Paper	Authors	Venue	Year	Task (domain) & risk / stakes	Study Setup, Intended audience (IA) & Participants	Crowd worker role / task	Incentive Design / Scheme & Time spent	Excerpt	Participant Motivation / Incentivization mentioned in Discussion / Limitations?	Notes	Link	Objective / Specific RQs
Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems	Brian Y. Lim, Anind K. Dey, Daniel Avrahami	CHI	2009	Activity recognition of exercise (Stakes: ?)	IA: Evaluator of a (simulated) wearable device (that detects activity) Participants: 53+158 (unclear who and how recruited)	Task: Learning section (interact with system), Understanding sections (Fill-in-the-Blanks Test and Reasoning Test), Survey section (explain how the system works, report their perceptions of the explanations and system in terms of understandability, trust and usefulness.)	Base pay: $1 per participant Bonus $2 per participant "to motivate performance", unclear how calculcated Time taken: avg. 33-34 minutes	"Participants were each given $3 for completing the study ($1 base and a $2 bonus to motivate performance). A further $2 was offered to a few participants who participated in interviews conducted soon (up to a few days) after completing the task."	None	System: model predicts whether a user is exercising based on factors such as body temp, participants evaluate explanations	https://dl.acm.org/doi/10.1145/1518701.1519023	"...a large controlled study comparing the provision of explanations addressing four intelligibility type questions (why, why not, how to, and what if) [on user"s understanding and trust in the system]."
Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err	Berkeley J. Dietvorst, Joseph P. Simmons, Cade Massey	Journal of Experimental Psychology	2015	Sales forecast Students' performance forecasting	Participants: S1,2,4: Students; IA: MBA admissions officers S3: MTurk, 400+1000	S3: Attention check Task: 10 unincentivized, 1 incentivized - Predicting the rank (1 to 50) of individual U.S. states in terms of the number of airline passengers that departed from that state in 2011.	S3: Base pay: $1 per participant Bonus: max. possible $1 per participant for correct forecast (- $0.15 for each unit of distance (here, rank) from correct forecast) Time taken: not mentioned	"Participants received $1 for completing the study and they could earn up to an additional $1 for accurate forecasting performance." "First, participants who were not in the control condition completed 10 unincentivized forecasts instead of 15 in the first stage of the experiment. Second, in the second stage of the study, all participants completed one incentivized forecast instead of 10. Thus, their decision about whether to bet on the model's forecast or their own pertained to the judgment of a single state. Third, we used a different payment rule to determine participants' bonuses for that forecast. Participants were paid $1 if they made a perfect forecast. This bonus decreased by $0.15 for each additional unit of error associated with their estimate. This payment rule is reproduced in Appendix B. Fourth, as in Study 2, participants learned this payment rule before starting the first stage of unincentivized forecasts instead of after that stage."	Varying incentives for varying treatment - associating incentives with making the correct forecast (what do people do when profit is associated with outcome)	3 in-person studies with similar incentive schemes (but higher pay) also conducted "Incentivized" tasks: bonus for correct predictions Combination of unincentivized and incentivized tasks sort of training vs. main task (stakes associated become different)	http://dx.doi.org/10.1037/xge0000033	"In 5 studies, participants either saw an algorithm make forecasts, a human make forecasts, both, or neither. They then decided whether to tie their incentives to the future predictions of the algorithm or the human."
"Why Should I Trust You?" Explaining the Predictions of Any Classifier	Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin	KDD	2016	Religion prediction (20news) (Stakes: ?)	IA: model evaluator? Participants: MTurk, 100 per setting	Task: 1. examine explanations, select best model; 2. pick words to remove to improve explanations	None mentioned	-	None		https://arxiv.org/pdf/1602.04938.pdf	"...propose LIME, a novel explanation technique that explains the predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction." "...show the utility of explanations via novel experiments, both simulated and with human subjects, on various scenarios that require trust: deciding if one should trust a prediction, choosing between models, improving an untrustworthy classifier, and identifying why a classifier should not be trusted."
Human-Centric Justification of Machine Learning Predictions	Or Biran and Kathleen McKeown	IJCAI	2017	Stock price prediction (Stakes: seem high)	IA: Investors Participants: CrowdFlower, 33	Task: answer questions (if they would buy stock, if explanation was helpful, etc), make as much money as possible betting on the stocks, avg. 59 questions per participant	Base pay: not mentioned Bonus: Outcome based, "large" bonus to top 2 highest virtual money earning participants Time taken: not mentioned	"To keep it interesting and encourage the annotators to behave like investors, we offered (relatively) large bonuses to the two annotators who made the most virtual money."	None	Bonus doesn't seem performance based but more outcome based as it depends on the experimental condition (effects aren't already known)	https://www.cs.columbia.edu/nlp/papers/2017/biran_human_centric_justification_ijcai	"...novel approach to producing [ML prediction] justifications that is geared towards users without machine learning expertise, focusing on domain knowledge and on human reasoning, and utilizing natural language generation."
Hafez: an Interactive Poetry Generation System	Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, Kevin Knight	ACL	2017	Poetry generation (Stakes: seem low)	IA: Poets? Participants: MTurk, 62 HITs (seems 1 per participant)	Task: Generate and re-generate poems using model	Base pay: not mentioned Bonus: Performance-based, based on if self-picked "best poem" aligns with expert picked "best poem" Time taken: not mentioned	"Improving the quality of adjusted poems over the default poem is not required for finishing the task, but it is encouraged." "For each task, Turkers can select the best generated poem, and if subsequent human judges (domain experts) rank that poem as "great", a bonus reward will be assigned to that Turker"	None		https://aclanthology.org/P17-4008.pdf	"...propose an automatic poetry generation system" "...enables users to revise and polish generated poems by adjusting various style configurations"
Insights into Human-Agent Teaming: Intelligent Agent Transparency and Uncertainty'	Kimberly Stowers, Nicholas Kasdaglis, Michael Rupp, Jessie Chen, Daniel Barber, and Michael Barnes	Advances in Human Factors in Robots and Unmanned Systems	2017	Military planning (monitor and direct unmanned vehicles) (Stakes: seem high)	IA: a UxV system operator Participants: recruitment / profile details not mentioned	Task: "monitor and direct vehicles to carry out missions given to them by a simulated commander" - interpret commander's intent, understand vehicle and environmental constraints, decide whether to follow the IA's recommendation.	None mentioned	-	None		https://kimberlystowers.files.wordpress.com/2017/08/stowers-et-al-2017-insights-into-human-agent-teaming.pdf	"...discusses two studies testing the effects of agent transparency in joint cognitive systems involving supervisory control and decision-making. Specifically, we examine the impact of agent transparency on operator performance (decision accuracy), response time, perceived workload, perceived usability of the agent, and operator trust in the agent"
Do Explanations make VQA Models more Predictable to a Human?	Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, Devi Parikh	ACL	2018	(Visual) Question answering	IA: Model assessor? Participants: MTurk, 280	Task: Evaluate 100 question-image (QI) pairs (50 train, 50 test) - one HIT (two conditions: Failure prediction (FP), Knowledge prediction (KP)) For FP, predict if model will answer correctly For KP, predict model response (With/without explanations)	Base pay: avg. $3 per participant Bonus: Performance based, resulting (avg) $0.44 Time taken: For FP: 10.11 ± 1.09 mins For KP: 24.49 ± 1.85 min	"Subjects were paid an average of $3 base plus $0.44 performance bonus, per HIT" "At the end, they are shown their score and paid a bonus proportional to the score."	None	Clear pay structure not mentioned (for different tasks FP and KP time taken was different) Single base pay mentioned but unclear how calculcated, (performance-based) bonus number mentioned calculcated "proportional" to score	https://aclanthology.org/D18-1128.pdf	"...pursuing research directions to help humans understand the strengths, weaknesses, quirks, and tendencies of AI. We instantiate these ideas in the domain of Visual Question Answering (VQA), by proposing two tasks that help measure how well a human 'understands'
The accuracy, fairness, and limits of predicting recidivism	Julia Dressel, Hany Farid	Science Advances	2018	Recidivism (Stakes: high)	IA: Judges Participants: MTurk, 462+449	Task: Predicting crime for 50 defendent profiles Attention checks	Base pay: $1 per participant Performance bonus: $5 per participant, performance-based on overall accuracy >65% Time taken: not mentioned	"The participants were paid $1.00 for completing the task and a $5.00 bonus if their overall accuracy on the task was greater than 65%. This bonus was intended to provide an incentive for participants to pay close attention to the task."	None	Running accuracy shown after making each prediction - point about realizing if they will earn the bonus affecting performance	https://advances.sciencemag.org/content/4/1/eaao5580/tab-pdf	"... [evaluate] whether these algorithms are any better than untrained humans at predicting recidivism in a fair and accurate way."
Comparing Automatic and Human Evaluation of Local Explanations for Text Classification	Dong Nguyen	ACL	2018	Review sentiment analysis (movie) Topic classification (20news) (Stakes: not mentioned)	IA: Content analyser? Participants: 406+445, CrowdFlower (Australia, Canada, Ireland, UK US with quality levels two or three)	Instructions, test questions Task: Forward prediction (guess output of model based on local explanations), avg 17-18 predictions Rate confidence	Base pay: $0.03 per judgement Time taken: not mentioned	"Each HIT (Human Intelligence Task) was carried out by five crowdworkers. We paid $0.03 per judgement. On the 20news dataset, we collected 7,200 judgements from 406 workers (mean nr of. judgements per worker: 17.73, std.: 7.21) and on the movie dataset we collected 8,100 judgements from 445 workers (mean nr of. judgements per worker 18.20, std: 7.24)."	None		https://aclanthology.org/N18-1097.pdf	"...evaluating local explanations for text classification."
Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories	Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, Noah A. Smith	IUI	2018	Creative writing (stories and slogans) (Stakes: seem low)	IA: Writers, Writing evaluators Participants: MTurk Writing: 36 total (9 per condition) Evaluating: "9 evaluations for 108 writing samples (3 per participant)"; total = 324 (>1000 tasks, >95% acceptance rate, US)	Task: Writing: Solo / MIL story writing (10 sentences), Solo / MIL slogan writing Each condition: 3 rounds of writing + self-evaluation Post survey Open-ended interview Evaluating: evaluate 3 pieces of writing (questionnaire)	Writing: Base pay: $20 gift card Evaluation: Base pay: $0.15 per task (evaluation) Resulting pay: $0.45 per participant Time taken: not mentioned	"Participants were compensated with a $20 Amazon gift card."	None	MIL: Machine-in-the-loop	https://doi.org/10.1145/3172944.3172983	"explore the possibility of machine-in-the-loop creative writing" • How can we design machine-in-the-loop systems to support diverse writing tasks and processes? • What effect do these systems have on people's writing, both as perceived by the writer and by other people? • What do people want to see in machine-generated suggestions and creative writing support systems?
Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them	Berkeley J. Dietvorst, Joseph P. Simmons, Cade Massey	Management Science	2018	Students' performance forecasting	Participants: S2, 3: MTurk, 800+800	S2,3: Attention check Task: Predicting scores of students on a test. S2: 10 incentivized forecasts S3: 10+10 incentivized forecasts (2 rounds after choosing a condition)	S2,3: Base pay: $1 per participant Bonus: max. possible $0.5 (S2) or $1 (S3) per participant for correct forecast (- $0.10 for each unit of distance from correct forecast on average (over all incentivized forecasts)) Time taken: not mentioned	S2 (similar for S3) "...earned $1 for completing the study and could earn up to an additional $0.50 depending on their forecasting performance." "Participants were paid a $0.50 bonus if their official forecasts were within five percentiles of students' actual percentiles. This bonus decreased by $0.10 for each additional five percentiles of error in participants' forecasts (this payment rule is reproduced in Appendix B). As a result, participants whose forecasts were off by more than 25 percentiles received no bonus."	Incentives used to encourage accurate predictions	Follow up to "Algorithmic Aversion"	https://www.semanticscholar.org/paper/Overcoming-Algorithm-Aversion%3A-People-Will-Use-If-Dietvorst-	"...present three studies investigating how to reduce algorithm aversion [modifiable algorithms]."
Investigating Human + Machine Complementarity: A Case Study on Recidivism	Sarah Tan, Julius Adebayo, Kori Inkpen, Ece Kamar	?	2018	Recidivism (Stakes: high)	IA: Judges Participants: MTurk	Task: Predict risk of re-offense	None mentioned	-	None		https://arxiv.org/pdf/1808.09123.pdf	"focused efforts on cases where humans and machines disagree as a potential area to enhance decision making."
'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions	Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, Nigel Shadbolt	CHI	2018	Loan approval, Selecting overbooked airline passengers for re-routing, Determining to freeze bank accounts due to money laundering suspicion, Dynamically pricing car insurance premiums, Job promotion (Stakes: all seem high)	IA: (context-specific) Participants: Prolific Academic, 325+65,	Task: 12 cases for a context - evaluate 4 explanation styles	Incentives: none mentioned Time taken: av. 8.1 mins	-	"Threats to validity: ...the scenarios considered were hypothetical, not affecting the participants directly, and therefore lacked the first-person consequences and significance of a real world decision"; point about using incentives to simulate stakes/consequences		https://doi.org/10.1145/3173574.3173951	"...undertake three experimental studies examining people's perceptions of justice in algorithmic decision-making under different scenarios and explanation styles." 1. How do explanations for algorithmic decisions affect justice perceptions regarding algorithmic decisions? In particular, do the positive correlations observed between informational, procedural and distributive justice in human decision-making settings also hold in algorithmic decision-making settings? 2. How do different styles of explanation affect such justice perceptions?
Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing	Nina Grgic-Hlaca, Christoph Engel, Krishna P. Gummadi	CSCW	2019	Recidivism prediction (Stakes: high)	IA: Jury Participants: 20*X, Prolific (United States, self-reported to have served on a jury)	Study 1: predict Study 2: predict with outcome feedback Study 3: predict with feedback, incentivized	S1: Base pay: £2 per participant Time taken: ~25 mins S2: Base pay: £2.5 per participant Time taken: ~30 mins S3: Base pay: £2 per participant Bonus: Performance-based gain / loss configured according to different 'treatments'	"In Study 1, participants receive £2 for their participation, but are not incentivized for the predictions they make. The experiment takes approximately 25 minutes." "In Study 2, participants receive £2.5 for their participation. Due to receiving feedback, this experiment took slightly longer than the one in study 1 – approximately 30 minutes." "Study 3 uses the exact same design as Study 1, except that choices are incentivized." "Respondents earn a base payment of £2 for completing the survey. At the beginning of the survey, participants are informed that they can earn an additional monetary reward, based on their performance." Correct Incorrect False Positive False Negative Aligned Not Aligned Aligned Not Aligned Aligned Not Aligned Baseline 0 0 0 0 0 0 Ground Truth .2 .2 -.2 -.2 -.2 -.2 False Positive .2 .2 -.5 -.5 -.1 -.1 False Negative .2 .2 -.1 -.1 -.5 -.5 Weak Alignment .1 .5 .1 -.2 .1 -.2 Strong Alignment .1 .2 .1 -.5 .1 -.5 The values indicate the size of the monetary incentives used, in £.	Study design: Use of incentives to simulate stakes - "For obvious reasons, we cannot manipulate any of these incentives [real life incentives for judges] in our experiment. But there is a straightforward way of making predictions more consequential for participants: we can give them financial incentives. In the field, incentives are of course more subtle. " Hypothesizing on the effects of incentives (and tying them to the real world) "If even a financial incentive that is immediate – and directly tied to the prediction made in the concrete case – is not effective, a fortiori the more subtle incentives that policymakers might use as levers will also be ineffective. In Study 3, we thus investigate whether machine advice can be made more effective by increasing stakes." "different mistakes are incentivized differently" Suggesting use of alternative incentives: "It would also be interesting to bring alternative incentives for accuracy, or for being sensitive to machine advice, to the lab."	Incentives to emulate stakes Directly studies the effect of incentives and incentivizing in different directions on participant performance and outcomes; Incentivizing different aspect / behaviours - like: finding ground truth or following machine advice Results might be relevant for solution design: "unbiased incentives have virtually no effect." "has an effect, though, if participants lose money for [false positive]" sensitivity to machine advice unchanged when incentivized for ground truth incentives to avoid fp/fn leads to less p/n predictions and lower fp/fn rates, but relying on machine advice unchanged and no increase in accuracy incentives to align with machine advice- > more likely to change decision strong alignment -> increase in accuracy	https://doi.org//10.1145/3359280	systematically investigating the conditions under which machine advice improves the accuracy of human decisions Study 1 uses a within-subjects design. It investigates whether access to machine advice improves human predictions. We find a small effect, which is biased in the direction of predicting no recidivism. From a policy perspective, it may be worrisome that machine advice has little effect, given that deploying the tool is costly. This finding motivates Studies 2 and 3. In Study 2 we test whether giving human decision makers feedback about the performance of the machine moderates the effect; it does not. In Study 3, we emulate higher stakes, by giving participants a financial incentive. It only proves effective if participants gain money by following the advice.
Progressive Disclosure: Designing for Effective Transparency	Aaron Springer, Steve Whittaker	IUI	2019	Emotion analysis (Stakes: unclear)	IA: Psychology experts? Participants: S1: MTurk (prev. completed a subset of the Psychological General Well-being Index) S2: University (students)	Study 1: Write about an experience,answer questions, give feedback on E-meter assessment Study 2: Semi-structured interviews: think aloud	Study 1: $3.33 per participant Time taken: avg. 14.68 mins Study 2: Course credit	"Users were recruited from Amazon Mechanical Turk and paid $3.33. The evaluation took 14.68 minutes on average." "They received course credit for participation."	None		https://doi.org/10.1145/3301275.3302322	explore empirical user-centric methods to better understand user reactions to transparent systems. • RQ1: Do users prefer to use transparent systems? Do they prefer systems providing transparent feedback or those that simply present overall predictions? (Study 1) • RQ2: Do cognitive load and distraction affect preferences for transparency? What other factors influence reactions to transparency? (Studies 1 and 2) • RQ3: How might we support effective transparency? (Studies 1 and 2)

Will you Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems	Rafal Kocielnik, Saleema Amershi, Paul N. Bennet	CHI	2019	Meeting scheduling assistance (Stakes: seem low)	IA: Scheduling app users/devs? Participants: Internal crowd-sourcing platform similar to MTurk (US, aged 18+)	Study 1: Tutorial, survey, attention check Task: experience a particular condition of systems Survey Study 2: similar with more questions, and perform task with AI assistance	Study 1: Base pay: $1.35 per task Time taken: avg. 5:21 mins Study 2: Base pay $2.45 per task Time taken: avg. 10:35 mins	"... took on average 5:21 min (SD : 3.45 min). Participants were compensated $1.35 per task." "Each task took on average 10:35 min (SD : 6.22 min). Participants were compensated $2.45 per task."	None	"Each participant was allowed to complete the Study only once." Unclear if per task is per participation	http://library.usc.edu.ph/ACM/CHI2019/1proc/paper411.pdf	RQ1. What is the impact of an AI system's focus on avoidance of different types of errors on user perception? RQ2. What are the design techniques for setting appropriate end-user expectations of AI systems? RQ3. What is the impact of expectation-setting intervention techniques on user satisfaction and acceptance of an AI system?
An Evaluation of the Human-Interpretability of Explanation	Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, Finale Doshi-Velez	arXiv	2019	Alien medicine recommendation (i: high-risk, a: could be low risk) Alien recipe recommendation (i: low-risk)	IA: "Aliens" / Makers of ML model Participants: 150*6 (900), MTurk, US/Canada, <50 y/o, Bachelor's degree	Tutorial, practice questions Task: Simulate model response, verify suggested responses, and determine whether the correctness of a suggested response changes under a change to the inputs	None mentioned	"[Participants] were told that their primary goal was accuracy, and their secondary goal was speed." "We excluded participants from the analysis who did not get all of one of the two sets of three practice questions correct. While this may have the effect of artificially increasing the accuracy rates overall—we are only including participants who could already perform the task to a reasonable extent—this criterion helped filter the substantial proportion of participants who were simply breezing through the experiment to get their payment."	None	"Participants were given a tutorial on each task and the interface, and were told that their primary goal was accuracy, and their secondary goal was speed." ^Incentivizing for such goals?	https://arxiv.org/pdf/1902.00006.pdf	"...investigated how the ability of humans to perform a set of simple tasks—simulation of the response, verification of a suggested response, and determining whether the correctness of a suggested response changes under a change to the inputs—varies as a function of explanation size, new types of cognitive chunks and repeated terms in the explanation."
The Principles and Limits of Algorithm-in-the-loop Decision Making	Ben Green, Yiling Chen	CSCW	2019	Recidivism prediction (high stakes) Loan approval (high stakes)	IA: Judges, Loan agents Participants: MTurk (US, acceptance rate >=75%), 1156+732	Tutorial, comprehension test, an intro survey (to obtain demographic information and other participant attributes), the primary experimental task comprising a series of predictions, and an exit survey (to obtain participant reflections on the task, in the form of both multiple choice and free response questions), attention checks. Task: 40 predictions (1 context, 1 condition)	Base pay: $2 per participant Bonus: Performance-bases on accuracy of predictions, Brier score, max. possible $2 Resultant pay: $15.20 per hour (recidivism), $17.18 per hour (loan) Time taken: not mentioned	"Participants were paid a base sum of $2 for completing the study, plus an additional reward of up to $2 based on their performance. We allocated rewards following a Brier score function: score = 1 − (prediction − outcome)2 , where prediction ∈ {0, 0.1, . . . , 1} and outcome ∈ {0, 1}. The Brier score is bounded between 0 (worst possible performance) and 1 (best possible performance), and measures the accuracy and calibration of predictions about a binary outcome.2 We mapped the Brier score for each prediction to a payment using the formula payment = score ∗ $0.05, such that perfect accuracy on all 40 predictions would yield a bonus of $2. Because the Brier score is a proper score function [32], participants were incentivized to report their true estimates of risk. We articulated this to participants during the tutorial and included a question about the reward structure in the comprehension test to ensure that they understood." "Participants reported in the exit survey that the experiment paid well, was clear, and was enjoyable. Considering both the base payment and the bonus payment, participants in the pretrial setting earned an average wage of $15.20 per hour and participants in the loans setting earned an average wage of $17.18 per hour."	None (did talk about crowdworkers: "A significant limitation of this paper is that our findings are based on the behaviors of Mechanical Turk workers rather than judges or loan agents, meaning that we cannot assume that the observed behaviors arise in practice.")		https://doi.org/10.1145/3359152	(1) What criteria characterize an ethical and responsible decision when a person is informed by an algorithm? (2) Do the ways that people make decisions when informed by an algorithm satisfy these criteria?
Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment	Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, Casey Dugan	IUI	2019	Recidivism prediction (high: "carries weight to elicit reaction on fairness")	IA: Judges Participants: MTurk, 160 (US completed > 1000 tasks, >= 98% approval rate)	Task: 6 fairness judgment trials, make prediction, view model prediction, rate agreement, justify judgement Attention checks, give feedback on explanations Survey: individual differences, attention checks	Base pay: $3 Time taken: avg. 18 mins	"On average the study took 18 min to complete, and each participant was compensated with $3."	None (mentioned limitation of using crowdworkers)		https://doi.org/10.1145/3301275.3302310	"...conducted an empirical study with four types of programmatically generated explanations to understand how they impact people's fairness judgments of ML systems." "...empirical insights on how diferent styles of explanation impact people's fairness judgment of ML systems, particularly the diferences between a global explanation describing the model and a local explanation justifying a particular decision."
Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance	Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, Eric Horvitz	AAAI	2019	Defective object pipeline (i: high stakes (simulated using rewards))	IA: QA (engineer) Participants: MTurk, 25*X	Task: 100 rounds, decide whether or not the objects going over the pipeline are defective - accept or override ML model recommendation	Unclear if base pay or based on game-performance Resulting pay: $20 per worker	"After submitting a choice, the human receives feedback and monetary reward based on her final decision. Table 1 shows the payoff scheme used across these experiments, which aims to simulate high-stake decision making (i.e., the penalty for an incorrect action is much higher than the reward for a correct one)." "Marvin Correct Marvin Wrong Accept $0.04 -$0.16 Compute 0 0 Table 1: Payoff matrix for the studies. As in high-stakes decisions, workers get 4 cents if they accept Marvin when it is correct, and lose 16 cents if they accept Marvin when wrong." "For every condition we hired 25 workers and on average workers were paid an hourly wage of $20."	None	Simulating high stakes decision-making using incentives (high penalty); a little unclear if this penalty is in actual pay or in-game		"...studied the role of human mental models on the human-AI team performance for AI-advised human decision making for situations where people either rely upon or reject AI inferences. "
Assessing the Local Interpretability of Machine Learning Models	Dylan Slack, Sorelle Friedler, Carlos Scheidegger, Chitradeep Roy	NIPS Workshop HCML	2019	Math questions (synthetic data) (seems low risk)	Participants: Prolific, 40 (pilot) + 1000 (main) (at least a high school education, rating > 75 / 100)	Instructions, descriptions Task: 24 (8 inputs, 3 models), calculate the output of a model, then determine the output of a perturbed input applied to the same model. Survey, attention check	Base pay: $3.50 per participant Time taken: estimated avg. 20-30 mins	"We used Prolific to distribute the survey to 1000 users each of whom was paid $3.50 for completing it." "...we asked each user at the end of the study to indicate whether they fully attempted to determine correct answers and that they would still be compensated in case they selected no." "The main takeaways from these pilot studies were that we estimated it would take users 20-30 minutes to complete the survey, but that some users would take much longer. "	In actuality, user might give up but study takers took time and kept at (likely for fear of lack of compensation) "The time taken to simulate neural networks might not be feasible in practice. The neural network simulation time was noticeably greater than that of the decision tree and logistic regression. In some cases, the time expended was greater than 30 minutes. A user attempting the simulate the results of a model might give up or be unable to dedicate that much time to the task. The study takers likely feared lack of compensation if they gave up. This result suggests that in time constrained situations, neural networks are not simulatable."		https://arxiv.org/pdf/1902.03501.pdf	"...to assess the simulatability and "what if" local explainability of machine learning models, and to study the extent to which the proposed metric works as proxy for local interpretability"
Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders	Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, Haiyi Zhu	CHI	2019	Student admission prediction (high - "important")	IA: admissions officers, applying students Participants: MTurk, 202 (approval rate >= 90%, US residents, age >= 18)	Background survey Task: Explore interface, answer questions (evaluating understanding and trust)	Base pay: $2 per participant Bonus: performance-based on accuracy of objective questions, upto $3 Resulting: avg $3 per participant Time taken: 20 mins	"The average time for completing the survey was 20 minutes. Each participant received a base payment of $2 and an additional bonus (up to $3) based on the number of correct answers they gave for the objective understanding questions. On average, each participant received a payment of $3, which is above the US minimum wage ($7.25/ hour at the time of writing)."	None	No decision-making task performed, should include?	https://doi.org/10.1145/3290605.3300789	"...the goal is to help users and other stakeholders understand the "algorithmic decision model", rather than the process of model training"
On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection	Vivian Lai, Chenhao Tan	FAccT	2019	Deception detection (stakes not mentioned; "challenging", "complex")	IA: Not mentioned; Content moderator? Participants: MTurk (US, English fluency, completed 50 HITs, >=99% approval rate	3 training tasks with outcome feedback Task: 20 predictions Exit survey: Estimation of own performance, demographics	Base pay: $0.05 per prediction ($1 per participant) Bonus: Performance-based, $0.02 per correct prediction Time taken: 11 mins	"To incentivize turkers to perform at their best, we provide 40% bonus for each correct prediction in addition to the 5 cent base rate for a review." "If the HIT is approved, the turker is compensated a dollar and bonuses depending on the number of reviews he correctly predicted. For example, if a turker makes 11 correct predictions, he is compensated $0.22 in addition to a dollar. The average duration for finishing our HIT is about 11 minutes (Figure 7 shows the CDF of the duration)."	None		https://doi.org/10.1145/3287560.3287590	"...conduct the first empirical study to investigate whether machine predictions and their explanations can improve human performance in challenging tasks such as deception detection"
Disparate Interactions: An Algorithm-in-the-loop Analysis of Fairness in Risk Assessments	Ben Green, Yiling Chen	FAccT	2019	Recidivism prediction (Stakes: high)	IA: Judges Participants: MTurk (US >= 75% approval rate)	Tutorial, comprehension test, intro survey Task: 25 predictions Exit survey, attention checks	Base pay: $2 per participant Bonus: Performance-based, upto $2 (avg. earned $1.54), Brier scoring on correct prediction Resulting pay: $3.54 per participant ($20 per hour) Time taken: 20 mins	"Participants were paid a base sum of $2 for completing the survey, with the opportunity to gain an additional reward of up to $2 based on their performance during the experiment. We allocated rewards according to a Brier score function, mapping the Brier reward (bounded [0,1], see Section 4.1) for each prediction to a payment using the formula pament = reward ← $0.08 (since the test population is restricted to defendants who were released before trial, we have ground truth data with which to evaluate each prediction). Because the Brier score is a proper score function [28], participants were incentivized to report their true estimates of crime risk. We explicitly articulated this to participants during the tutorial and included a question about the reward structure in the comprehension test to ensure that they understood." "During the exit surveys, participants reported that the experiment paid well, was clear, and was enjoyable. Participants earned an average bonus of $1.54 (median=$1.56), making the average total payment $3.54. Participants completed the task in an average of 20 minutes (median=12), and earned an average wage of $20 per hour (median=$18). Out of 213 participants who responded to a free text question in the exit survey asking for any further comments, 32% mentioned that the experiment length and payment were fair."	None	Same as row 45 [28] Tilmann Gneiting and Adrian E. Raftery. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation	https://doi.org/10.1145/3287560.3287563	"...sheds new light on how risk assessments influence human decisions in the context of criminal justice adjudication." "...study how people make predictions about risk, both with and without the aid of a risk assessment."
What can AI do for me: Evaluating Machine Learning Interpretations in Cooperative Play	Shi Feng, Jordan Boyd-Graber	IUI	2019	Quizbowl (Stakes: seem low; "challenging")	IA: Quizbowl player Participants: 40 Experts (through forum); Novices: MTurk, 40	Task: Play quizbow (QnA game), at least 20 questions	None mentioned	"Experts are free to play as many questions as they want (but each player can only play a question once), and we encourage them to play more by offering monetary prizes for those who finish the whole question set." "Turkers usually stopped after answering the required twenty questions, but many experts kept on playing."	None		https://doi.org/10.1145/3301275.3302265	"...propose an evaluation of interpretation on a real task with real human users, where the effectiveness of interpretation is measured by how much it improves human performance."
Understanding the Effect of Accuracy on Trust in Machine Learning Models	Ming Yin, Jennifer Wortman Vaughan, Hanna Wallach	CHI	2019	Speed dating (i: both high and low stakes - high stakes simulated using reward for correct predictions)	IA: Not mentioned - domain a testbed; Dating app devs? Participants: MTurk, 1994 + 757 + 1042, (US)	Task: 40 predictions (initial and final) Exit survey	Base pay: $1.50 per participant Exp 1: Bonus: Performance-based, $0.10 per correct prediction Time taken: not mentioned	Exp 1: "As a robustness check to guard against the potential criticism that any null results might be due to a lack of performance incentives, we randomly selected some subjects to receive a monetary bonus for each correct prediction. We also posited and pre-registered two additional hypotheses: • [H3] The amount at stake has a significant effect on people's trust in a model before seeing the feedback screen. • [H4] The amount at stake has a significant effect on people's trust in a model after seeing the feedback screen." "...to test whether the effect of stated accuracy on trust varies when people have more "skin in the game."" "Subjects were randomly assigned to either low or high stakes. Subjects assigned to low stakes were paid a flat rate of $1.50 for completing the experiment. Subjects assigned to high stakes also received a monetary bonus of $0.10 for each correct (final) prediction(3) in addition to the flat rate of $1.50" Results: "...the amount at stake does not have an effect on laypeople's trust in a model, at least for the limited range of stakes used in our experiment." "Because our first experiment revealed that the amount at stake does not affect people's trust in a model, we did not select any subjects to receive monetary bonuses, nor did we pre-register any hypotheses about the amount at stake." Exp 2&3: "with a flat rate of $1.50 for completing the experiment."	Discussed in task design - to simulate high stakes Found result: Amount at stake did not have an effect on people's trust in the model	(3) The highest possible bonus was 40 × $0.10 = $4—i.e., substantially more than the flat rate of $1.50, thereby making the bonus salient [11]. [11] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. Incentivizing High Quality Crowdwork.	https://doi.org/10.1145/3290605.3300509	"examine whether laypeople's trust in a model, measured in terms of both the frequency with which they revise their predictions to match those of the model and their self-reported levels of trust in the model, varies depending on the model's stated accuracy on held-out data and on its observed accuracy in practice."
A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy	Joon Sung Park, Rick Barber, Alex Kirlik, Karrie Karahalios	CSCW	2019	Jellybean counting (stakes: seem low, chosen for simplicity and more reasons)	IA: N/A Participants: MTurk, 140 + 200 (US, 18+, 100 HITs, >= 95% approval rate)	Task: Estimate number of jellybeans with AI assistance	Study 1 & 2: Base pay: $1.50 per participant Bonus: To adjust for minimum wage (actual time taken exceeded expectation), $0.30 per participant (based on avg time taken) Resulting pay: Minimum wage, $ 7.25 per hour Time taken: 14.3 mins (S1), 13.9 mins (S2)	Study 1: "...the participants were initially paid $1.50 for their time through the standard payment system of MTurk. Our post-study analysis revealed, however, that the participants in Study 1 took longer than our expectation. Therefore, following the recent practice of using MTurk's bonus system that allows requesters to pay the workers extra money after the initial payment, we paid every participants in this study extra $0.30 (for an example, see [35]). This ensured us that participants were paid at least the US Federal minimum wage of $7.25 per hour."	None	[35] Mark E. Whiting, Grant Hugh, and Michael S. Bernstein. 2019. Fair Work: Crowd Work Minimum Wage with One Line of Code.	https://doi.org/10.1145/3359204	"...[study] the impact of an algorithm's speed on how users incorporate the algorithm's advice when making judgments in the context of simple visual recognition tasks." (RQ). Would a slow algorithm improve users' assessments of the algorithm's accuracy?
What You See Is What You Get? The Impact of Representation Criteria on Human Bias in Hiring	Andi Peng, Besmira Nushi, Emre Kiciman, Kori Inkpen, Siddharth Suri, Ece Kamar	HCOMP	2019	Hiring (Stakes: seem high)	IA: Hiring associates Participants: MTurk, 300*~17 (US, >90 % approval rate)	Task: Pick a candidate for hiring recommendation from a slate (one HIT) Exit survey	Resulting pay: $15 per hour Time taken: 5-10 minutes	"...compensate workers at a wage of $15 per hour. "	None	Each participant does only one HIT Not AI-assisted decision-making, Human-In-The-Loop perf compared to AI	https://arxiv.org/pdf/1909.03567.pdf	Research Question 1: Does balancing the gender distribution in candidate slates mitigate bias? How does this effect vary across different professions? Research Question 2: For professions where this intervention is not enough, does over-representation help? Research Question 3: How do personal features of the decision-maker, such as gender, impact human decisionmaking in hiring recommendations?

Algorithm Appreciation: People Prefer Algorithmic To Human Judgment	Jennifer M. Logg, Julia A. Minson, Don A. Moore	Organizational Behavior and Human Decision Processes	2019	Song rank order prediction Person weight estimation Attractiveness estimation	Participants: Experts and MTurk	Several contexts, prediction tasks	Base pay: not mentioned Study 3: Bonus: max possible $1 (-$0.15 for each unit of distance from correct answer) All other studies: Bonus: Performance-based on correct final predictions, entry into raffle of $10	"Each exact correct final answer entered them into a raffle for $10." "Accurate final responses entered participants into a raffle for a $10 bonus." "We incentivized mTurk participants by entering them into a raffle for $10. Participants who made more accurate forecasts received more entries into the raffle" "The more accurate participants' answers, the greater their bonus payment (from $0.10 for an answer within 6 ranks of the truth, increasing in 15 cent increments for each closer rank, to $1.00 for a correct answer).	"...weight on advice, especially under incentivized conditions, captures actual change in the participant's own judgment as a function of exposure to the advice" Use of incentives to make the decision "more important" - incentivized prediction	Follow up to "Algorithmic Aversion" studies by diff. researchers Only one study followed similar incentive schemes	bs.edu/ris/Publication%20Files/17-086_610956b6-7d91-4337-90cc-5bb5245316a8.pdf	Examine algorithmic appreciation under various conditions
Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff	Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, Eric Horvitz	AAAI	2019	Defective object pipeline (Stakes: Simulated high using reward)	IA: QA (engineer) Participants: MTurk, 25*X	Familiarize Task: 100 rounds, label objects defective or non-defective with AI assistance Self-report model accuracy	Unclear if base pay or based on game-performance Resulting pay: $20 per worker	"Accept Compute AI right $0.04 0 AI wrong -$0.16 0 Table 1: Reward matrix for the user studies. To mimic highstakes domains, penalty for mistakes is set to high." "The matrix is designed in a way that it imitates a high-stakes scenario, i.e., the monetary penalty for a wrong decision is much higher than the reward for a correct decision. We found this design choice to be a good incentive for workers to learn and update their mental model on Marvin." "Workers were paid on average $20/hr, over the minimum wage in line with ethical guidelines for requesters (Dynamo 2017)."	Used to simulate high stakes	Similar to "Beyond Accuracy" (that one seems like the follow up to this) (Dynamo 2017) Guidelines for Academic Requesters.	https://ojs.aaai.org//index.php/AAAI/article/view/4087	"...studied how updates to an AI system can affect humanAI team performance and introduced methods and measures for characterizing and addressing the compatability of updates."
COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations	Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, Brian Y. Lim	CHI	2020	Property price prediction (Stakes: seem low)	IA: Real estate experts? Participants: MTurk (>5000 completed HITs with >97% approval rate)	Introduction, tutorial, memory test Task: Reading, memorization, recall, recognition, counterfactual Post survey questionnaire	Base pay: $2.50 per participant Time taken: 30-40 mins	"Participants were compensated US$2.50 and completed the survey in about 30 to 40 minutes."	None		https://doi.org/10.1145/3313831.3376615	AI framework to generate explanations with desired cognitive load and accuracy
The limits of human predictions of recidivism	Zhiyuan "Jerry" Lin, Jongbin Jung, Sharad Goel, Jennifer Skeem	Science Advances	2020	Revidivism Prediction (Stakes: high)	IA: Judges/jury Participants: MTurk	Task: Prediction Attention checks	Base pay: $1 per participant Bonus: Performance based - Brier scoring on accuracy Resulting pay: ~$25 per hour Time taken: not mentioned	"Each participant received $1 for completing the study and a bonus of up to $5 based on performance. Following previous work (30), we measured performance via Brier scoring, an incentive-compatible payment scheme for eliciting probabilities (31–33). For each question i, the Brier score is 1 − (Yi − pi)2, where pi is the probability of recidivism reported by the participant, Yi = 1 if the individual described was indeed arrested for a new (violent) crime within 2 years of release, and Yi = 0 otherwise. A participant's final score was computed by summing the Brier scores earned for the 50 substantive questions, excluding the two attention checks. Across participants and experiments, the average hourly compensation was approximately $25."	None	(30): An algorithm-in-the-loop analysis of fairness in risk assessments (31). G. W. Brier, Verification of forecasts expressed in terms of probability. (32). K. Rufibach, Use of Brier score to assess binary predictions. (33). J. Hernández-Orallo, P. Flach, C. Ferri, A unified view of performance metrics: Translating threshold choice into expected classification loss	https://www.science.org/doi/pdf/10.1126/sciadv.aaz0652	Test the impact of three conditions on the relative accuracy of human judgment and RAIs in predicting reoffense; designed to illuminate both the situations in which humans can predict recidivism as accurately as algorithms and settings in which algorithms can provide better estimates than humans.
Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making	Yunfeng Zhang, Q. Vera Liao, Rachel K. E. Bellamy	FAccT	2020	Income prediction (Stakes: low)	IA: ? Participants: MTurk	Instructions, additional info, training tasks Task: 40 Prediction tasks Demographic survey	Base pay: $3 per participant Performance bonus: $0.05 per correct and -$0.02 per incorrect final prediction Resulting pay of ~$4.16 per participant Time taken: 30 mins	"We took two measures to improve the ecological validity. First, the decision performance was linked to monetary bonus, with a reward of 5 cents if the final prediction was correct and a loss of 2 cents if otherwise (in addition to a base pay of $3). Prior research showed that such a reward design is effective in motivating participants to optimize the decision outcome [2, 31]" "As discussed, participants received a base pay of $3 in addition to the performance-based bonus payment (plus 5 cents if correct and minus 2 cents if wrong). On average, each participant received $1.16 bonus, and a total of $4.16 compensation for completing the half-hour long experiment."	Rewards as mitigation for limitation of no participant 'responsibility': "Another limitation is that we use a contrived prediction task where the participants would not be held responsible. We mitigated the problem by introducing an outcome based bonus reward, which prior studies suggest could effectively motivate optimizing the decision-making. While future study could experiment with scenarios with more significant real-world impact, we note that they have to be executed with caution to avoid ethical concerns."	Rewards to improve ecological validity, motivate performance [2] Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff, AAAI 2019 [31] Twelve-choice probability learning with payofs. Psychonomic Science 7, 10 (1967)	https://arxiv.org/pdf/2001.02114.pdf	(RQ1): How does showing AI's prediction versus not showing, afect trust, accuracy of AI-assisted predictions, and the efect of conidence score on trust calibration? (RQ2): How does knowing to have more domain knowledge than the AI afect humans' trust, accuracy of AI-assisted predictions, and the effect of conidence score on trust calibration?
Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems	Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, Elena L. Glassman	IUI	2020	Nutrition prediction (Stakes: low)	IA: Nutritionist? Participants: MTurk (200, US adults)	Information about study Task: 24 questions, divided into 2 blocks Mid and end task questionnaires	Base pay: $2 per participant Time taken: avg. 7 mins	"The study lasted 7 minutes on average. Each worker was paid 2 USD."	None		https://doi.org/10.1145/3377325.3377498	Evaluate two currently common techniques for evaluating XAI systems: (1) using proxy, artificial tasks such as how well humans predict the AI's decision from the given explanations, and (2) using subjective measures of trust and preference as predictors of actual performance
Mental Models of AI Agents in a Cooperative Game Setting	Katy Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Maria Ruiz, Sarah Miller, David Millen and Werner Geyer	CHI	2020	Word guessing (Stakes: low)	IA: Game player Participants: Study 1: IBM employees Study 2: MTurk, 113	Study 1: Play game with In-person think aloud, semi-structured interview Study 2: Play game, survey	Base pay: $3 per participant Resulting pay: $15 per hour Time taken: est. avg. 11-13 mins	"In pilot studies, the average time for completion for the 5 game condition was 11 minutes and for the 10 game condition was 13 minutes. Based on this, all participants were paid $3 for the task, or about $15/hour.	None		https://www.katygero.com/papers/2020_MentalModelsofAIAgents.pdf	study people's mental models of AI in a cooperative word guessing game RQ1 What should conceptual models of AI systems include? RQ2 How do users develop mental models of AI systems? RQ3 What encourages accurate mental models of AI systems?
Do I Look Like a Criminal? Examining how Race Presentation Impacts Human Judgement of Recidivism	Keri Mallari, Kori Inkpen, Paul Johns, Sarah Tan, Divya Ramesh, Ece Kamar	CHI	2020	Recidivism prediction (Stakes: high)	IA: Judges/jury Participants: 1600, MTurk	Task: Rate for recidivism Catch trials (attention checks)	Base pay: $1.50 per task completion Bonus: $5 per task completion for overall accuracy >=65% Resulting pay: $9-$39 per hour Time taken: est. 10 mins	"Our second modification was increasing the payment to users to be $1.50 (from $1.00 paid by Dressel and Farid) for completion of the task, with a $5.00 bonus (same as Dressel and Farid) for reaching an overall accuracy of 65% or higher. This change was made to ensure that we were paying ethical wages [33]. On average, the task was expected to take 10 minutes to complete, so workers would earn $9 or $39 per hour(2), depending on whether they received the bonus." (2)If bonus is received, the estimated hourly rate is $39=$(1.50+5)*6.	None	Dressel and Farid: 2018. The accuracy, fairness, and limits of predicting recidivism [33]: Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. Paying an ethical wage	http://dx.doi.org/10.1145/3313831.3376257	investigate the validity and generalizability of Dressel and Farid's results and the implications it has on the design of evaluation systems, as well as studies utilizing Mechanical Turk workers.
Leveraging Rationales to Improve Human Task Performance	Devleena Das and Sonia Chernova	IUI	2020	Chess playing (Stakes: seem low)	IA: Chess game players Participants: 68, MTurk	Qualification test (rules of game) Task: Play chess with hints, rationales given by AI	Base pay: $2 (for day 1) + $4 (for day 2) + $6 (for day 3) per participant Time taken: ~15-20 mins *3	"Each daily session took approximately 15-20 minutes, and participants were compensated $2.00, $4.00, and $6.00 on days 1, 2 and 3, respectively."	None	Multi-day study; Increasing base amount over the days - seemingly to encourage retention	https://arxiv.org/pdf/2002.04202.pdf	Given a computational system whose performance exceeds that of its human user, can explainable AI capabilities be leveraged to improve the performance of the human?
No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML	Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Dan Weld, Leah Findlater	CHI	2020	Email topic classification (Stakes: seem low)	IA: E-mail users? Participants: MTurk ("Masters" qualification, located in the US, completed > 500 HITs, approval rate >=98%)	Task: Review a text classification model's predictions with / without explanations	Base pay: not mentioned Bonus: Fixed, $2 per participant Time taken: avg. 22.6 mins	"Remote study sessions took on average 22.6 minutes (SD = 15.3). " "To motivate quality work, participants were told that at least the top 50% of participants would be given a $2 bonus based on the thoroughness of their evaluations; unbeknownst to them, all ultimately received the bonus."	User motivation as a limitation: "However, the degree of their frustration would likely vary along with their actual desire and ability to provide feedback in more realistic settings. All are likely affected by task and model complexity, task importance (and therefore user motivation), and domain expertise."	Incentives to motivate performance Base pay not mentioned Incentive communication: participants told it would be paid to only top 50% of them but was actually paid to all	http://dx.doi.org/10.1145/3313831.3376624	investigates how explanations shape users' perceptions of ML models with or without the ability to provide feedback to them: (1) does revealing model flaws increase users' desire to "fix" them; (2) does providing explanations cause users to believe—wrongly—that models are introspective, and will thus improve over time
"Why is 'Chicago' deceptive?" Towards Building Model-Driven Tutorials for Humans	Vivian Lai, Han Liu, Chenhao Tan	CHI	2020	Deception detection (Stakes: low)	IA: Content moderators? Participants: User study: University mailing list Experiment: 480*3, MTurk (US, >= 50 HITs completed, 99% of HITs approved)	In-person semistructured user study: tutorial, think aloud, interviews, feedback on design to gather insights on how system is being used Experiment: explanation of task, attention check, tutorial (training phase), predicting labels of 20 reviews, exit survey: demographic, feedback	User study: $10 per half hour Experiment: Base pay: $2.5 per participant Bonus: Performance-based, $0.05 for each correct label Time taken: 10-19 mins	In person-user study: "Participants were compensated between $15 and $20 for $10 every 30 minutes." Experiment: "Each participant was compensated $2.50 and an additional $0.05 bonus for each correctly labeled test review."	None		https://arxiv.org/abs/2001.05871	• RQ1: Do model-driven tutorials improve human performance without any real-time assistance? • RQ2: How do varying levels of real-time assistance affect human performance after training? • RQ3: How do model complexity and explanation methods affect human performance with/without training?
Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study	Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, Nadia Berthouze	IUI	2020	Image classification (Stakes: seems low)	IA: Model user/dev? Participants: Prolific (approval rate > 95%, normal / corrected to normal vision, fluent in English, 18+, technical background (i.e. a degree in computing or engineering)	Tutorial Task: predict the classification outcome of the model, answer certain questions	Base pay: £8 per participant Bonus: Performance-based £0.5 per correct answer Time taken: est. max. 40 mins	"To increase participants engagement in the study, in addition to an £8 payment for their time, participants received an additional performance-based bonus of £0.5 for each correct answer as an incentive."	None	Incentives to encourage engagement	https://doi.org/10.1145/3377325.3377519	RQ1 Do saliency maps allow users to develop a better understanding of how the CNN model classifes a class of images? We measured this by the participant success to predict the system outcome on the task images. RQ2 Do Scores infuence the participant ability to predict the system outcome on the task images? RQ3 When saliency maps are present, do users pay more attention to detailed features?
Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity	Samuel Carton, Qiaozhu Mei, Paul Resnick	ICWSM	2020	Toxicity classification (Stakes: ?)	IA: Content analyzers/moderators Participants: MTurk (US-based, completed at least 1000 HITs with 95% acceptance)	Phase 1: Ground truth re-collection, task: label each comment, attention checks Phase 2: predict the (majority) outcome of phase 1 with AI assistance	Phase 1: Base pay: $1.50 Bonus: $0.50 per correct attention check Phase 2: Base pay: $1.25 Bonus: $0.50 per correct prediction (relative to ground truth collected in phase 1) Time taken: not mentioned	"Phase 1 workers were compensated with a base paymentof $1.50 plus a bonus of $0.50 for each attention check they marked correctly." "Workers in Phase 2 were given a base payment of $1.25 plus a bonus of $0.05 for each item they predicted correctly relative to the aggregated results of Phase 1. We didn't use any other quality assurance mechanism for two reasons. First, we were relying on the natural desire of our subjects to maximize their earnings under the stipulation of the error-prone model. Second, because we were interested in measuring speed, we wanted to simulate a smoother perceived trade-off between effort and reward. If we had included attention checks on this task, subjects would have been strongly incentivised to carefully read every token, which would have potentially masked any effect on subject speed arising from the presence of explanations."	None	Incentives for quality assurance - Relying on desire to maximize earnings No attention checks - not incentivizing careful consideration (measuring speed) Unclear if base pay per participant or per task (probably the former) There were 2 attention checks	https://ojs.aaai.org/index.php/ICWSM/article/view/7282	experimental assessment of the impact of fea-ture attribution-style explanations on human performance inpredicting the consensus toxicity of social media posts with advice from an unreliable machine learning model RQ1: Presence of model predictions. How does the adviceof an unreliable predictive model affect subject performancein predicting consensus toxicity of social media comments? RQ2: Presence of explanations. Do (attribution-style) ex-planations help users make better use of advice from suchan unreliable model? RQ3: Explanation type. Do more minimal "partial" orsparser "keyword" explanations exhibit different perfor-mance properties from explanations optimized for completeness?
An Empirical Study on the Perceived Fairness of Realistic, Imperfect Machine Learning Models	Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez	FAccT	2020	Bail outcomes prediction (Stakes: high)	IA: Judges/jury Participants: MTurk (18+, US, 95%+ rating over 500+ HITs)	Task: Survey-based - compare two models for deciding whether to grant bail to criminal defendants, rating-based questions	Base pay: $2.50 per participant Time taken: med. 14 mins	"We paid $2.50 for the survey, which took a median time of 14 minutes. We excluded data from participants whose free responses were off-topic or nonsensical. Exclusion happened after data collection, and all participants were paid regardless of whether we excluded their data from analysis."	None		https://doi.org/10.1145/3351095.3372831	gauge perceptions of the fairness of such realistic, imperfect models • RQ1 When choosing between models exhibiting the two sides of a difficult trade-off, which do people prioritize? • RQ2 What models that encapsulate difficult, yet realistic, trade-offs do people perceive as fair or biased? • RQ3 Do people prefer to use an imperfect model or rely on a human judge? • RQ4 To what extent do responses vary based on which racial group the model disadvantages?
I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI	Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, Andreas Butz	IUI	2021	Loan risk prediction (Stakes: high)	IA: Loan risk managers Participants: S1: Moderated, 40 educated lay-users (university mailing list) S2: Unmoderated, 107 crowdworkers, Prolific (100% approval rate, 10 prev submissions, UG degree)	Task 1: understand how AI forms its predictions Task 2: explore the decision-making behavior of the model Task 3: write a detailed explanation of their global understanding of the model's prediction behavior Task 4: simulate the prediction of the ML model Task 5: (participants with incorrect predictions shown results) re-examine the ML model behavior After each: rate perceived understanding Post-session questionnaire, report demographics	Base pay: £3.75 per completion (=£7.09/hour) Time taken: avg. 28.5 mins	"£3.75 per completion (=£7.09/hour)"	None	- Crowdworkers shown timer to see how much time they had already spent on a task	https://dl.acm.org/doi/10.1145/3397481.3450644	Whether non-technical users of such XAI systems are prone to an IOED (illusion of explanatory depth) • RQ1: How robust is a self-reported global understanding gained from local explanations when examined? • RQ2: What do non-technical XAI users do to construct a global understanding from local explanations?

Explainable Active Learning (XAL): An Empirical Study of How Local Explanations Impact Annotator Experience	Bhavya Ghai, Q. Vera Liao, Yunfeng Zhang, Rachel Bellamy, Klaus Mueller	CSCW	2021	Predicting annual income (Stakes: low?)	"interaction elicitation study" (map out desired interactions for people to teach models based on its explanations) IA: ? Participants: MTurk (98% approval rating, each participant only once)	Training: Look at link to supporting docs, practice trials with [outcome feedback] Task: 20 instances, make own / judge model prediction, (optionally) explain judgement, rate explanantion, (optionally) explain; Some info presented to keep engagement, attention checks, post task survey: subjective perception of the ML model, report demographic information and factors of individual differences	Base pay: $4 per participant Bonus: Performance-based, among top 10% -> 10% chance of $2 Time taken: 20-40 mins	"Participants spent about 20-40 min on the study and was compensated for $4 with a 10% chance for additional $2 bonus" "Incentivized with a $2 bonus if the consistency between their predictions and similar cases reported in the Census survey were among the top 10% of all participants."	Limitation: small-scale crowdsourcing study; A Mitigation: Rewards to improve ecological validity: "However, we attempted to improve the ecological validity by carefully designing the domain knowledge training task and reward mechanism (participants received bonus if among 10% performer)."		https://arxiv.org/pdf/2001.09219.pdf	• RQ1: How do local explanations impact the annotation and training outcomes of AL? • RQ2: How do local explanations impact annotator experiences? • RQ3: How do individual factors, specifically task knowledge, AI experience, and Need for Cognition, impact annotation and annotator experiences with XAL? • RQ4: What kind of feedback do annotators naturally want to provide upon seeing local explanations?
To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making	Zana Buçinca, Maja Barbara Malaya, Krzysztof Z. Gajos	CSCW	2021	Nutrition prediction (Stakes: low)	IA: Nutritionist? MTurk (199 participants in 3 batches, US residents, only once)	Consent form, instructions, (optional) demographic survey Task: shown meal images and asked to replace the ingredient highest in carbohydrates on the plate, with an ingredient that was low in carbohydrates, but similar in flavor; 26 questions Diff conditions: no AI, on demand, update (decide twice), wait (30 seconds) Post session questionnaire: subjective experience Mid session: Need for Cognition questionnaire	Base pay: $2.5 per participant ($10 / hour) Bonus: Performance-based, top performer of batch -> 1.1% chance of $3 Time taken: avg. 15 mins	"The study took 15 minutes on average to complete. Each participant was paid $2.5 (USD) for an estimated rate of $10 per hour." "To motivate participants to perform well on the task, the top performer of each batch was rewarded with a bonus of $3"	None	- studied people with different levels of Need for Cognition (i.e., motivation to engage in effortful mental activities)	https://scholar.harvard.edu/files/zbucinca/files/bucinca2021trust.pdf	examine whether cognitive forcing functions are successful in reducing human overreliance on the AI when working on a decision-making task
Manipulating and Measuring Model Interpretability	Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, Hanna Wallach	CHI	2021	Property price prediction (Stakes: seem low)	IA: Real estate expert? MTurk (1250 participants, country: US, 97% approval rating)	Instructions, understanding check, training Task: Guess model prediction, state confidence, see actual prediction, state confidence, make own prediction	Base pay: $2.5 per participant Time taken: not mentioned	"Each participant received a flat payment of $2.50."	None	No mention of time taken to complete the task	https://arxiv.org/pdf/1802.07810.pdf	number of features and the transparency of the model - investigated how these factors affected three measurable outcomes: (1) How well can people simulate a model's predictions? (2) To what extent do people follow a model's predictions when it is beneficial for them to do so? (3) How well can people detect when a model has made a mistake and correct for it?
Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance	Gagan Bansal, Tongshuang Wu, Joyce Zhu, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, Daniel S. Weld	CHI	2021	Lsat question answering Review sentiment analysis (Stakes: seem low)	Mixed methods: 2 studies IA: LSAT test takers, content analysis expert? Participants: MTurk (508 (filtered 100+100), country: US, 97% approval rating, min. 1000 approved tasks)	Screening phase Task: Label 50 samples Post task survey	Base pay: $0.5 per participant Main task: Performance bonus (combination of linear and step functions on accuracy): For sentiment classification: $0.05 per every correct decision + $0.50 if the total accuracy exceeded 90% or $1.00 if it exceeded 95% For LSAT: $0.30 for every correct decision + $1.00, $2.00, and $3.00 for reaching an overall accuracy of 30%, 50%, and 85% Survey: Fixed bonus, $0.25 Resulting pay: S1: avg. $3.35 ($15.77 per hour), S2: avg $6.30 ($23.34 per hour). Time taken: S1: avg. 13 mins, S2: avg. 16 mins	Study 1 (Sentiment classification): "Participants received a base pay of $0.50 for participating, a performance-based bonus for the main task, and a fixed bonus of $0.25 for completing the survey. Our performance-based bonus was a combination of linear and step functions on accuracy: we gave $0.05 for every correct decision in addition to an extra $0.50 if the total accuracy exceeded 90% or $1.00 if it exceeded 95%. The assigned additional bonuses were intended to motivate workers to strive for performance in the complementary zone and improve over the AI-only performance [33]. Since we fixed the AI performance at 84%, humans could not obtain the bonus by blindly following the AI's recommendations. Participants spent 13 minutes on average on the experiment and received an average payment of $3.35 (equivalent to an hourly wage of $15.77)." Study 2 (LSAT): "Participants received a base pay of $0.50 for participating, a performance-based bonus of $0.30 for each correct answer in the main task, and a fixed bonus of $0.25 for completing an exit survey. They received an additional bonus of $1.00, $2.00, and $3.00 for reaching an overall accuracy of 30%, 50%, and 85% to motivate workers to answer more questions correctly and perform their best. The average completion time for the LSAT task was 16 minutes, with an average payment of $6.30 (equals an hourly wage of $23.34)."	None	Rewards for motivating performance Unclear if bonus for reaching a certain overall accuracy paid overall or per correct decision (probably the former, need to refer to [33] to cross-check) [33] "Incentivizing High Quality Crowdwork" https://doi.org/10.1145/2736277.2741102	https://arxiv.org/pdf/2006.14779.pdf	"We ask if AI explanations help achieve complementary team performance, i.e.whether the team is more accurate than either the AI or human acting independently. We conducted large-scale experiments with more than 1,500 participants. Importantly, we selected our study questions to ensure that our AI systems had accuracy comparable to humans and increased the opportunity for seeing complementary performance."
Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making	Xinru Wang, Ming Yin	IUI	2021	Recidivism prediction (Stakes: high) Forest cover prediction (Stakes: low)	IA: Judges/jury, forest cover experts Participants: MTurk (country: US, only once)	Background survey (demographics, technical literacy, expertise in ML), tutorial, qualification questions, training tasks Task: 32 decision making tasks: initial prediction, review model prediction (and explanation), and make a final prediction Exit survey	Base pay: $1.80 (Recidivism), $2.00 (Forest cover) per participant Bonus: Performance based, $0.03 for each correct final prediction if overall final accuracy >=60% Survey: $0.10 for each correct answer Time taken: not mentioned	"The base payment of the experiment was $1.80 for the recidivism prediction tasks and $2.00 for the forest cover prediction tasks[4]. To incentivize participants to carefully read about the model's explanation in each task and adjust their behavior accordingly, we further provided them with additional performance-contingent bonuses—if the overall accuracy of the participant's final predictions on the 32 tasks was at least 60%, she can earn a bonus of $0.03 for each of her correct final predictions; and for each correct answer the participant submitted to a multiple-choice question about the model behavior in the exit survey, she could also earn a $0.10 bonus. The maximum amount of bonuses a participant could earn in this study was $2.26." "[4]The base payment for the forest cover prediction tasks was higher because participants spent more time on them due to the addition of training tasks."	None	Incentives to encourage careful consideration Paying for training tasks	https://mingyin.org/paper/IUI-21/iui21.pdf	RQ1: How do different types of explanation impact people's understandings of an AI model? RQ2: How do dierent types of explanation inuence people's capability of dierentiating a model's high confidence predictions from the low condence ones? RQ3: How do dierent types of explanation change people's ability of calibrating their trust in an AI model?
Human reliance on machine learning models when performance feedback is limited: Heurstics and risks	Zhuoran Lu, Ming Yin	CHI	2021	Speed dating (Stakes: seem low)	IA: Dating app users/developers Participants: MTurk (country: US, only once)	Instructions, qualification questions, attention check Task: Sequence of 30 decision making tasks - perform 20 instances then reflect then 10 instances Mid and post task questions	Base pay: $1.5 per participant Main task: (Random) performance bonus, $1 if a randomly selected final prediction was correct Time taken: not mentioned	"The base payment of our HIT was $1.5. To motivate subjects to carefully consider whether and how much to rely on the ML model when making their predictions, we also provided a performance-based bonus to subjects: after the subject completed the HIT, we randomly selected one prediction task in the sequence to check whether the subject's final prediction on that task was correct. If so, the subject would receive a $1 bonus on top of the base payment. Note that in this experiment, we never provided any feedback to subjects on the accuracy of the ML model on any of the tasks."	None	"Randomized" performance bonus	https://mingyin.org/paper/CHI-21/reliance.pdf	• When people receive no information about an ML model's performance, does the level of agreement between people and the model on tasks that people have high condence in aect people's reliance on the model? • If so, does it continue to do so after people have had the opportunity to obtain some aggregate information about the model's performance (e.g., the model's overall accuracy on a set of decisionmaking tasks) in practice? • In the real world, people may encounter both cases that they feel confident and cases that they are not confident when interacting with an ML model. How does the people's own confidence in those cases that they agree or disagree with the model change their reliance on the model?
Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors	Niels van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, Mikael B. Skov	CHI	2021	Recidivism prediction Loan approval (Stakes: high)	IA: Judges/jury, loan approval officers Participants: Prolific Academic (80 participants, acceptance rate >=95%, US nationality)	Introduction + background explanation Task: Assess 10 'predictor'-'group filter' pairs, explore data Post task (SGL) questionnaire	Base pay: $3 per participant ($7.25 per hour) Time taken: est. 20 mins	"Participants received a predetermined amount of money for the full completion of the task. Following the US minimum wage of $7.25 per hour at the time of our study and an expected completion time of 20 minutes (as based on our pilot data), we compensated each participant with $3."	Lack of motivation as a possible reason for attention check failures: "While it is impossible to state whether the excluded participants were unable to comprehend our explanation, unmotivated to read the background text, or were perhaps automated survey completion bots, we consider all three as harmful to data collection."		https://doi.org/10.1145/3411764.3445365	Motivated by the increasing need to capture people's perspectives on AI fairness, and the lack of knowledge of the factors that affect the perceptions of fairness among non-ML experts • How does the presentation of data points afect perceived algorithmic fairness? • What is the efect of presenting / not presenting the outcome of the predicted variable on perceived algorithmic fairness? • How do the demographic factors of education and gender affect perceived algorithmic fairness?
Exploring the Effects of Machine Learning Literacy Interventions on Laypeople's Reliance on Machine Learning Models	Chun-Wei Chiang and Ming Yin	IUI	2022	Property price prediction (Stakes: seem low)	IA: Real estate evaluator? Participants: 498, MTurk (US, only once)	Pre-survey, Instructions Task: Tutorial, 20 predictions in 2 phases Exit-survey	Base pay: $1 per participant Bonus: Performance-based on average percentage error (APE), APE < 10%: $0.15, 10% ≤ APE < 20%: $0.10, 20% ≤ APE < 30%, $0.05, max. possible $3 Resulting pay: avg. $1.8 ($10.7 per hour) Time spent: avg. 10 mins	"The base payment of our HIT was $1.0. To motivate subjects to make accurate predictions and carefully consider how much to rely on the ML model in their predictions, we further provided performance-based bonus opportunities to subjects—In Phase 2, if the average percentage error (APE) of a subject's initial prediction was less than 30%, she could earn extra bonuses (APE < 10%: $0.15, 10% ≤ APE < 20%: $0.10, 20% ≤ APE < 30%, $0.05). The same bonus rule also applied to the subject's final prediction in each Phase 2 task. Thus, the max amount of bonuses a subject could earn in our HIT was $3.0." "On average, a subject spent 10 minutes on our HIT and was compensated $1.8, leading to an effective hourly wage of $10.7."	None		https://dl.acm.org/doi/10.1145/3490099.3511121	"...focus on an ML-assisted decision-making setting and conduct a human-subject randomized experiment to explore how providing different types of user tutorials as the machine learning literacy interventions can influence laypeople's reliance on ML models, on both in-distribution and out-of-distribution example"
Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning	Krzysztof Z. Gajos and Lena Mamykina	IUI	2022	Nutrition (Stakes: low)	IA: Nutritionists, lay people? Participants: 251+268+221+270 (over 3 exps), LabInTheWild & MTurk	Demographic survey, instructions Task: 24 questions - AI assisted nutrition knowledge test In Exp 2, Need For Cognition survey Exit survey	LabInTheWild: Curiosity, opportunity for social comparison MTurk: Base pay: $1 per participant ($10 per hour) Time taken: med. 6 mins	"LabintheWild is a platform for conducting online experiments with unpaid participants [58]. Instead of being paid, participants are incentivised by the promise that at the end of the study they will see their own results and compare themselves to other test takers. Both curiosity and opportunities for social comparison have been shown to increase engagement of online participants [32, 42] and multiple validation studies demonstrated that data collected on LabintheWild and other similar platforms are valid and lead to the same conclusions as data collected in traditional laboratory settings [26, 31, 43, 44, 58]. While some LabintheWild studies attracted tens of thousands of participants [25, 57], experimenters have little control over the rate at which participants arrive. Thus, we supplemented recruitment with MTurk, which is also an effective choice collecting valid behavioral data [37]. We paid MTurk participants $1 (US) aiming for $10/hour (the median time to complete the study was 6 minutes)."	None	At the end: "shown their own results, the average accuracy of other test takers, and the correct responses (and explanations) for all the questions in the test.": communication "intervention targets the person's motivation to exert cognitive effort": useful?	https://doi.org/10.1145/3490099.3511138	"How do people process the information and advice they receive from AI, and do they engage with it deeply enough to enable learning?"
Counterfactual Explanations for Prediction and Diagnosis in XAI	Xinyue Dai, Mark T. Keane, Laurence Shalloo, Elodie Ruelle, and Ruth M.J. Byrne	AIES	2022	Grass growth on dairy farms Alien planet (Stakes: seem low)	IA: Farmers / Model evaluators? Participants: 243 (193+), Prolific (native English speakers, over 18, no experience with agriculture)1	Exp 1: Agriculture Task: 10 trials, Make predictions about model's predictions Exp 2: Alien planet Task: 10 trials, Make predictions about model diagnosis (reasoning)	Base pay: £1.75 per participant Time taken: ~12 mins	"The experiment took approximately 12 minutes to complete, and each participant was paid £1.75 UK sterling for their participation"	None	Alien planet domain chosen to make "task more difficult" by removing any scope of background knowledge	https://doi.org/10.1145/3514094.3534144	"...compared two sorts of explanations for decisions made by an AI system: counterfactual explanations about how an outcome could have been different in the past, and prefactual explanations about how it could be different in the future. We examined the effects of these alternative explanation strategies on the accuracy of users' judgments about the AI app's predictions about an outcome (inferred from information about the causes), compared to the accuracy of their judgments about the app's diagnoses of a cause (inferred from information about the outcome)."
Understanding Decision Subjects' Fairness Perceptions and Retention in Repeated Interactions with AI-Based Decision Systems	Meric Altug Gemalmaz and Ming Yin	AIES	2022	Loan approval prediction (Stakes: gamified, simulated to be high)	IA: Small business owner Participants: 809+636, MTurk (US workers, only once)	Instruction, tutorial, understanding test Task: Play game (max. 9 rounds, min 1) - decide whether to keep applying for loan at a bank (coin loss/reward system) Exit survey (demographics, fairness perceptions)	Base pay: $1.5 per participants Bonus: Outcome based, max. possible $4.8 Resulting pay: med. $3.3 ($9.9 per hour) Time taken: med. 20 mins	"After completing the exit-survey, we would reveal to the subject the amount of bonus payment she received in this game—We converted the amount left in the subject's account to her bonus payment using a rate of 500 coins to $1.5. Thus, together with the base payment of $1.5 of this HIT, the subject could earn a maximum of $4.8 from this game(7). "(7)The median value of time that subjects spent on our HIT was 20 minutes, and the median payment to subjects was $3.3, leading to an effective hourly wage of $9.9."	None	Bonus was the amount left - impacts players' willingness to pay (keep playing, cash out); also different treatments (disadvantage of receiving loan) results in possible bonus to be diff for all participants i.e. bonus is outcome-based not fully performance based (although participants can "learn" to stop applying, outcome of applying is pre-decided)	https://doi.org/10.1145/3514094.3534201	"...taking a perspective of repeated interactions—We ask that when decision subjects interact with an AI-based decision system repeatedly and can strategically respond to the system by determining whether to stay in the system, what factors will affect the decision subjects' fairness perceptions and retention in the system and how"
st	~~Yiwei Lyu, Paul Pu Liang, Zihao Deng, Ruslan Salakhutdinov, and LouisPhilippe Morency~~	~~AIES~~	~~2022~~	~~Visual (multimodal) reasoning (QnA)~~ ~~(Stakes: unclear)~~	~~IA: Researchers~~ ~~Participants: 5, recruitment details not mentioned~~	~~Task: Classify datapoints into categories~~	~~None mentioned~~	-	~~None~~		~~https://doi.org/10.1145/3514094.3534148~~	~~"...focus on advancing the state-of-the-art in interpreting multimodal models"~~ ~~"...DIME generates accurate disentangled explanations, helps users of multimodal models gain a deeper understanding of model behavior, and presents a step towards debugging and improving these models for real-world deployment."~~

Do Humans Trust Advice More if it Comes from AI?: An Analysis of Human-AI Interactions	Kailas Vodrahalli, Roxana Daneshjou, Tobias Gerstenberg, and James Zou	AIES	2022	Art period determination City identification Sarcasm detection Income prediction (Stakes: low risk)	IA: several Participants: 1100, (US, UK, Asia)	Task: 32*4, make predictions with and without (human/AI) advice	Base pay: $10 per hour (resulting: $1.67 per participant) Bonus: Performance-based (calculated for both initial and final decision), bonus percentage calc. as: 0.3 * avg. performance over all tasks (performance for a task is +1 when correct, -1 when incorrect) awared when avg. performance >=0.3 (makes bonus range possible $0.9 to $3) Time take: ~10 mins	"Participants were incentivized to perform well on the task by receiving a bonus based on their averaged judgments across all task instances (both before and after advice). Participants were informed of how the bonus was calculated in the instructions." "Participants were compensated at a rate of $10.00 per hour as per the Prolific recommended rates. The survey was estimated to take 10 minutes based on several trial runs by the authors. Participants were compensated with this assumption." "Participants were also informed that they could receive up to a 30% bonus. This bonus was calculated as bonus = (0 if S < 0.3 S 0.3 otherwise, where S is the average performance of the participant across all tasks, both before and after receiving advice. Performance for a single task is computed as S * sign (correct response) · response1 where −1 ≤ response1 ≤ 1. Note that this performance metric penalizes incorrect responses." "The total cost of running all of our experiments (including the participants we used to calibrate the advice) was around $2, 500."	In justifying concerns around reinforcing biases: Participants exhibiting a certain kind of (gender/racial) bias would (implicitly) be penalized with low bous (through balanced dataset) "The census dataset may raise concerns of reinforcing gender or race-related biases in participants. However, as income levels were balanced across race and gender, biased participants would actually receive negative feedback as they would necessarily receive a low bonus payment due to their bias."	No feedback on performance during experiment	https://doi.org/10.1145/3514094.3534150	"...characterize how humans use AI suggestions relative to equivalent suggestions from a group of peer humans across several experimental settings."
Supporting Serendipitous Discovery and Balanced Analysis of Online Product Reviews with Interaction-Driven Metrics and Bias-Mitigating Suggestions	Mahmood Jasim, Christopher Collins, Ali Sarvghad, and Narges Mahyar	CHI	2022	Product recommendation / purchase decision (Stakes: unclear)	IA: Online buyers Participants: 100 (25*4), MTurk (North America, Master workers)	Pre-study questionnaire, tutorial Task: Explore (880 available) reviews (using model) and make purchase recommendation Post-study questionnaire	Base pay: $15 per participant Time taken: 19 +- 11 mins for one treatment (max.), rest not mentioned	"Each participant was compensated with USD $15."	None	Attention checks to decide whether to compensate	https://doi.org/10.1145/3491102.3517649	"...investigated interventions that are intended to support serendipitous discovery and analysis of product reviews to help readers to explore reviews more comprehensively in a balanced way, prior to making purchase decisions"
Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation	Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan	CHI	2022	Toxicity detection (Stakes: "relatively" low)	IA: Content moderators Participants: 240, MTurk (at least 1000 HITs, approval rate 99%, US residents, adult content qualifcation)	Introduction, tutorial Task: Create keyword-based rules (at least 10) to identify toxic comments (with/without AI assistance) Exit survey	Base pay: not mentioned Bonus: Performance-based, +-$0.10 for every 100 comments correctly/incorrectly reported by their created rules (possible range: $0 to $2) Resulting pay: avg. $11.80 per hour Time taken: ~10 mins	"Reward. To motivate quality work, in addition to a base payment, we design a bonus incentive as follows: participants will be awarded $0.10 for every 100 toxic comment their rules correctly reported, and penalize them $0.10 for every 100 nontoxic comment their rules mistakenly reported (lower bounded by $0 and upper bounded by $2). This bonus thus rewards both precision–how likely comments under the rule (for the manual condition) or conditional delegation with the rule are correctly classifed as toxic, and coverage–the quantity of comments covered by the rule. To make the calculation easy to understand for participants, this reward makes a simplifed assumption that the cost of wrongly reported non-toxic comments (false positive) equals the benefit of correctly reported toxic comments (true positive). This reward mechanism is explained to participants, and we include one question in attention check to ensure they understood it. We also explicitly suggest that, to optimize for the reward, their goal should be to come up with keywords that meet the following criteria: (1) that occur in a lot of comments; (2) with which the model makes accurate predictions on the comments, and (3) that are a diverse set so they may cover diferent kinds of toxic comments." "Participants were paid an average wage of $11.80 per hour"	Defined reward as a metric and used this metric as the performance incentive: "Reward (number of reported toxic comments - number of reported non-toxic comments), a measure combining precision and coverage "A user could achieve a quite high reward (and outperform the model) simply by reporting all comments with either of these two words." "This metric is highly volatile because a small number of keywords can achieve much higher rewards than others, especially out-of-distribution. We believe that precision is the more reliable measure of efficacy given that our participants tended to only choose about 10 rules."	Then based on how incentives are tied to actual task factors (like how much behaving in a certain way leads to a specific outcome), performance can be manipulated to achieve high reward?	https://doi.org/10.1145/3491102.3501999	"...develop novel interfaces to assist humans in creating conditional delegation rules and conduct a randomized experiment with two datasets to simulate in-distribution and outof-distribution scenarios. Our study demonstrates the promise of conditional delegation in improving model performance and provides insights into design for this novel paradigm, including the effect of AI explanations."
For What It's Worth: Humans Overwrite Their Economic Self-interest to Avoid Bargaining With AI Systems	Alexander Erlei, Richeek Das, Lukas Meub, Avishek Anand, and Ujwal Gadiraju	CHI	2022	Ultimatum bargaining game (Stakes: ?)	IA: Negotiator in economic setting? Participants: 280, Prolific (English-speaking, approval rate >= 90%	Attention check, questionnaire, comprehension test Task: Responders - indicate beliefs for 18 scenarios, choose proposer type, indicate minimum allocation; Proposers: indicate 18 beliefs, learn about their type, choose offer Post-task questionnaire	Base pay: £1 per participant Bonus: Outcome-based - in-game rewards (1 Coin = 0.01£); Max. possible £7.5 pp Resulting pay: avg. £4.9 pp (17.3£/h) (Proposer) and £4.8 pp (14.4£/h) (Responder) Time taken: avg. 17 mins (Proposer), avg. 20 mins (Responder)	"Throughout the experiment, we used "Coins" as an experimental currency unit. Coins were later converted into Pound sterling, with a conversion rate of 1 Coin = 0.01£. Subjects earned a base payment of 1£ with a maximum bonus payment of additional 7.5£. The average proposer earned 4.9£ ($6.8) in about 17 minutes (17.3£/h), the average responder earned 4.8£ ($6.7) in about 20 minutes (14.4£/h)."	Study design: Economic expectations through incentivizing responder beliefs "we map these preferences to economic expectations by eliciting incentivized responders beliefs about the behavior of each proposer type" Results: Incentives not enough to manage algorithmic systems "the introduction of algorithmic systems cannot simply be managed through monetary incentive schemes or the promise of efficiency gains." Potential reasoning behind a statement: Non-incentivized work (answering an open-ended question) considered to not have received enough thought by workers: "There are two important caveats to the analysis. First, it relies on non-incentivized self-reported data near the end of the experiment. Thus, we cannot verify that subjects reflected carefully on their answers..." "our results also concern the design of incentive schemes and market mechanism": generalizability not claimed	Incentives used to emulate "economic self-interest" Suggests that incentive schemes are not enough to manage the introduction of algorithmic systems Supposes that non-incentivized tasks possibly do not receive enough careful thought from workers	https://doi.org/10.1145/3491102.3517734	"...analyze whether economic expectations can explain differences between human-AI and humanhuman interactions."
You Complete Me: Human-AI Teams and Complementary Expertise	Qiaoning Zhang, Matthew L. Lee, and Scott Carter	CHI	2022	Object identification (Stakes: seem low)	IA: (artificial task) Participants: MTurk , 178+395 (approval rating >= 95%, min. 1,000 approved tasks)	S1: Instructions and scoring scheme, training trials Task: 42 trials, identify a shape (initial and final decision, with/without AI advice), outcome feedback Rate trust & reliance, demographic survey S2: Similar, with diff. typesof explanations	Base pay: $1.50 per participant Bonus: Performance-based on correct initial and final answers, $0.02 per point (max.possible: 42*2*0.02); resulting bonus: med. $1.20 (S1) / avg. $1.26 (S2) Resulting pay: $14.57 per hour (S1), $11.78 per hour (S2) Time spent: med. 11.12 mins (S1), 14.06 mins (S2)	S1: "For participating in the study, each respondent received a payment consisting of a base rate of $1.50 and a performance-based bonus payment, which depended on the total points earned (1 point=$0.02) for correctly identifying shapes. Participants received a median bonus of $1.20 and took a median of 11.12 minutes to complete the task, for a median hourly rate of $14.57." S2: "Participants were paid a base rate ($1.50) and a performance bonus payment based on the final points earned (1 point =$0.02). Participants received an average bonus of $1.26 and took a median of 14.06 minutes to complete the task, for a median hourly rate of $11.78."	None		https://doi.org/10.1145/3491102.3517791	"...investigate how people trust and rely on an AI assistant that performs with different levels of expertise relative to the person, ranging from completely overlapping expertise to perfectly complementary expertise."
Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making	Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein	CHI	2022	Civil protection / Air defense (Stakes: high)	IA: Drone operator Participants: Prolific, 428	Tutorial, comprehension questions Task: Play game, 2 missions (4 decision problems each) - interact with human/AI expert and make decision (maximize utility / minimize risk) Post-task questionnaire	Base pay: £3.75 per participant Time taken: avg. 31 mins	"Each participant was paid GBP 3.75 for completing our survey. On average, people took 31 minutes to participate."	None	Simulating ethical decision making and high stakes: through gamification, incentives could matter	https://doi.org/10.1145/3491102.3517732	"...investigate how the expert type (human vs. AI) and level of expert autonomy (adviser vs. decider) influence trust, perceived responsibility, and reliance"
AI-Moderated Decision-Making: Capturing and Balancing Anchoring Bias in Sequential Decision Tasks	Jessica Maria Echterhof, Matin Yarmand, and Julian McAuley	CHI	2022	Purchase decisions (Stakes: low)	IA: Product buyers Participants: MTurk	Task: Decide whether or not to buy a product based on description, several sequential trials	None mentioned	-	None	Crowdworkers do not perform an AI-assisted decision-making task, only decide based on a decision (crowdsourcing used for data collection); exclude?	https://doi.org/10.1145/3491102.3517443	"...investigate a specific type of anchoring bias, in which decision-makers are anchored by their own recent decisions" "...propose an algorithm that identifes existing anchored decisions, reduces sequential dependencies to previous decisions, and mitigates decision inaccuracies post-hoc"
Not Just a Preference: Reducing Biased Decision-making on Dating Websites	Zilin Ma and Krzysztof Z. Gajos	CHI	2022	Dating simulation (Stakes: unclear)	IA: Users of dating apps Participants: LabInTheWild, 1907	Demographic survey, check-in questions Task: Dating simulation - match (3 profiles shown) Check-out questions, debrief	No monetary compensation; Feedback for performance (LabInTheWild)	"The study was conducted on LabintheWild.org [80], a crowd-sourcing platform where participants voluntarily access the study in exchange for feedback on how they performed in the study. Participants on LabintheWild do not receive monetary compensation."	None	[80] Katharina Reinecke and Krzysztof Z. Gajos. 2015. LabintheWild: Conducting Large-Scale Online Experiments With Uncompensated Samples.	https://doi.org/10.1145/3491102.3517587	Study the effect of certain design choices for dating websites (like swiping to like, viewing match scores) on amplifying racial biases
Will AI Console Me when I Lose my Pet? Understanding Perceptions of AI-Mediated Email Writing	Yihe Liu, Anushk Mittal, Diyi Yang, and Amy Bruckman	CHI	2022	E-mail writing (Stakes: unclear)	IA: E-mail writers? Participants: MTurk, 229 (approval ratings > 90%, Master-Mturkers.)	Task: Read 12 emails written by human/AI/both and answer survey questions Closing survey, attention checks	Base pay: US minimum wage (amount not mentioned) Time taken: 30 mins	"To provide fair compensation to our participants, Mturkers were offered an equivalent of United States federal minimum wage in terms of the time expected to complete the survey"	None		https://doi.org/10.1145/3491102.3517731	"...investigates people's perceptions of AI-mediated communication in the context of writing emails"
Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems	Cecilia Panigutti, Andrea Beretta, Dino Pedreschi, and Fosca Giannotti	CHI	2022	Disease prediction (Stakes: high)	Participants (& IA): Healthcare providers recruited from Prolific (pre-screened), high approval & submission rate	Task: Estimate chances of a specific disease occurring in a patient, with AI advice (with/without explanation)	Base pay: 6.20£ per participant Time taken: not mentioned	"Each participant was asked to perform a task (detailed below) and answer a set of questionnaires and received a compensation of 6.20£ for it."	None	Participants crowdworkers but screened to be healthcare providers	https://doi.org/10.1145/3491102.3502104	"[investigate] impact of receiving an explanation for an algorithmic suggestion in the healthcare context."
"There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making	Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski	FAccT	2022	Loan approval (Stakes: high)	IA: Loan applicant Participants: Prolific	Task: Read loan application decision and answer questionnaire	Base pay: >$6.50 per hour (exact unclear) Time taken: not mentioned	"SPs were monetarily compensated above the recommended min. pay of $6.50 per hour."	None		https://doi.org/10.1145/3531146.3533218	"...conduct a human subject study to assess people's perceptions of informational fairness (i.e., whether people think they are given adequate information on and explanation of the process and its outcomes) and trustworthiness of an underlying ADS when provided with varying types of information about the system"
It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy	Andrew Bell, Ian René Solano-Kamaiko, Oded Nov, and Julia Stoyanovich	FAccT	2022	Student performance estimation Property price prediction (Stakes: mentioned as limitation - no "real" stakes for users)	Participants (& IA): In education / real estate recruited from Prolific (pre-screened), 168*2 (US, over 18, fluent English)	Task: interact with 2 ML models, predict system output, predict most important feature, system understanding questionnaire Post-task background survey	Base pay: $15 per hour Resulting: avg. $28.73/hr (education); avg. $27.44/hr (housing) (unclear how resulting reward is higher than base pay - no bonus scheme mentioned) Time taken: 8m50s (education), 10m14s (housing)	"Participant compensation was set at $15/hour; participants were estimated to complete the study in 15 minutes or less (the education study average reward was $28.73/hr and the housing study average reward was $27.44/hr)."	Mentioned that no "real" stakes for users (could be simulated through incentives?) "The methods developed in this paper have only been tested in a lab environment where there are no real stakes for users, which may impact their robustness or validity (e.g., the teachers from our survey did not have a connection with a real student for which the AI system was making a prediction)."		https://doi.org/10.1145/3531146.3533090	"...study the trade-off between accuracy and explainability for the end users of black-box and interpretable models in public policy use-cases and seek to answer two related research questions: (1) how can we quantify explainability? and (2) how can we quantify the trade-off between accuracy and explainability?"
What People Think AI Should Infer From Faces	Severin Engelmann, Chiara Ullstein, Orestis Papakyriakopoulos, and Jens Grossklags	FAccT	2022	Advertising (Stakes: low) Hiring (Stakes: high)	IA: Moral judges(?) Participants: 3745, MTurk (approval rating > 95%)	Task: 8 tasks, rate agreement with AI decision on an "inference" (judgements about social media users/job applicant based on a photo), written justification	Base pay: minimum wage (amount not mentioned) Time taken: avg. 10.4 mins	"Following recommended principles of ethical crowdsourced research [104], we first ran a pre-study with 120 Turkers to determine the average time it would take to complete the survey and used this reference time to determine a payout above the US minimum wage (mean=8.03 min). In our study (N = 3745), the mean was 10.4 min (min = 3.35 min, max = 31.55 min)."	None	[104] Vanessa Williamson. 2016. On the ethics of crowdsourced research.	https://doi.org/10.1145/3531146.3533080	"...understanding how "non-experts" in AI ethically evaluate facial AI inference-making"

Taking Advice from (Dis)Similar Machines:The Impact of Human-Machine Similarity on Machine-Assisted Decision-Making	Nina Grgić-Hlača, Claude Castelluccia, & Krishna P. Gummadi	HCOMP	2022	Dating preference prediction Recidivism prediction Age estimation (Stakes: varying)	IA: several, task specific Participants: 901, Prolific (approval rate >= 95%, completed >= 100 studies)	Exp 1: Prediction without AI assistance Exp 2: Task: 25+25 instances, initial prediction, observe machine advice, final prediction; first test phase with outcome feedback then prediction phase without outcome feedback Questionnaire, attention check	Base pay: £2 (Exp1), £2.5 (Exp2) Bonus: Performance-based, +-$0.1 bonus per correct/incorrect prediction (Exp2) Resulting pay: $11 / hour (Exp 1), $9.3 / hour (Exp 2) Time taken: avg. 14 mins (Exp 1), avg. 21 mins (Exp 2)	"To incentivize respondents to put effort into building a mentalmodel of the machine's predictions in the test-drive phase of the experiment, we informed them that they could earn monetary rewards in the prediction phase." "Following the approach of Dietvorst, Simmons, and Massey (2015), we used monetary incentives only in the prediction phase." "For each correct prediction, we rewarded respondents with a $0.10 bonus, and penalized them the same amount for each incorrect prediction. Similar financial incentives have been shown to encourage respondents to provide accurate responses (Chittilappilly, Chen, and Amer-Yahia 2016; Harris 2011)" Exp 1: "The average completion time for this set of surveys was 14 minutes, and respondents were paid a base fee of £2 for taking part in this experiment (i.e., slightly above $11 per hour)" Exp 2: " On average, participants took21 minutes to complete the survey. Respondents were paida base fee of £2.5 for taking part in the experiment (i.e.,slightly above $9.3 per hour). Additionally, the respondentscould earn bonus payments based on their performance, asdescribed in Figure 7b in the SM"	Mentioned as possible reasoning for an outcome: "...people were more likely to take machine advice in the second experimental phase(5)" "(5)This difference could be caused by the introduction of monetary incentives and removal of feedback about performance in the second phase, or other factors, such as learning or fatigue effects."	Dietvorst, Simmons, and Massey (2015) (Chittilappilly, Chen, and Amer-Yahia 2016; Harris 2011) SM: incentives and bonus structure clearly communicated	https://arxiv.org/pdf/2209.03821.pdf	"In a series of human-subject experiments with a total of 901 participants, we study how the similarity of human and ma-chine errors influences human perceptions of and interactions with algorithmic decision aids"
Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making	Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett	CSCW	2022	Student performance estimation (Stakes: simulated high (through incentives))	IA: Educators Participants: 47 + 479, MTurk (US, approval rating >= 98%, min. 100 approved tasks.)	Training: 15 trials with outcome feedback Task: 36 (exp 1), 40 (exp 2) trials, make predictions with/without AI advice in diff. settings	Base pay: $3.5 (exp 1), $4.125 (exp 2) per participant Bonus: Performance-based (called as outcome based in paper), $1 for accuracy (unclear how calculcated) Resulting pay: $4.5 ($10 / hour) (exp 1), $5.125 ($10.25 / hour) (exp 2) Time taken: avg. 27 mins (exp 1), avg. 30 mins (exp 2)	Exp 1: "The average completion time for this user study was 27 minutes, and each participant received compensation of $4.5 (roughly equals an hourly wage of $10). The participants received a base pay of $3.5 and a bonus of $1 (to incentivize accuracy)." Exp 2: "The average completion time for this user study was 30 minutes, and participants received compensation of $5.125 on average (roughly equals an hourly wage of $10.25). The participants received an average base pay of $4.125 and bonus of $1 (to incentivize accuracy)."	Incentives used to increase stakes: "We used a non-critical decision-making task where the participants would not be held responsible for the consequences of their decisions. This problem was mitigated by introducing an outcome-based bonus reward which motivates optimal decision-making."	Bonus for incentivizing accuracy Incentives used to motivate optimality; attach consequences to decision-making -> increase stakes, basically Controlling for time given to perform tasks: then paying for time taken: fair?	https://doi.org/10.1145/3512930	"...focus on anchoring bias and the associated anchoring-and-adjustment heuristic that is important towards optimizing team performance. We validate the use of time as an effective strategy for mitigating anchoring bias through a user study. Furthermore, through a time-based resource allocation formulation, we provide an optimal allocation strategy that attempts to achieve the "best of both worlds" by capitalizing on the complementary knowledge presented by the decision-maker and the AI model"
Categorical and Continuous Features in Counterfactual Explanations of AI Systems	Greta Warren, Ruth M.J. Byrne, and Mark T. Keane	IUI	2023	Legal driving alcohol limit prediction (Stakes: high)	IA: ? Participants: 127+211, Prolific (native English speakers)	Training + test phase, 40 or 16 trials (exp or exp2) each Task: Predict whether someone is over the legal blood alcohol (BAC) limit to drive (model predicts BAC) with AI assistance (diff. kinds of explanations & features). Questionnaire, attention checks	Base pay: £2.61 per participant Time taken: ~28 mins	"Participants were paid £2.61 for their time. The experiment took approximately 28 minutes to complete."	None		https://doi.org/10.1145/3581641.3584090	"...carried out two user studies to (i) test a fundamental distinction in feature-types, between categorical and continuous features, and (ii) compare the relative effectiveness of counterfactual and causal explanations" "The studies used a simulated, automated decisionmaking app that determined safe driving limits after drinking alcohol, based on predicted blood alcohol content, and user responses were measured objectively (users' predictive accuracy) and subjectively (users' satisfaction and trust judgments)"
Subgoal-Based Explanations for Unreliable Intelligent Decision Support Systems	Devleena Das, Been Kim, and Sonia Chernova	IUI	2023	Restaurant (kitchen) planning (Stakes: seem high)	IA: Chefs ParticipantsL 120, MTurk	Tutorial Task: Gamified- 5 games, prepare meals with AI assistance, deliver meals on time and identify suboptimal AI suggestions	Base pay: $5 per participant Time taken: avg. 40 mins	"The task took on average 40 minutes and participants were compensated $5.00."	None		https://doi.org/10.1145/3581641.3584055	"...examine novice user interactions with a non-robust IDS system – one that occasionally recommends suboptimal actions, and one that may become unavailable after users have become accustomed to its guidance." "...introduce a new explanation type, subgoal-based explanations, for plan-based IDS systems, that supplements traditional IDS output with information about the subgoal toward which the recommended action would contribute."
Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction	Patrick Hemmer, Monika Westphal, Max Schemmer, Sebastian Vetter, Michael Vössing, and Gerhard Satzger	IUI	2023	Image classification (Stakes: varying, here seem low)	IA: ? Participants: 196, Prolific Academic	Unrelated "test" task , attention check, familiarization tasks Task: 20, Classify images (with/without AI assistance) Follow-up questions	Base pay: $1.5 per participant Time taken: ~10 mins	"Participants received $1.5 for their participation in the task that took approximately 10 minutes."	None		https://doi.org/10.1145/3581641.3584052	"...investigate how and why AI delegation affects human task performance and task satisfaction."
It Seems Smart, but It Acts Stupid: Development of Trust in AI Advice in a Repeated Legal Decision-Making Task	Patricia K. Kahr, Gerrit Rooks, Martijn C. Willemsen, and Chris C. P. Snijders	IUI	2023	Jail time estimation (Stakes: high)	IA: Member of law enforcement Participants: 171, Prolific (British citizenship, 18+, experience in the field of law (half of all participants)	Task: 20 trials, predict jail time with AI assistance and explanations (varying accuracy, type), initial and final decision with outcome feedback Questionnaire	Base pay: avg. £5.87 per participant Time taken: ~20 mins	"Each participant received on average £5.87 as compensation."	None		https://doi.org/10.1145/3581641.3584058	"...focus on how trust develops over time in a humanAI-interaction scenario. In a 2x2 between-subject experiment, we test how model accuracy (high vs. low) and type of explanation (human-like vs. not) affect trust in AI over time"
How does Value Similarity affect Human Reliance in AI-Assisted Ethical Decision Making?	Saumik Narayanan, Guanghui Yu, Chien-Ju Ho, and Ming Yin	AIES	2023	Medical resource (here: kidney) allocation (Stakes: high)	IA: Medical practitioners / policy makers Participants: 303, MTurk	Task: Stage 1: 9 scenarios - express and assess own and AI's ethical preferences; Stage 2: Decision-making: 18 scenarios - resolve dilemma with AI assistance	Base pay: ~$10 per hour	"Median pay for workers was approximately $10 per hour"	Mentioned as a possible solution for overcoming limitation of using crowdsourcing in general but no further comment made: "The common approaches to improve the quality of crowdsourced data collection include...designing proper incentives [21, 22, 24, 30, 44]"	Check out references: [21, 22, 24, 30, 44]	https://doi.org/10.1145/3600211.3604709	"...explores the impact of value similarity between humans and AI on human reliance in the context of AI-assisted ethical decision-making."
How do you feel? Measuring User-Perceived Value for Rejecting Machine Decisions in Hate Speech Detection	Philippe Lammerts, Philip Lippmann, Yen-Chia Hsu, Fabio Casati, and Jie Yang	AIES	2023	Hate speech detection (Stakes: high)	IA: Users of SM platform / Content moderators Participants: 160, Prolific (18+, fluent in English, approval rating > 90%), experience using SM	Warm-up task, manipulation checks Task: 40 tasks, classify instance, indicate agreement/disagreement with AI decision	Base pay: £9 per hour Time taken: not mentioned	"Every participant is paid an hourly wage of 9 GBP, exceeding the UK minimum wage at the time of the study."	None		https://doi.org/10.1145/3600211.3604655	"...propose a value-sensitive rejection mechanism that automatically rejects machine decisions for human moderation based on users' value perceptions regarding machine decisions."
Comparing Sentence-Level Suggestions to Message-Level Suggestions in AI-Mediated Communication	Liye Fu, Benjamin Newman, Maurice Jakesch, and Sarah Kreps	CHI	2023	E-mail writing (Stakes: unclear)	IA: S1: Staffers from legislators' offices S2: E-mail recipients Participants: S1: 120, Prolific (US, fluent in English, listed "politics" as a hobby) S2: 1000	S1: Task: Respond to e-mails with/without (varying) AI suggestions Answer questions S2: Subjective study, evaluate e-mail responses by answering questions	S1: Base pay: $5 per participant Time taken: est. 20 mins	"We pay $5.00 for each task session based on an estimated completion time of 20 minutes."	None	S2 pay not mentioned	https://doi.org/10.1145/3544548.3581351	"...explores the trade-offs between sentence vs. message-level suggestions for AI-mediated communication"
Exploring the Use of Personalized AI for Identifying Misinformation on Social Media	Farnaz Jahanbakhsh, Yannis Katsis, Dakuo Wang, Lucian Popa, and Michael Muller	CHI	2023	Misinformation prediction (Stakes: high)	IA: Social media users Participants: 61, MTurk (> 500 HITs approved, approval rate > 98%, US citizen/resident, 18+, at least occasionally read news online, fluent in English,	Task: 78 tasks, interact with feed, assess tweet with/without AI assistance Answer questions	Base pay: $17 per participant Time taken: ~1 hour	"From our pilot studies with our research group, we determined that the average time for completing the task was approximately an hour. Therefore, we set a compensation of $17 for the task."	None		https://doi.org/10.1145/3544548.3581219	"explore how human assessments and AI predictions can be combined to identify misinformation on social media"
Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems	Gaole He, Lucie Kuiper, and Ujwal Gadiraju	CHI	2023	Logical reasoning (Stakes: unclear)	IA: (Reasoning) test taker? Participants: 6+249 (Prolific)	Task: 16 trials, Initial and final decision, with/without tutorial and XAI	Main study: Base pay: £2.5 per participant (£7.5 per hour) Bonus: Performance-based, £0.1 per correct decision Time taken: est. 20 mins Pilot: Base pay: £7.5 per hour Bonus: Performance-based, £0.05 per correct decision Time taken: est. 33 mins Follow-up study: Base pay: £1.5 per participant (£9 per hour) Bonus: Performance-based, £0.1 per correct decision Time taken: est. 10 mins	Pilot: "All participants were rewarded with hourly wage of £7.5 (estimated completion time was 33 minutes), and extra bonus of £0.05 for each correct decision." Main: Compensation. All participants were rewarded with £2.5, amounting to an hourly wage of £7.5 (estimated completion time was 20 minutes). We rewarded participants with extra bonuses of £0.1 for every correct decision in the 16 trial cases. By incentivizing participants to reach a correct decision, we operationalize the concomitant "vulnerability" discussed by Lee and See [47] as a contextual requirement to encourage appropriate system reliance."	Limitations: Monetary compensation introduces self-interest bias "Self-interest bias is possible, because crowd workers we recruited from the Prolifc platform are motivated by monetary compensation."	Incentives as a way of operationalizing "vulnerability" to encourage appropriate reliance Incentives can introduce self-interest bias [15] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. 2021. A checklist to combat cognitive biases in crowdsourcing.	https://doi.org/10.1145/3544548.3581025	"addresses an under-explored problem of whether the Dunning-Kruger Efect (DKE) among people can hinder their appropriate reliance on AI systems"
"Should I Follow the Human, or Follow the Robot?" — Robots in Power Can Have More Influence Than Humans on Decision-Making	Yoyo Tsung-Yu Hou, Wen-Ying Lee, and Malte F Jung	CHI	2023	Client consultancy (Stakes: seem high)	IA: Member of a consulting team Participants: 120, MTurk (approval rate > 99%, US)	Instructions Task: 10 client consultation questions, initial + fianl decision, with human & AI in team Post-task survey, attention checks, debrief	Base pay: $5.5 per participant Bonus: Fixed, $2 per participant Time taken: not mentioned	"All participants were paid $7.50, including the base value $5.50 plus a $2.00 bonus payment, which was for power manipulation and will be explained in the following section." "They were also told that they had the chance to win a bonus payment base on their team performance: the more their team answer was like professional consultants' answer, the more they could win." Human/Robot leader: "The participants were informed that the leaders were chosen as the team leaders, they decided the final answer for the teams, and they also decided how to split the bonus payment after the game." All equal: "They were also told in advance that they would decide on bonus allocation together." "At the end of this study, the participants were given the debriefing that there were actually no other teammates, and they would get the maximum possible $2.00 bonus payment."	In study design: Control of bonus allocation as a tool for "power manipulation" among peers, reward power: "and the right to allocate monetary bonus (to gain reward power, the power to control desirable resources)." "For h-lead and r-lead, the participants were also informed about who (the robot or the human) was selected as the leader, as well as the team leader's responsibility (deciding the team answer and the bonus allocation), and thus were exposed to the manipulation of power." To encourage a certain kind of behavior (here: not giving extreme answer): "To prevent participants from giving extreme answers as their initial suggestions and made the symmetry impossible, we let participants know beforehand that "None of the correct answers from the professional consultants are extreme. So submitting a very low or very high answer as your suggestion is probably not a good idea to win the bonus payment."	Right to control bonus allocation as reward power: manipulation of power; responsibility; Incentives used to encoourage a certain kind of behaviour Communication: Bonus payout communicated as outcome based but in actuality fixed bonus paid to everyone Participants also asked how they would allocate bonus, but only as part of power manipulation but not included in paper focus	https://doi.org/10.1145/3544548.3581066	"investigate the efect of power on people's perception and behavior in situations where they interact with humans and robots simultaneously"
Disentangling Fairness Perceptions in Algorithmic Decision-Making: the Effects of Explanations, Human Oversight, and Contestability	Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, Alessandro Bozzon	CHI	2023	Loan approval (Stakes: varying: low & high)	IA: Fairness assessor? Participants: 267, Prolific (18+, fluent English)	Task: 1, Receive loan approval scenario, model explanations, and answer questions Questionnaire	Base pay: $12 per hour Time taken: med. 7m41s	"Participants were rewarded based on a $12 hourly rate and the median completion time was 7 minutes and 41 seconds"	None	No "performing a task"; exclude?	https://doi.org/10.1145/3544548.3581161	"...a user study (N = 267) investigating the individual and combined efects of explanations, human oversight, and contestability on informational and procedural fairness perceptions for high- and low-stakes decisions in a loan approval scenario"

Are Two Heads Better Than One in AI-Assisted Decision Making? Comparing the Behavior and Performance of Groups and Individuals in Human-AI Collaborative Recidivism Risk Assessment	Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin	CHI	2023	Recidivism prediction (Stakes: high)	IA: Judges / Jury Participants: MTurk (US)	Phase 1: Pre-task survey Task: 9 tasks, predict recidivism Exit survey Phase 2: Task: 15 (9 practice + 6 formal) tasks, Predict risk of reoffense with AI assistance; For formal tasks Iindividually or group; use chatroom to discuss for group setting, answer questions	Phase 2: Base pay: $1 per participant Bonus: Performance-based, $0.4 per correct final prediction of formal tasks, max. possible $2.4 Time taken: not mentioned	"The base payment was $0.3 for Phase 1 and $1.0 for Phase 2. In addition, to motivate subjects to carefully deliberate (and discuss with other members in their group if applicable) about what predictions to make in the formal task, we further informed each subject at the beginning of the Phase 2 HIT that they could earn a $0.4 bonus for each correct fnal prediction made on the formal task. Thus, the maximum amount of bonuses a subject could receive in Phase 2 was $2.4."	None	Bonus for "motivation"	https://doi.org/10.1145/3544548.3581015	"...conduct a case study to compare groups and individuals in human-AI collaborative recidivism risk assessment along six aspects, including decision accuracy and confdence, appropriateness of reliance on AI, understanding of AI, decision-making fairness, and willingness to take accountability"
Comparing Zealous and Restrained AI Recommendations in a Real-World Human-AI Collaboration Task	Chengyuan Xu, Kuo-Chin Lien, and Tobias Höllerer	CHI	2023	Video anonymization (Stakes: high)	Participants (& IA): Professional annotators, 78, "in-house"	Task: 24+12 videos, Annotate videos with/without AI assistance	Employees' usual salary	"they are paid at their regular hourly rate, so participants are not motivated by compensation to work faster"	Reasoning for finding: Due to a lack of performance based incentives, people not be motivated to work harder "veteran workers are overall signifcantly slower than novice workers...if we consider how people are paid, this result would be a reasonable optimization given the incentives – veteran workers know the acceptable work pace, so they do not need to work faster than necessary" "when paid at a flat hourly rate, people are not necessarily motivated to work faster. When lacking a quality-based performance evaluation mechanism, people are not necessarily motivated to push for "better-than-sufficient" quality" Limitations: System can be helpful given the right incentives "Incentives for users to actively perform better. We discussed in Section 6.3 observations that methods with bette performances are not necessarily favored by the users. I.e., the users were involuntarily pushed to have higher performance by their AI teammates. From a system designer's perspective, the AI teammate should help users to voluntarily perform better given the right incentives."	Suggests that performance based incentives necessary to encourage higher quality of work and adoption of (high performing) systems	https://doi.org/10.1145/3544548.3581282	"...investigate a real-world video anonymization task for which recall is paramount and more costly to improve. We analyze the performance of 78 professional annotators working with a) no AI assistance, b) a high-precision "restrained" AI, and c) a high-recall "zealous" AI"
Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations	Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes	CHI	2023	Logical reasoning (Stakes: unclear)	IA: Users of AI systems? Participants: 204, Prolific	Demographic survey Task: 10 tasks, Determine logical validity of statements with/without diff. kinds of AI feedback Post-task questionnaire	None mentioned	-	None		https://doi.org/10.1145/3544548.3580672	"...presents the novel idea of AI-framed Questioning that turns information relevant to the AI classifcation into questions to actively engage users' thinking and scafold their reasoning process. We conducted a study with 204 participants comparing the efects of AI-framed Questioning on a critical thinking task; discernment of logical validity of socially divisive statements"
Overcoming Algorithm Aversion: A Comparison between Process and Outcome Control	Lingwei Cheng and Alexandra Chouldechova	CHI	2023	Student performance prediction (Stakes: ?)	IA: Educators Participants: 47 + 479, MTurk (US, approval rating >= 97%, min. 1000 completed HITs.)	Attention checks Task: 20 tasks, predict reading test percentile scores between [0, 100] for high school sophomores under diff. conditions Post-survey questions	OG scheme: Base pay: $2 per participant Bonus: Performance-based, Max. possible $5 (-$1 for every 5 units of distance from correct answer on avg.) Alternative scheme (adjust-by-10): Base pay: $1 per participant Bonus: Performance-based, Max. possible $5 (-$1 for every 3 units of distance from correct answer + 10 on avg.) Study 2&3: Base pay: $4 per participant Time taken: avg. 19.6 +- 9.5 mins (across 3 studies)	"Bonus Original Scheme (Dietvorst et al) Proposed Scheme $5 within 5 points within 14 points of students' actual performance on average $4 within 10 points within 17 points of students' actual performance on average $3 within 15 points within 20 points of students' actual performance on average $2 within 20 points within 23 points of students' actual performance on average $1 within 25 points within 26 points of students' actual performance on average One-time participation fee is $2 under original scheme and $1 under proposed scheme during study 1"	Alternative bonus scheme: "constructed to yield approximately the same expected total reward as the original, but offers greater incentive to choose the model" "We find that participants are not sensitive to incentive structures. They are just as likely to use the model if the model's stated performance is sufficient to achieve a bonus payout in the middle-to-top of the bonus range vs. if it is only sufcient to achieve a minimum bonus payout, a result that is surprising when viewed in the context of loss aversion." "[One participant's response] indicates that the change in bonus scheme did not go without notice. However, the majority of the participants were not aware of the implication of the proposed bonus scheme." Design of incentives in studies: "In addition to the debatable replicability of loss aversion indicated by recent studies [54, 67], our null fnding in this case highlights the challenges of conducting such studies online and in the unique context of human-AI collaboration" "Because reallife end-users can have very diferent demographics characteristics and non-monetary incentives and operate in higher-stake environments, we cannot reliably generalize the finding to real workplaces but acknowledge that our finding has implications on designing incentives for crowdworker studies."	Incentives to promote a certain behaviour (choosing the model) (and study the effects on loss aversion?) Suggest that participants are not sensitive to incentive structures (but seems like people didn't understand the implications, then is it fair to say this?) Identify incentive construction as a challenge in conducting human-AI crowdsourcing studies	https://doi.org/10.1145/3544548.3581253	"We study whether algorithm aversion is mitigated by process control, wherein users can decide what input factors and algorithms to use in model training. We conduct a replication study of outcome control, and test novel process control study conditions on Amazon Mechanical Turk (MTurk) and Prolific"
Watch Out for Updates: Understanding the Effects of Model Explanation Updates in AI-Assisted Decision Making	Xinru Wang and Ming Yin	CHI	2023	Poisonous mushroom prediction Loan default prediction (Stakes: high)	IA: ? , loan assessment officer Participants: 475 + 394 + 412, MTurk (US)	3 exps Pre-task questionnaire, tutorial Task: 15+15 tasks (2 phases w/ diff. treatments), initial and final prediction Mid & exit questionnaire, attention checks	Exp 1: Base pay: $1.80 Bonus: Performance-based, for final predictions: if overall accuracy > 55%, $0.4 for each correct prediction, max. possible $1.20 Resulting pay: med. $11 per hour Time taken: med. 12.5 mins Exp 2: Resulting pay: med. $10.3 per hour (exp 2.1), med. $11.9 per hour (exp 2.2) Time taken: med. 12.8 mins (exp 2.1), med. 11.3 mins (exp 2.2)	Exp 1: "The base payment of the experiment was $1.80. To incentivize participants to carefully read about the model's explanation in each task and adjust their trust accordingly, we further provided them with additional performance-contingent bonuses—if the overall accuracy of a participant's fnal predictions on the 30 tasks was at least 55%, they could earn a bonus of $0.04 for each of their correct fnal predictions. Thus, the maximum amount of bonus a participant could earn in this experiment was $1.20." "The median time participants spent on the experiment was 12.5 minutes, leading to a median hourly wage of $11.00." Exp 2: "In Experiment 2.1, the median time participants spent on the experiment was 12.8 minutes, and the median hourly wage participants earned was $10.3. In Experiment 2.2, the median completion time and median hourly wage were 11.3 minutes and $11.9, respectively"	None	Incentives for motivation; Kind of unclear description	https://doi.org/10.1145/3544548.3581366	"...study how varying levels of similarity between the AI explanations before and after a model update afects people's trust in and satisfaction with the AI model"
Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making	Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma	CHI	2023	Income prediction (Stakes: "relatively" low)	IA: ? Participants: 293, Prolific	Tutorial, attention check, training examples Task: 20+20 predictions (with then without AI advice) Exit survey	Base pay: not mentioned Bonus: Performance-based, $0.50 per participant if overall accuracy > 80% Resulting pay: avg. $9.34 per hour Time taken: 20 mins	"To motivate high quality work, in addition to the base payment, we gave participants a $0.50 bonus if their overall accuracy exceeded 80%. The entire study lasted about 20 minutes. The average wage for participants was about $9.34 per hour."	None	Incentives for motivation	https://doi.org/10.1145/3544548.3581058	"In the first phase, we explore how to model humans' capability (correctness likelihood) on a given task instance. We propose a human decision-making model approximation method with an interactive decision rule modifcation interface. In the second phase, we explore how to leverage human-AI capabilities to promote appropriate trust in AI-assisted decision-making. Based on theories of people's cognitive processes, we propose three CL exploitation methods and investigate their efects on humans' trust appropriateness, task performance, and user experience."
Questioning the ability of feature-based explanations to empower non-experts in robo-advised financial decision-making	Astrid Bertrand, Winston Maxwell, and James R. Eagan	FAccT	2023	Life insurance planning (financial decision making) (Stakes: varying risk)	IA: Life insurance robo-advisor users Participants: ~30*4, Lucid (interested in life insurance)	Questionnaire Task: View life insurance plan recommendations by AI (diff. treatments) and accept/reject Post questionnaire, feedback	Base pay: ~€3.50 per participant Time taken: ~10 mins	"The whole study lasted around 10 minutes. Participants were paid around €3.50(6) for completing the study" "(6)Lucid goes through several suppliers to gather participants. Each supplier receives 3.50€ for each study completed, takes a commission and pays the rest to the participant."	None		https://doi.org/10.1145/3593013.3594053	"we carried out a qualitative study to understand what end-users and consumer protection experts—regulators—say about feature-based explanation requirements. We then presented the results of a large-scale study to investigate if different formats of feature-based explanations help novice users appropriately rely on, trust and understand recommendations of life-insurance plans."
Algorithmic Decisions, Desire for Control, and the Preference for Human Review over Algorithmic Review	Henrietta Lyons, Tim Miller, and Eduardo Velloso	FAccT	2023	Fraud detection by rideshare companies Employee performance assessment Hiring (Stakes: varying)	IA: Impacted user (context-specific: driver, employee, job applicant) Participants: 260, Prolific (US, 18+)	Demographic survey Task: Pick decision for review from human/AI reviewer under diff. conditions	Base pay: £2.47 per participant Time taken: ~15 mins	"The study was expected to take approximately 15 minutes, and participants were paid £2.47."	None		https://doi.org/10.1145/3593013.3594041	"we explore why decision subjects generally express a preference for human reviewers of algorithmic decisions over algorithmic reviewers. We theorise that decision subjects desire control over the decision-making process in order to increase their chance of receiving a favourable outcome."
Humans Forgo Reward to Instill Fairness into AI	Lauren S. Treiman, Chien-Ju Ho, Wouter Kool	HCOMP	2023	Ultimatum bargaining game (Stakes: ?)	IA: Economic negotiator? Participants: 217+339, Prolific	Task: Play game - decide whether to accept or reject proposals of monetary splits made by human/AI knowing/not knowing their feedback will be used to train AI	Base pay: $8.50 per hour Bonus: Outcome-based (in-game earnings) Resulting pay: med. $10 per hour Time taken: 6 mins	"This experiment took 6 minutes to complete and the median pay rate for participants was approximately $10 per hour (all participants were paid $8.50 per hour before receiving a bonus)"	Study design - Incentive schemes to create stakes, encourage them to care about AI training: "To incentivize choice behavior, participants were in-formed that one trial would be randomly selected and re-solved at the end of the experiment. They would receive abonus of 5% of the amount they earned from the trial se-lected, and were informed that the bonus would increase to15%in the follow-up session to encourage them to return(3)" "The purpose of the follow-up session is mainly to provide stakes for participants to care about the AI trained on their data. Because the questions asked in this paper do not apply to this second session, we do not report the results here." Potential explanation of observation: "However, participants in the AI training con-dition knew they would return for a follow-up session, facingthe AI they trained with more rewards at stake. Therefore,the changes in behavior in the AI training condition may reflect a strategy to increase personal gains in this follow-up session rather than a genuine desire to foster fairness" Limitations: smaller rewards "nother conjecture is that the reward itself could influ-ence people's inclination to train AI for fairness. Specifi-cally, individuals may train AI to exhibit fair behavior sincethe rewards were relatively small. Indeed, people acceptmore unfair offers when more reward is at stake (Ander-sen et al. 2011; Slonim and Roth 1998; Novakova and Flegr2013). Thus, the small rewards in our study (maximum of15¢) may not have been enticing enough for individuals toprioritize personal gains. It would be interesting to deter-mine whether individuals would still be willing to forgo theirrewards in high stake scenarios. Systematically manipulat-ing the reward at stake might reveal a threshold at whichpersonal rewards outweigh the act of training AI to be fair."	Incentives used to encourage behaviour: Used to encourage participants to care about training the AI (higher bonus on returning) Studying the effect fairness, AI training, and game partner on accepting offers (economic self-interest) Suggest that people care about training AI even when monetary benefit not involved	https://ojs.aaai.org/index.php/HCOMP/article/view/27556/27329	"This study used the ultimatum game to examine whether in-dividuals are inclined to train AI to make equitable offers."
Decision Making Strategies and Team Efficacy in Human-AI Teams	Imani Munyaka, Zahra Ashktorab, Casey Dugan, J. Johnson, and Qian Pan	CSCW	2023	Word association game (Stakes: low)	IA: Game player Participants: 125, MTurk	Task: Play game with human/AI partner with diff. decision-making styles Post-survey	Base pay: $2.50 per participant Time taken: avg. 15 mins	"In pilot studies, the average time of completion was 15 minutes. Based on this, all participants were paid $2.50, commensurate with federal minimum wage."	None	Minimum wage	https://doi.org/10.1145/3579476	"We investigate how the decision-making of a team member in a human-AI team impacts the outcome of the collaboration and perceived team-efficacy" "We study how decision-making styles impact player behavior, perception of the team and perception of self, and game outcomes"

Improving Human-AI Collaboration With Descriptions of AI Behavior	Ángel Alexander Cabrera, Adam Perer, and Jason I. Hong	CSCW	2023	Fake Review (Deception) Detection Satellite Image Classification Bird Classification (Stakes: ?)	IA: Context-specific (role: classification) Participants: 225, MTurk (US, completed > 1,000 tasks, approval rating > 98%)	Introduction, examples Task: 30 tasks, label with/without AI with/without behavioural descriptions Survey, attention checks	Base pay: $2 per participant ($8 per hour) Time taken: 15 mins	"We paid participants $2 for the task, which lasted 15 minutes for an hourly compensation of $8 an hour." "Although we considered providing bonuses as an incentive for accurate responses, we found that the incentive to have the task approved was sufficient to get good results without a bonus. We confirmed this by finding similar average accuracy on the control condition for the reviews task, 56%, to that reported by Lai and Tan [34] on the same task using a bonus, 51%."	None	Intentional choice to not provide bonus, based on observing similar accuracy for similar tasks with/without bonus (compared with prev. work) [34] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection.	https://doi.org/10.1145/3579612	"To help people appropriately rely on AI aids, we propose showing them behavior descriptions, details of how AI systems perform on subgroups of instances"
Explanations Can Reduce Overreliance on AI Systems During Decision-Making	Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna	CSCW	2023	Maze solving (Stakes: low)	IA: Puzzle solvers? Participants: 340+170+31+114+76, Prolific (50 submissions, US, native English speakers, approval rating >= 95%)	Task: Solve mazes with/without AI assistance and explanations, number of trials depending on difficulty level	Study 1,2,3: Base pay: $4 per participant Time taken: 20 mins Study 4: Base pay: (not mentioned; probably same as others) Bonus: Performance-based, manipulated for diff. conditions: $0.10 (low bonus condition), $0.5 (high bonus condition) per correct completion Study 5: Base pay: $4 per participant Bonus: Performance-based, $0.10 (medium) $0.05 (easy) per 100 credits of in-game rewards	Study 1,2,3: "All participants received a payment of $4 for 20 minutes of their time with no bonuses, unless otherwise noted in a specific study's procedure section. We estimated the time required to complete the tasks through pilot experiments." Study 4: Study 5: "Participants receive a base payment of $4 for 20 minutes of their time, the same as Study 1 and Study 2. However, participants are also given opportunities for bonuses in this study when they answer a question correctly. To equalize participant perception of monetary reward and mitigate income effects, we use a credit system, where participants earn "credits" instead of dollar amounts. Participants are informed during the study that 100 credits are equivalent to $0.10 in the medium task and 100 credits are equivalent to $0.05 in the easy task (for participants who see the easy task)."	Design of study 4: To see the effect of high vs low bonus on overreliance: "First, we adopted a blocked, within-subjects study design in which participants are randomly placed in one bonus condition for the first half of the study and the other condition for the second half4. We chose this study design because prior work [86] shows that people calculate their judgements of monetary gains and losses relative to a certain reference point, which was indicative to us that we ought to have within-subjects comparisons of monetary value. Second, we add one question where the AI is incorrect to the collaboration phase as to have an even amount of questions with the low- and high- bonus." Result: Increased incentives reduce overreliance on AI "when the benefit (in this case, via monetary bonus) of doing a task properly increases, overreliance decreases." "Taken together, Study 3 demonstrates that overreliance is responsive not just to costs, but also to the benefit accrued for expending these costs. These findings also have implications on the designs of crowd-sourced studies, for example, as findings may differ when providing a bonus or not in a study." "(1) Monetary reward: Although this is not true of all tasks, some tasks, like crowd-sourcing, could benefit from reward systems where people are bonused for their joint work with AIs. Since we found an impact of rewards on overreliance, there may also be a difference in the findings of work that bonuses crowd-workers versus work that does not" "We find that people explicitly forego monetary rewards for an AI in a harder task or for an AI that gives more understandable explanations."	Study 4 directly studies the effect of monetary compensation on overreliance Suggests that studies may have different findings based on the bonus scheme [86] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases	https://doi.org/10.1145/3579605	"we formalize this strategic choice in a cost-benefit framework, where the costs and benefits of engaging with the task are weighed against the costs and benefits of relying on the AI. We manipulate the costs and benefits in a maze task, where participants collaborate with a simulated AI to find the exit of a maze. Through 5 studies (N = 731), we find that costs such as task difficulty (Study 1), explanation difficulty (Study 2, 3), and benefits such as monetary compensation (Study 4) affect overreliance. Finally, Study 5 adapts the Cognitive Effort Discounting paradigm to quantify the utility of different explanations, providing further support for our framework."
Eye into AI: Evaluating the Interpretability of Explainable AI Techniques through a Game with a Purpose	Katelyn Morrison, Mayank Jain, Jessica Hammer, and Adam Perer	CSCW	2023	Image classification (Stakes: ?)	IA: Game player / AI evaluator? Participants: Prolific (English as first language, US, approval rate > 95%, min. 50 submissions)	Task: 3 rounds, 3 times, round 1 (explainer): choose explanations, round 2&3 (guesser): guess AI prediction (with explanations) Post survey	Base pay: $1.60 per participant Bonus: Performance-based, $1 per participant for top 50% in-game scoring participants Time taken: avg. 12 mins	"Each worker was compensated $1.60 USD for their participation; Prolific reported an average pay of $9.50 USD per hour. Participants were assigned an anonymous ID for analyzing their responses and, on average, took 12 minutes to complete the game. To incentivize active, high-quality participation, we offered a bonus of $1 USD for those who scored within the top 50% of all participants. Out of the 50 participants, 25 participants received bonuses."	None	Incentivize high-quality participation Mentions GWAPs and how they don't need to rely strongly on financial incentives	https://doi.org/10.1145/3610064	"We designed Eye into AI, a GWAP, to help researchers collect data to evaluate and compare the interpretability of visually distinct and algorithmically different XAI techniques. Through an empirical study, we evaluated how well our GWAP achieves that goal by exploring the significance of player data..."
Being Trustworthy is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust	Nikola Banovic, Zhuoran Yang, Aditya Ramesh, and Alice Liu	CSCW	2023	Game of chess (Stakes: ?)	IA: Chess players Participants: 120, MTurk (US, 18+)	Qualification: Solve 15 chess puzzles Task: Play 3 games of chess with two AI coaches' assistance Questionnaire	Base pay: Qualification: $0.25 Completing 3 games: $15 per participant Bonus: Performance-based, $1.50 per participant for not losing any games Time taken: avg. 1 hour	"We compensated each participant $0.25 for completing the qualication task and $15.00 for completing three games of chess. Participants who did not lose any games received $1.50 bonus."	Earning bonus affected by behaviour (trusting untrustworthy AI): "Our results show evidence that participants reduced their reliance on the untrustworthy coach by the last game. However, by that time the damage would have already been done (i.e., they could have lost one of their first two games and thus a chance to earn the bonus)."		https://doi.org/10.1145/3579460	"...we conducted an experiment with 120 participants to test if untrustworthy AI can deceive end users to gain their trust"
How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System	Gaole He, Stefan Buijsman, and Ujwal Gadiraju	CSCW	2023	Loan approval prediction (Stakes: seemingly high)	IA: Loan approval officer Participants: 281+248, Prolific (English-speaking, 18+, >= 90% approval rate)	Main study: Pre-task questionnaire, 2 example cases Task: 10 trials, initial and final decision (with AI advice) Post-task questionnaire, attention checks Pilot: Similar, 10 trials but without AI advice Follow-up: Similar, 12 trials	Main study: Base pay: £1.5 per participant (£7.5 per hour) Bonus: Performance-based, £0.1 per correct decision Time taken: est. 12 mins Pilot: Base pay: £0.88 per participant (£7.5 per hour) Bonus: Performance-based, £0.1 per correct decision Time taken: avg. 8.5 mins Follow-up study: Base pay: £1.5 per participant (£9 per hour) Bonus: Performance-based, £0.1 per correct decision Time taken: est. 10 mins	"Compensation. All participants were rewarded with £1.5, amounting to an hourly wage of £7.5 deemed to be "good" payment by the platform (estimated completion time was 12 minutes). We rewarded participants with extra bonuses of £0.1 for every correct decision in the 10 trial cases. By incentivizing participants to reach a correct decision, we operationalize the concomitant "vulnerability" discussed by Lee and See[39] as a contextual requirement to encourage appropriate system reliance." Pilot: "In addition to the basic reward of £0.88 (equivalent to an hourly wage of £7.5), we set up a bonus of £0.1 for every correct decision to incentivize and encourage participants to concentrate on their individual decisions. On average, the pilot study was completed in 8.5 minutes, with an average accuracy of 0.43 (SD = 0.13)" Follow-up: " All participants were rewarded with £1.5, amounting to an hourly wage of £9 deemed to be "good" payment by the platform (estimated completion time was 10 minutes). Similar to the main study, we rewarded participants with extra bonuses of £0.1 for every correct decision in the 12 trial cases."	None	Incentives as a way of operationalizing "vulnerability" to encourage appropriate reliance; For pilot, incentives to focus on decisions Platform-determined "good" payment [39] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance.	https://doi.org/10.1145/3610067	'How does the understanding of stated system accuracy affect reliance of users on the AI system?' and 'How does explaining stated system accuracy using analogies affect the reliance of users on the AI system?'
When Does Uncertainty Matter?: Understanding the Impact of Predictive Uncertainty in ML Assisted Decision Making	Sean McGrath, Parth Mehta, Alexandra Zytek, Isaac Lage, Himabindu Lakkaraju	arXiv, TMLR	2020 (Pub. 2023)	Apartment price prediction (Stakes: low)	IA: ? Participants: Students & researchers (mailing lists/reaching out) + Prolific (approval rate >=98%, completed min. 250 submissions, US	Tutorial, practice run Task: estimate rental price once then again (with model prediction) End survey	Base pay: $3 per participant Bonus: Performance-based, 3 participants (in each of the two batches) received $30, based on alignment with model prediction Time taken: med. ~10 mins	"Participants were paid $3 for completing the study and could earn up to an additional $30 based on their performance in the study. Participants were told that their performance is measured based on the average distance between each apartment's true price and their first and second estimates. For the purposes of distributing bonus payments, we set the true price of the apartments to the model predictions, although recall that users were not told the true price of the 7 apartments at any point in the study. In each of the two phases of our study (see Section 3.2.5), the three participants with the lowest average distance received the $30 bonus. The median time to complete the study was approximately 10 minutes."	Limitations and Future Directions: "To help mitigate potential self-priming effects, we offered a $30 incentive for users to make the most accurate apartment price predictions. Furthermore, this study design has been suggested as good practice by prior work (Poursabzi-Sangdeh et al., 2021)."	Incentive communication: Performance assessment for bonus told to be distance with true price, calculcated as distance with model prediction Incentives to mitigate potential self-priming effects	https://arxiv.org/abs/2011.06167	systematically assess how people with differing levels of expertise respond to different types of predictive uncertainty in the context of ML assisted decision making for predicting apartment rental prices. • Does showing predictive uncertainty affect how closely people follow model predictions? • Does the effect of showing predictive uncertainty depend on the type of uncertainty – either the shape or the variance of the distribution? • Do participant with more expertise, either in the domain or in ML, more closely follow model predictions?
Do Explanations Help Users Detect Errors in Open-Domain QA? An Evaluation of Spoken vs. Visual Explanations (Prev: Human evaluation of spoken vs. visual explanations for open-domain qa)	Ana Valeria Gonzalez, Gagan Bansal, Angela Fan, Robin Jia, Yashar Mehdad, and Srinivasan Iyer	ACL IJNCLP	2021, preprint (2020)	Question answering (Stakes: seem low, simulated through penalty?)	IA: ? Participants: MTurk (7*75)	Task: 40 questions - accept/reject model prediction Post-task survey	Base pay: $10 for participation Bonus: Performance-based, $0.15 per accepting a correct answer, -$0.15 per accepting an incorrect answer, $0 for any rejection (max cumulative reward: $2.70, no deductions if -ve bonus) Resulting pay: $15 per hour + variable bonus Time taken: 35-45 mins	"Incentive scheme. In addition to providing a fixed upfront pay of $10 for participating in the task, to encourage workers to engage, we also used a bonus-based strategy (Bansal et al., 2019) — When users accept a correct answer, we provide a 15 cent bonus, but when they accept an incorrect answer they lose the same amount. When they reject an answer, however, they do not receive any bonus.(5) This aims to simulate the real-world cost and utility of users choosing to believe answers of an ODQA model. The maximum cumulative reward is $ 2.70. These values were chosen to ensure workers earned minimum a $15 hourly wage." "(5) if bonus is negative, no deductions re made from base pay. Bonus is instead set to zero"	None	Incentives to simulate real-world cost and utility (ecological validity) Bansal et al., 2019: The role of mental models in human-ai team performance.	https://aclanthology.org/2021.findings-acl.95.pdf	explore the effectiveness of explanations for Open-Domain Question Answering models, which involves answering users' questions using a large corpus