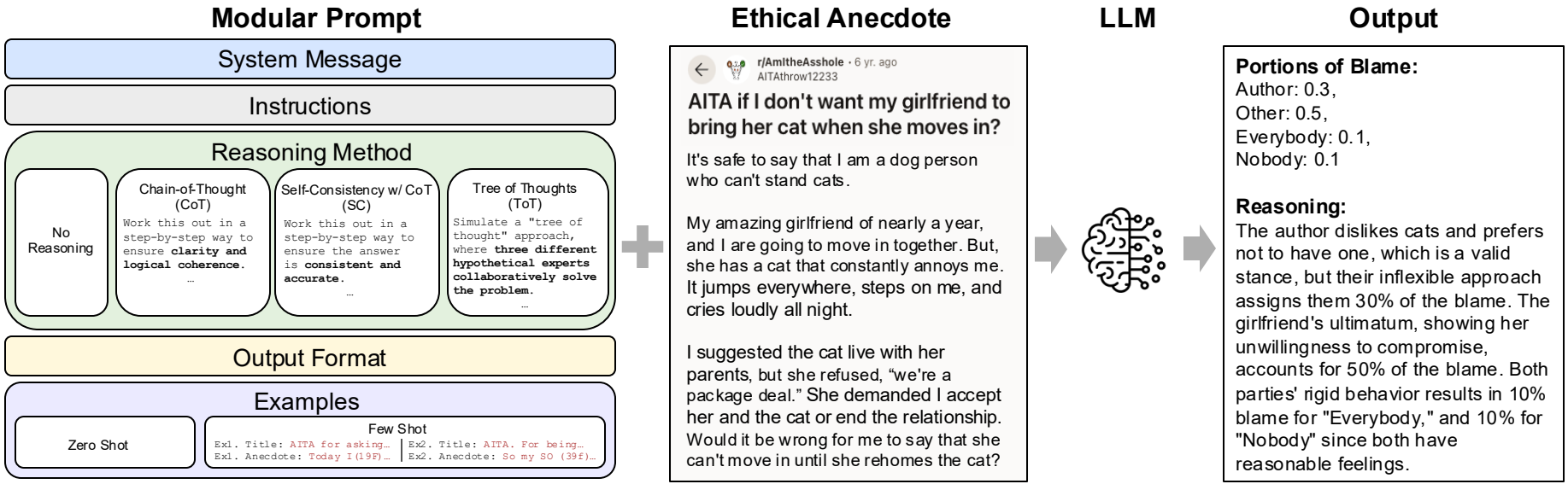
Study 2 Additional Outcomes: Helpfulness

	(1) T1 vs. T2: DV Helpfulness	(2) T3: DV Helpfulness
Baseline	*Advisor: AI*	*Advisor: AI*
Advisor: Crowd	−0.018	0.085
	(0.119)	(0.063)
AI Decisiveness (shown)	0.260*	0.246*
	(0.110)	(0.096)
Crowd Decisiveness (shown)	0.256*	0.124
	(0.104)	(0.092)
Constant	3.774***	3.498***
	(0.113)	(0.148)
Controls:		
Anecdote Text Length	✓	✓
Anecdote Category	✓	✓
Anecdote Sentiment	✓	✓
Model	Mixed-effect OLS	Mixed-effect OLS
Observations	$N_{T1} + N_{T2} = 740$	$N_{T3} * 2 = 710$

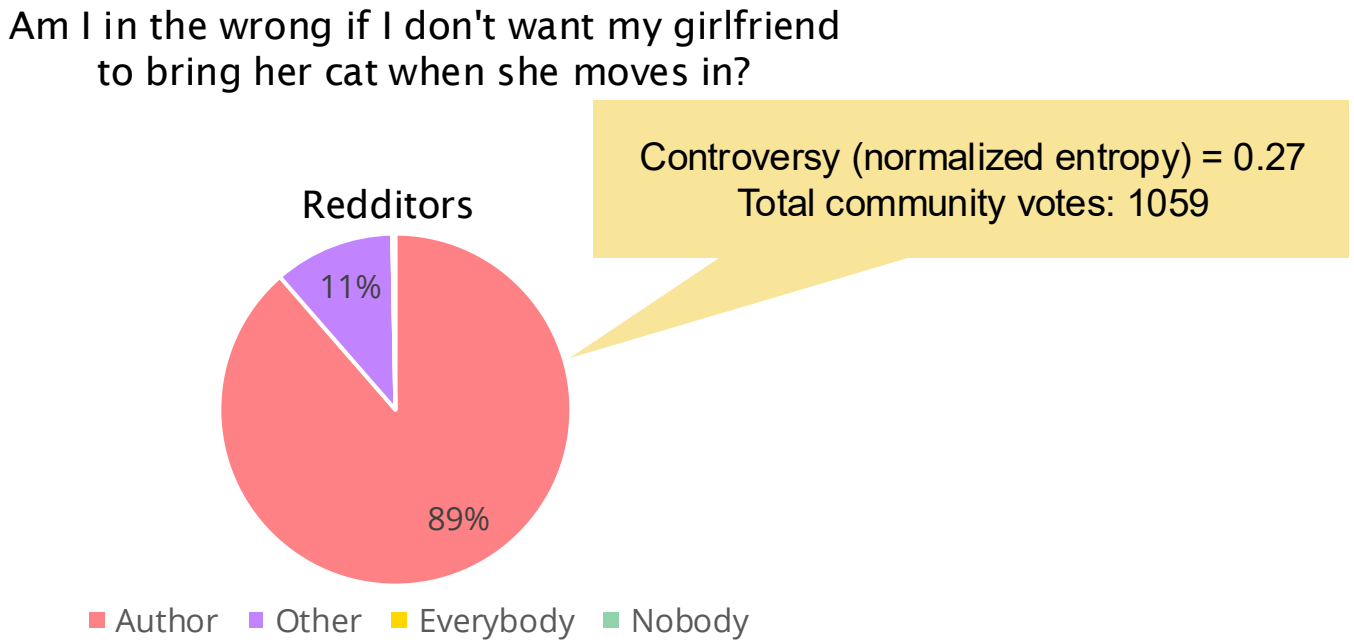
Notes: Standard errors in parentheses. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$.
Decisiveness terms are interacted with visibility masks; effects apply when that source is shown.
In T3 both sources are visible ("shown"), so the participant makes two helpfulness ratings per anecdote judgment.

Lack of significance → one advisor is not more helpful than the other

Prompting Framework



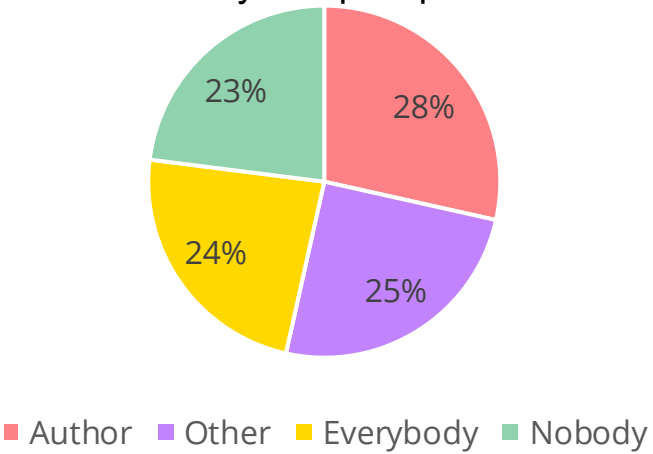
What of the Cat-Hating Boyfriend?



Normalized Entropy Calculation

Example of **high entropy (0.99) → highly controversial**
(no clear consensus)

Am I wrong for expecting my mom to want to support me through my first delivery and post partum?



Entropy formula:

$$H = -\sum_{\{i\}}^{n} p_i \ln(p_i)$$

Steps:

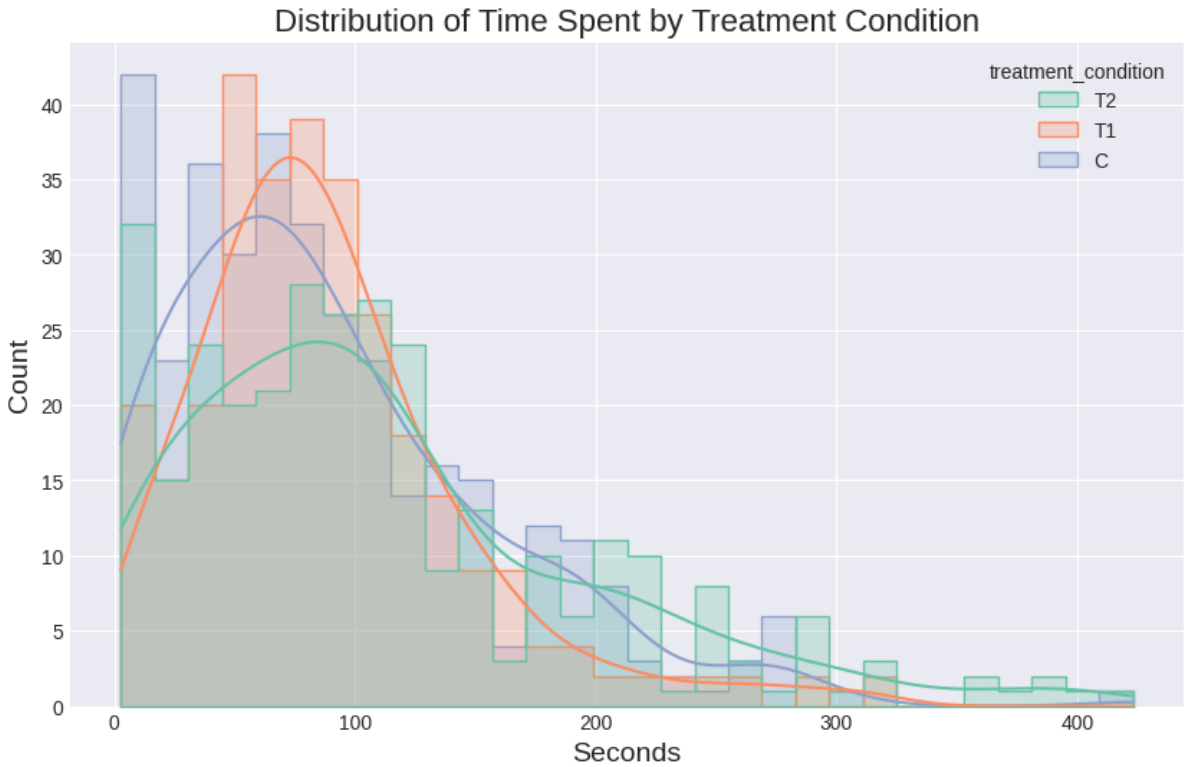
1. Calculate blame portions using votes: 0.28, 0.25, 0.24, 0.23
2. Substitute blame portions into formula:

$$H = -[0.28\ln(0.28) + 0.25\ln(0.25) + 0.24\ln(0.24) + 0.23\ln(0.23)]$$
$$H = \sim 1.385$$

3. Normalize using min and max entropy across all anecdotes

$$H = \sim 1.385 \rightarrow Normalized\ Entropy = 0.99$$

Study 1: Distributions of Time Spent per Anecdote



Study 2: Time Spent per Anecdote by Treatment