

Limits of Interventions in AI-Assisted Decision Making

Shikha Shrestha¹, Eunhye Ahn^{1,2}, Yevgeniy Vorobeychik¹

¹Washington University in St. Louis
St. Louis, Missouri, USA

²University of Wisconsin-Madison
Madison, Wisconsin, USA

dshrestha@wustl.edu, eunhye.ahn@wisc.edu, yvorobeychik@wustl.edu

Abstract

As AI decision-support systems are increasingly deployed in high-stakes domains, understanding how to prevent automation bias is critical. This study examines whether three common interventions, including accountability requirements, consequence framing, and sequential disclosure of AI recommendations, reduce over-reliance on AI in socially consequential decisions. We conducted a controlled experiment using simulated case studies of AI-assisted child maltreatment risk assessment.

We found that accountability requirements increased deliberation time but did not reduce over-reliance on AI or improve alignment with expert judgment. Sequential disclosure revealed that participants frequently abandoned an *expert-aligned* initial assessment in favor of the AI recommendation, producing no meaningful improvement in appropriate reliance. Consequence framing, by contrast, reduced over-reliance without increasing under-reliance and produced the highest rate of decisions consistent with expert judgment.

Across conditions, greater deliberation and more elaborate moral reasoning were associated with *increased AI reliance and lower alignment with expert judgment*. Furthermore, results from the sequential disclosure condition show that AI recommendations can lead participants to revise a prior independent judgment. These findings highlight the limits of deliberation-based interventions and point to consequence framing as a more promising approach for calibrating human reliance on AI in high-stakes settings.

Introduction

As algorithmic decision-support systems are deployed across increasingly consequential domains, from criminal sentencing and hiring to medical diagnosis and child protective services (Angwin et al. 2016; Barocas and Selbst 2016; Obermeyer et al. 2019), a shared institutional assumption has taken hold: that requiring human oversight of algorithmic recommendations will protect against the errors and biases these systems can produce (Green 2022; Agudo et al. 2024; European Parliament and Council of the European Union 2016). Regulatory frameworks, professional guidelines, and organizational policies converge on the principle that humans should remain meaningfully “in

the loop,” retaining both the authority and the responsibility to accept or override algorithmic outputs (Green 2022; Government of Canada 2021; Shneiderman 2022). Yet the empirical basis for this assumption remains incompletely developed. Even when AI operates at the level of recommendation rather than direct action, the mere presence of an algorithmic suggestion can systematically shift human judgment toward the algorithm’s conclusion, a phenomenon known as automation bias (Goddard, Roudsari, and Wyatt 2012; Alon-Barkat and Busuioc 2023). Critically, the institutional mechanisms most commonly proposed to counteract this shift may be less effective than assumed (Green 2022; Agudo et al. 2024; Coskun and Leidner 2024; Stevenson and Doleac 2024). The stakes of this question are not abstract. In child welfare, for example, algorithmic risk assessment tools are being adopted to assist caseworkers in screening decisions that determine whether a family is investigated for alleged maltreatment (Chouldechova et al. 2018; Vaithianathan et al. 2013). These tools generate risk scores that caseworkers are expected to integrate with their own judgment, but the conditions under which this integration reliably preserves independent judgment in high-stakes professional contexts remain inconsistently supported across studies. In child welfare and clinical settings, the question is not simply whether users trust AI appropriately. It is whether the humans nominally in the loop retain the capacity for genuine independent judgment. Empirical evidence from real clinical settings shows that when AI outputs conflict with professional judgment but provide no underlying reasoning, professionals often cannot evaluate which is correct (Lebovitz, Lifshitz-Assaf, and Levina 2022), yet they bear full professional and legal accountability for outcomes the algorithm influenced but does not share responsibility for (Busuioc 2021). Compounding this, practitioners routinely draw on information such as family dynamics observed in person, nonverbal cues, and cultural context that may be difficult to translate into the feature values an algorithm operates on (Kawakami et al. 2022; Saxena and Guha 2024). Consequently, the human and the algorithm are frequently not reasoning from the same evidence, even when they receive the same case file. If institutional safeguards designed to promote careful reasoning instead formalize deference to algorithmic outputs, the consequence is not merely suboptimal prediction; it is a transfer of effective decision authority

from accountable professionals to opaque statistical models, precisely the outcome that human oversight is meant to prevent (Green 2022). How decision-support interfaces are designed may matter, but the conditions under which design choices preserve or further erode independent judgment remain inconsistently supported across studies. We address this gap through a between-subjects experiment in a child welfare risk assessment context, comparing four conditions: a standard AI recommendation baseline, consequence framing (stakes salience), accountability (written justification requirement), and sequential disclosure (independent judgment before AI exposure). When an algorithmic recommendation is available before or during deliberation, it may function as a cognitive anchor, meaning an initial reference point that disproportionately shapes subsequent judgment, such that deliberation elaborates rather than evaluates the AI's conclusion (Green and Chen 2021; Rastogi et al. 2022; Carter and Liu 2025). Two prominent institutional responses to this problem, requiring justification and delaying AI disclosure, each attempt to counteract anchoring through different means. Consequence framing motivates a different possibility: heightening the salience of decision stakes before the recommendation is processed may change what receives attention. We investigate four questions within this framework. First, under what conditions do individuals follow versus resist AI recommendations in ethically significant decisions, and do interventions targeting accountability or consequence salience shift that balance? Second, when people are pushed to think more carefully about their decisions, does that careful thinking help them rely on AI more appropriately, or does it make them more likely to defer to it? Third, how is the requirement to provide a written justification reflected in participants' decision outcomes and alignment with AI recommendations, and in the framing of their written rationales? Fourth, do participants' initial judgments remain stable when AI is disclosed after an independent judgment, or do they revise toward the recommendation? Across conditions, we examine both decision outcomes and decision process: what participants chose, how long they deliberated, and how they reasoned in writing. Requiring written justification increased deliberation time but did not reduce AI deference. Participants in the sequential disclosure condition sometimes revised toward AI even when it diverged from expert-majority judgment, and more elaborate moral reasoning was associated with lower expert alignment. Consequence framing showed the most consistent protective pattern without requiring more deliberation. These results are compatible with multiple explanations, including anchoring, perceived authority, and attentional re-framing; the design does not establish any one mechanism.

Related Work

Automation Bias and AI-Assisted Decision-Making

Automation bias refers to the tendency for human decision-makers to defer to automated recommendations even when contradictory information is available or when independent judgment would produce better outcomes (Carter and Liu 2025; Parasuraman and Manzey 2010). First documented in

aviation and process control (Parasuraman and Riley 1997; Mosier et al. 1998), the phenomenon has since been demonstrated across medical diagnosis (Goddard, Roudsari, and Wyatt 2012; Lyell and Coiera 2017), judicial risk assessment (Stevenson and Doleac 2024), and organizational decision-making (Logg, Minson, and Moore 2019). The core finding is broadly consistent: when humans receive an algorithmic recommendation, they defer to it at rates exceeding what would be warranted by the algorithm's demonstrated accuracy, even when available evidence contradicts the recommendation (Skitka, Mosier, and Burdick 1999; Goddard, Roudsari, and Wyatt 2012).

Three explanations have been proposed for why people defer to algorithmic recommendations. First, algorithms serve as a mental shortcut that reduces cognitive effort, so people default to agreement unless they are motivated to think carefully (Evans 2008; Kahneman 2011). Second, people treat algorithms as authoritative experts and defer to them rather than weighing the evidence themselves (Sundar 2008). Third, the algorithm's recommendation acts as an anchor, meaning an initial reference point that shapes all subsequent reasoning. Anchoring is particularly important because it explains why thinking harder does not fix the problem: once an anchor is set, additional deliberation tends to build a case for the anchor's conclusion rather than question it (Tversky and Kahneman 1974; Mussweiler and Strack 1999).

Cognitive Anchoring and the Limits of Deliberation

The anchoring-and-adjustment heuristic specifies that when a salient anchor is provided, subsequent reasoning is selectively biased toward anchor-consistent information, and adjustments are systematically insufficient (Tversky and Kahneman 1974; Mussweiler and Strack 1999). In AI-assisted decision-making, the algorithmic recommendation functions as such an anchor (Rastogi et al. 2022). The selective accessibility model (Mussweiler and Strack 1999) proposes that the anchor primes semantically consistent knowledge, such that additional deliberation deepens rather than corrects the anchor's influence. Green and Chen (2021) provided direct support in AI contexts, showing that algorithmic risk assessments systematically altered how participants weighted risk, consistent with AI recommendations shifting the frame through which case information was evaluated.

This mechanism generates a counterintuitive prediction: accountability interventions may produce more elaborately justified deference rather than independent judgment (Kunda 1990). The locus of effective intervention is therefore attentional frame, not deliberative depth. Prior work provides some support for disclosure-based interventions: Bućinca, Malaya, and Gajos (2021) demonstrated that cognitive forcing functions reduce over-reliance on AI relative to standard simultaneous disclosure. Whether these protective effects persist when AI explicitly surfaces disagreement with the user's prior judgment remains the question the sequential disclosure condition is designed to answer.

Accountability as an Institutional Safeguard

Accountability, defined as the expectation of having to justify one's decisions to others, is among the most widely studied mechanisms for improving judgment quality (Lerner and Tetlock 1999). Tetlock's social contingency model (Tetlock 1983, 1992) provides the theoretical foundation: when decision-makers anticipate evaluation by an audience whose views are unknown, they engage in more effortful, integrative, and self-critical reasoning.

However, the accountability literature contains an important caveat. Tetlock's model specifies that accountability improves judgment primarily when the decision-maker has not yet committed to a position. When decision-makers know or can infer the evaluator's preferred conclusion, accountability can produce the opposite effect, motivating justification of the expected position (Tetlock 1983; Lerner and Tetlock 1999). In AI-assisted contexts, the algorithm's recommendation may function both as an anchor and as a signal of the expected conclusion, converting accountability pressure from a spur to independent thinking into a demand for rationalization.

Consequence Framing and Attentional Reframing

Consequence framing, which makes the potential real-world outcomes of a decision vivid and personally salient, may operate through at least two mechanisms. Under an attentional reframing account, consequence framing shifts the decision-maker's frame from *responding to a recommendation* to *predicting an outcome*, preserving independent judgment without requiring additional cognitive effort (Loewenstein et al. 2001; Lerner and Keltner 2001). When potential consequences are vivid and personally contingent, decision-makers may attend more to their own causal agency and less to the algorithm's assessment (Lerner et al. 2015; Lerner and Keltner 2000), disrupting anchoring by changing the object of attention before deliberation begins. Under the alternative risk-as-feelings hypothesis (Loewenstein et al. 2001; Slovic et al. 2004), affective responses to vivid outcome representations drive decision behavior by increasing motivational engagement.

Child Welfare as an Empirical Setting

Child welfare decision-making offers an informative testbed for studying intervention effects on AI-assisted decision-making. Unlike domains with clear, verifiable outcomes, decisions about whether evidence substantiates maltreatment involve contested evidence and rely heavily on professional judgment (Berrick 2018; Ahn et al. 2025). Compounding this, errors carry serious consequences in both directions: failing to intervene where maltreatment has occurred risks serious harm to a child, while unwarranted intervention can disrupt families. Despite the increasing deployment of algorithmic risk assessment tools in this domain (Chouldechova et al. 2018; Vaithianathan et al. 2013), experimental research on how interventions shape human-AI interaction in such high-stakes contexts remains limited.

Our controlled analogue task uses lay participants and therefore tests a broader assumption about human oversight

rather than how child welfare professionals would behave in practice. Professional expertise, institutional routines, and accountability structures may produce different reliance patterns, making replication with caseworkers essential.

Methods

Participants

This study was approved by the Washington University in St. Louis Institutional Review Board. Participants were recruited through Prolific (U.S. residents, $\geq 95\%$ approval rate, \$11/hr) and randomly assigned via Qualtrics to one of four between-subjects conditions. Among 280 participants recruited, we excluded participants with prior child welfare experience, either professional (e.g., social work, child protective services) or personal (e.g., having had a child involved with the system), to isolate intervention effects from variation in domain familiarity and to avoid drawing inference from participants whose lived experience could make the case material distressing. We further excluded those who failed an attention check ($n = 34$), completed the study in under 180 seconds ($n = 3$), exceeded 3 SD above the mean duration ($n = 7$), participated in multiple conditions ($n = 6$), or had participated in a prior pilot ($n = 7$). Because some participants met multiple criteria, these exclusions removed 55 unique participants. The final analytic sample comprised 224 participants (57 baseline, 57 consequence, 52 accountability, 58 No AI) contributing 1,344 case-level observations. A descriptive check retaining respondents across CPS-experience categories included 189 participants with no prior CPS experience, 18 with professional CPS exposure, and 20 with personal or direct CPS experience. The professional-exposure subgroup was small, but descriptively similar to the no-prior group: expert-majority alignment was 63.0% versus 62.7%, AI matching was 61.1% versus 63.8%, over-reliance was 17.6% versus 19.6%, and under-reliance was 19.4% versus 17.7%. This check is descriptive only and is not used for inferential claims.

Task and Materials

Each participant reviewed six child welfare case vignettes randomly assigned from a pool of 16 fictional cases. Each vignette included a referral narrative and a structured assessment covering family composition, living situation, child conditions, caregiver factors, and prior CPS history. Cases spanned a range of complexity and ambiguity.

An AI risk recommendation was generated for each case by inputting case feature values into a logistic regression classifier previously trained on the National Child Abuse and Neglect Data System (NCANDS) data, achieving 78 percent accuracy on held-out data. The recommendation was displayed as "The AI system predicts HIGH RISK [or LOW RISK] for maltreatment." During training, participants were told the AI was accurate about 80% of the time (rounded from the held-out accuracy for participant communication) and that it does not replace their judgment. They completed two practice cases, one in which AI aligned with the expert majority and one in which it did not, to calibrate expectations of fallibility. Of the 16 cases, 6 received high-risk

and 10 received low-risk AI recommendations. To establish an analytical benchmark for evaluating participants' AI reliance, three child welfare experts independently assessed the risk level of all cases. We use the majority decision as the benchmark.

Following informed consent, participants completed demographic and AI experience items. In the main task, participants reviewed the six cases in randomized order, with an attention check embedded midway. Participants were randomly assigned with equal allocation via Qualtrics to one of four between-subjects conditions. In three conditions, participants reviewed the case alongside the AI recommendation and made a binary risk assessment (High Risk or Low Risk); in the fourth, they made an initial assessment before seeing the AI recommendation. The interface presented each case in a linear flow: narrative case vignette, structured assessment, AI recommendation, and user decision, ensuring participants engaged with case evidence before encountering the AI output (Figure 1).

Experimental Conditions

Baseline. Participants reviewed the case and AI recommendation, then made their decision.

Consequence. Same as baseline, plus a reminder: "Your decision carries meaningful consequences. Screening out a risk case could leave a child unprotected. Screening in a safe case could cause unnecessary stress for a family."

Accountability. Same as baseline, but participants were required to provide a written justification for each decision immediately afterward.

Sequential Disclosure (referred to as the *No AI* condition in all figures and results). This condition was designed to test whether forming an independent judgment before AI exposure provides a sufficient anchor to reduce over-reliance, that is, whether prior commitment to a decision makes participants more resistant to deferring to the algorithm. Participants first made a decision without seeing the AI recommendation. They were not informed that an AI recommendation would follow; they were told only that they were reviewing child welfare cases and providing risk assessments. Only if their decision differed from AI were they shown a disagreement alert and given the opportunity to revise; participants whose initial judgment agreed with AI never saw the AI output at all.

Because initial decisions were made before the algorithmic recommendation was provided, they serve as the *No AI* reference across all analyses. This reference should be interpreted with one important caveat: *No AI* decisions derive from control condition participants prior to AI exposure rather than from a fully independent, naive *No AI* group. Although the design minimized demand characteristics by withholding any mention of an upcoming AI recommendation, a separate group with no AI exposure at any point would have provided a stronger baseline. *No AI* should further be distinguished from the baseline condition, in which participants saw the AI recommendation but received no additional intervention.

To trace how AI exposure shifted judgment within this condition, we decompose observations into three nested an-

alytic distributions: *Control: No AI*, comprising all unaided decisions before AI exposure; *Disagreed subgroup*, the portion of that group whose initial decisions differed from AI and who were therefore shown the recommendation and given the opportunity to revise; and *Final Combined*, the full group again with any revisions from the disagreed subset incorporated.

Expert Benchmark

Three independent raters with prior professional experience in child welfare assessed each case on a four-point risk scale (1 = lowest risk, 4 = highest risk). Ratings were converted to a binary classification matching the format of the AI recommendation and participant decision: scores of 1–2 were classified as low risk and scores of 3–4 as high risk. The majority vote across the three binary ratings constitutes the *expert majority judgment* for each case.

Agreement on the binary ratings was moderate (Fleiss' $\kappa = 0.59$). Experts were unanimous on 13 of 16 vignettes and split 2–1 on the remaining 3, with 87.5% pairwise agreement.

In child welfare, as in many high-stakes domains, definitive ground truth is rarely available; risk assessment involves weighing ambiguous and sometimes conflicting indicators, and reasonable professionals may reach different conclusions from the same evidence (Dawes, Faust, and Meehl 1989). Expert majority judgment does not resolve this ambiguity, and we do not treat it as a proxy for the correct answer. Rather, it serves as a stable reference point against which we can assess whether participants adjusted their reliance based on the quality of the AI recommendation or followed it regardless.

We further characterize cases by the degree of rater consensus. Cases with unanimous agreement (3–0) have higher expert consensus; split decisions (2–1) indicate expert ambiguity because trained professionals reached different conclusions.

Measures

Outcome Variables. The primary outcome of interest is whether participants' decisions matched expert majority judgment, operationalized as *expert alignment* (binary: 1 = decision matched expert majority, 0 = did not match). This is the central criterion against which we evaluate decision quality across conditions, since it captures whether participants made the choice that trained professionals would most likely endorse, regardless of what the AI recommended.

Secondary outcomes include *AI reliance*, *decision latency*, *justification effort*, and *decision revision*, each defined below.

AI reliance (binary) captures whether a participant's final risk assessment matched the AI recommendation (1 = followed AI, 0 = overrode AI). To allow finer-grained analysis, each case-level observation is further classified into one of four *reliance types* based on the intersection of AI following and expert alignment: *appropriate reliance* (followed AI when experts agreed), *appropriate deviation* (overrode AI when experts also disagreed), *over-reliance* (followed AI

(a) Baseline

AI Recommendation: The AI system predicts **HIGH RISK** for maltreatment.

Based on the information provided, how would you rate the overall risk to the child?

☐ High Risk

☐ Low Risk

(b) Consequence

AI Recommendation: The AI system predicts **HIGH RISK** for maltreatment.

Reminder: Your decision carries meaningful consequences.

- Screening out a risky case could leave a child unprotected.
- Screening in a safe case could cause unnecessary stress for a family.

Based on the information provided, how would you rate the overall risk to the child?

☐ High Risk

☐ Low Risk

(c) Accountability

AI Recommendation: The AI system predicts **HIGH RISK** for maltreatment.

Please note: Your decision and reasoning will be reviewed by Child Welfare experts. You'll be asked to explain why you accepted or rejected the AI's advice.

Based on the information provided, how would you rate the overall risk to the child?

☐ High Risk

☐ Low Risk

----- ↓ Next page -----

In a few words, describe the specific evidence from the case report that supports your final decision.

0/500

(d) Sequential Disclosure

Based on the information provided, how would you rate the overall risk to the child?

☐ High Risk

☐ Low Risk

----- ↓ Next page -----

There is a disagreement between your previous decision and the AI's recommendation. Would you like to revise your answer?

AI's assessment: High Risk **Your assessment: Low Risk**

☐ High Risk

☐ Low Risk

Figure 1: Experimental interface for each condition (case report and structured assessment, shared across all conditions, are omitted). (a) Baseline: AI recommendation and decision. (b) Consequence: adds a reminder of screening consequences. (c) Accountability: adds a review notice before the decision and a justification prompt on the next page. (d) Sequential Disclosure: participants first decide without AI; if they disagree, a disagreement alert is shown on the next page with the option to revise.

when experts disagreed), and *under-reliance* (overrode AI when experts agreed).

Decision latency (continuous) is the time in seconds from page load to decision submission, log-transformed prior to analysis to address right skew. We treat decision latency as a behavioral proxy for deliberation effort: in the No AI condition it captures unaided deliberation, while in treatment conditions it captures post-recommendation processing.

Justification effort (accountability condition only) is operationalized along two dimensions: *justification time* (seconds, log-transformed) and *word count* (log-transformed), both measured from the written rationale participants provided after each decision.

Decision revision (No AI condition only; binary) indicates whether a participant who initially disagreed with the AI recommendation changed their decision after seeing it (1 = switched, 0 = maintained). Only participants in the disagreed subgroup contribute observations to this outcome.

Predictor and Moderator Variables. *Experimental con-*

dition is a four-level factor (No AI, baseline, consequence, accountability). *AI-expert alignment* (binary) indicates whether the AI recommendation matched expert majority judgment for a given case. *Expert consensus* (binary) distinguishes unanimous (3–0) from split (2–1) expert agreement. Individual-level covariates included *AI familiarity* (three levels: 0, 1, or 2+ domains of daily use), *AI literacy* (three levels: 0–1, 2, or 3 items endorsed), *gender*, and *education level*.

Justification Coding

Justification content was coded on three dimensions corresponding to the theoretical frameworks introduced in the section on accountability. Reasoning quality (1–4) captured the depth of inferential reasoning, from simple restatement of the decision to identification of specific case factors linked through a logical chain, following frameworks of argument structure and elaboration depth (Toulmin 1958; Tetlock 1983). Moral framing (binary) indi-

cated whether the justification explicitly invoked concepts of harm, safety, duty, welfare, or caregiver responsibility, drawing on care ethics and harm-based moral reasoning frameworks (Noddings 1984; Haidt 2001). Case specificity (binary) indicated whether the justification referenced particular details from the vignette rather than generic risk language, operationalizing the distinction between individuating and base-rate categorical reasoning (Kahneman and Tversky 1973; Pennington and Hastie 1992). Coding was applied to all 312 justification texts using the Claude API (claude-sonnet-4-20250514) with a structured prompt specifying each dimension and its response format. To assess validity, one human rater independently coded a randomly selected subset of 30 justifications (approximately 12% of the sample) blind to participant condition and reliance outcome. Agreement between the human rater and the API-generated codes was high across all dimensions (reasoning quality: $\kappa = .86$; moral framing: $\kappa = .92$; case specificity: $\kappa = .89$; AI reference: $\kappa = .83$), supporting the reliability of the automated coding scheme.

Analytic Approach

Statistical Models. We organize our analysis around three dimensions of human-AI decision-making: what participants decided (AI reliance and expert alignment), how long they took (decision latency), and, for the accountability condition, how they reasoned about it (justification effort). For each dimension, we fit Bayesian multilevel models using the *brms* package in R (Bürkner 2017), with weakly informative priors (Normal(0, 2) on fixed effects and Student-t(3, 0, 2.5) on variance components) to regularize estimates away from extreme values. All models include participant- and case-specific random intercepts to separate individual tendencies from case characteristics, and use 4 chains of 4,000 iterations (2,000 warmup).

AI Reliance. To examine how experimental condition and AI-expert alignment shaped reliance, we regress the binary AI reliance indicator on experimental condition (No AI, baseline, consequence, accountability), AI-expert alignment, and their interaction, with a Bernoulli likelihood. This interaction allows us to test whether participants in different conditions were more or less sensitive to whether the AI recommendation matched the expert majority, that is, whether condition affected not just overall following rates but calibrated following. A sensitivity model adds gender, education, AI familiarity, and AI literacy as covariates to assess whether condition effects are robust to individual differences in background and prior AI experience.

Decision Latency. To examine deliberation effort, we regress log-transformed decision latency on experimental condition, AI-expert alignment, and their interaction, with a Gaussian likelihood. A covariate-adjusted variant parallels the sensitivity model above.

Justification Effort. Within the accountability condition, we fit two additional models characterizing the written justifications: one regressing log-transformed justification time, the other log-transformed word count, each on whether the participant followed the AI recommendation, AI-expert alignment, and their interaction. We complement these mod-

els with the content coding variables described in the Justification Coding subsection above, specifically reasoning quality, moral framing, case specificity, and AI reference, examined both as predictors of reliance and as descriptive summaries aggregated by reliance type. This allows us to assess not only how much effort participants invested in justifying their decisions, but also what kinds of reasoning were associated with different patterns of AI reliance.

Decision Revision. Finally, we examine decision revision among participants in the No AI condition who were shown an AI disagreement alert after their initial unaided decision and given the opportunity to revise. Because only participants whose initial decision differed from the AI recommendation were ever shown the recommendation, only this disagreeing subgroup could meaningfully revise; revisions here directly index how much weight participants assigned to the AI recommendation when it conflicted with their own judgment. We regress the binary decision revision indicator for this subgroup on AI-expert alignment, that is, whether the AI recommendation the participant was shown matched the expert majority, to assess whether switching was associated with this benchmark. We also model log-transformed revision time as a function of whether the participant switched and whether AI was expert-aligned, to assess whether switching versus maintaining one's initial judgment carries distinct temporal signatures.

We extend the revision analysis by adding expert consensus as a moderator, regressing the binary decision revision indicator on AI-expert alignment, expert consensus (unanimous vs. split expert agreement), and their interaction, to assess whether switching patterns differed between higher-consensus and more ambiguous cases. This model includes participant- and case-level random intercepts and uses the same prior specification as the primary revision model.¹

Results

What They Decided: AI Reliance Across Conditions

Participants who saw the AI recommendation followed it more often than those who decided without it. All intervals reported below are 95% credible intervals (CIs). In the No AI condition, participants' decisions matched the AI recommendation 54% [43, 66] of the time, reflecting the baseline rate of agreement between independent judgment and the algorithm on the same cases. When participants saw the AI recommendation alongside the case, this rate rose to 74% [63, 83] in the baseline condition and 69% [56, 80] in the accountability condition; both increases over the No AI reference were credible (baseline: $b = 1.04$ [0.47, 1.64]; account-

¹ Average marginal predictions from each model are computed by averaging posterior predictions across observed covariate distributions within each subgroup rather than at fixed reference values. For instance, to characterize reliance in the accountability condition, we compute $\Pr(\text{followed AI})$ for each accountability-condition observation using its expert-consensus category and AI-expert alignment, then average within the group. This yields estimates that reflect where each condition's behavior sits given the joint distribution of covariates participants actually encountered.

	Primary	With Covariates
<i>Fixed effects</i>		
Intercept	−0.34 [−1.29, 0.56]	−0.58 [−1.63, 0.50]
Baseline	1.04 [0.47, 1.64]	1.09 [0.50, 1.68]
Consequence	0.04 [−0.54, 0.64]	0.11 [−0.50, 0.71]
Accountability	0.61 [0.01, 1.21]	0.61 [0.01, 1.23]
AI-expert aligned	0.82 [−0.34, 1.98]	0.82 [−0.39, 1.96]
Baseline × AI-expert aligned	−0.19 [−0.89, 0.50]	−0.17 [−0.88, 0.53]
Consequence × AI-expert aligned	0.51 [−0.19, 1.21]	0.54 [−0.18, 1.27]
Acct. × AI-expert aligned	0.11 [−0.62, 0.82]	0.15 [−0.59, 0.89]
Male		0.27 [−0.03, 0.58]
Some college		−0.06 [−0.54, 0.43]
Bachelor's degree		−0.22 [−0.69, 0.22]
Graduate degree		−0.04 [−0.57, 0.49]
AI fam.: 1 domain		0.40 [0.06, 0.75]
AI fam.: 2+ domains		0.08 [−0.32, 0.47]
AI lit.: 2 items		0.10 [−0.36, 0.56]
AI lit.: 3 items		0.01 [−0.43, 0.43]
<i>Random effects</i>		
<i>SD_{case}</i>	1.06 [0.69, 1.63]	1.07 [0.70, 1.66]
<i>SD_{pid}</i>	0.51 [0.20, 0.77]	0.54 [0.22, 0.79]
Observations	1,344	1,344
Participants	224	224

Table 1: Bayesian multilevel models of AI reliance. Reference levels: No AI (condition), AI-expert not aligned (AI-expert alignment), Female (gender), High school or less (education), 0 domains (AI familiarity), 0–1 items (AI literacy). Estimates are posterior means with 95% credible intervals. Bold indicates intervals excluding zero.

Note. This model predicts whether the participant followed the AI recommendation. Estimates for the effect of AI exposure on expert-aligned decision rates by condition are derived from the reliance-type prediction distributions (see main text); those contrasts are not reproduced here for space.

ability: $b = 0.61 [0.01, 1.21]$; Table 1). The consequence condition showed no credible increase in AI following relative to No AI (61% [48, 74]; $b = 0.04 [−0.54, 0.64]$).

The central outcome of interest is whether participants' decisions matched expert majority judgment—that is, whether following or overriding AI produced more expert-aligned choices. The consequence condition produced the highest rate of expert-aligned decisions (67.1%), compared with 61.7% in baseline, 59.8% in accountability, and 61.0% in No AI (Figure 2). This pattern held despite the consequence condition's lower overall AI-following rate, suggesting that reduced AI following in this condition was concentrated on cases where deviation from AI was warranted.

Across all conditions, none of the interventions credibly moderated the association between AI-expert alignment and participants' tendency to follow AI. The interaction terms for all conditions were near zero with intervals spanning it (baseline × AI-expert aligned: $−0.19 [−0.89, 0.50]$; consequence: $0.51 [−0.19, 1.21]$; accountability: $0.11 [−0.62, 0.82]$; Table 1). No condition made participants meaningfully more likely to follow AI when it was expert-aligned and to deviate when it was not.

On cases where the AI recommendation did not match expert majority, baseline participants were credibly less likely to make an expert-aligned choice than those deciding without AI ($b = −0.99 [−1.55, −0.44]$), as were accountability

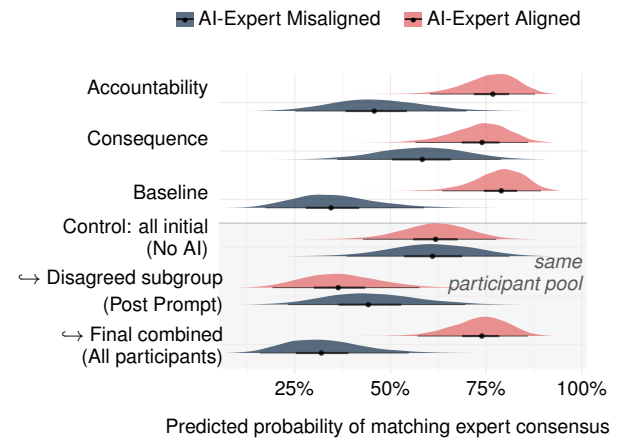


Figure 2: Predicted probability of matching expert majority by condition and AI-expert alignment. Blue distributions represent AI-expert not aligned cases; red distributions represent AI-expert aligned cases. Points and thick bars indicate posterior means and 50% credible intervals; thin bars indicate 95% credible intervals. The sequential disclosure condition is decomposed into three distributions: *Control: all initial* (unaided decisions before any AI exposure); *Disagreed subgroup* (decisions among the subset who initially disagreed with AI and were subsequently shown it); and *Final Combined* (combined final decisions across the full control group after the revision opportunity).

condition participants ($b = −0.55 [−1.13, −0.02]$); consequence participants did not show this pattern ($b = −0.04 [−0.61, 0.51]$), performing at a level indistinguishable from the No AI reference on these cases.

How They Engaged: Decision Latency Across Conditions

Decision latency varied by condition. Participants in the accountability condition spent an estimated 39 seconds [32, 47] on each decision page—approximately 34% longer than the No AI (29 s [24, 35]) and baseline (29 s [24, 35]) conditions.² The consequence condition showed a modest, non-credible increase (31 s [26, 37]) that overlapped substantially with baseline (Figure 3).

Decision latency also varied by case characteristics. In the No AI condition, participants spent noticeably longer on cases where AI and expert assessments diverged. In the baseline and consequence conditions, deliberation time was roughly equivalent across cases regardless of AI-expert alignment, so case-related variation in deliberation was attenuated when AI was visible. The accountability condition was the sole exception, producing the only credible interaction with AI-expert alignment ($b = 0.16 [0.01, 0.32]$; Table 2): participants took disproportionately longer on AI-

²Decision latency was measured as time on the decision page prior to submission. In the accountability condition, justification writing occurred on a subsequent page and is not included in this measure.

	Primary	With Covariates
<i>Fixed effects</i>		
Intercept	3.40 [3.17, 3.64]	3.34 [3.02, 3.67]
Baseline	−0.03 [−0.26, 0.20]	−0.02 [−0.24, 0.20]
Consequence	0.06 [−0.16, 0.28]	0.11 [−0.12, 0.33]
Accountability	0.20 [−0.04, 0.43]	0.20 [−0.04, 0.43]
AI-expert aligned	−0.06 [−0.31, 0.18]	−0.07 [−0.32, 0.19]
Baseline × AI-expert aligned	0.06 [−0.09, 0.21]	0.07 [−0.08, 0.22]
Consequence × AI-expert aligned	0.03 [−0.12, 0.18]	0.04 [−0.11, 0.19]
Acct. × AI-expert aligned	0.16 [0.01, 0.32]	0.17 [0.02, 0.32]
Male		−0.05 [−0.20, 0.09]
Some college		0.13 [−0.11, 0.37]
Bachelor's degree		−0.08 [−0.30, 0.13]
Graduate degree		−0.12 [−0.40, 0.15]
AI fam.: 1 domain		0.04 [−0.13, 0.22]
AI fam.: 2+ domains		0.07 [−0.13, 0.26]
AI lit.: 2 items		0.18 [−0.05, 0.42]
AI lit.: 3 items		0.01 [−0.20, 0.22]
<i>Random effects</i>		
SD_{case}	0.21 [0.14, 0.32]	0.21 [0.14, 0.32]
SD_{pid}	0.53 [0.48, 0.59]	0.53 [0.47, 0.58]
$SD_{residual}$	0.45 [0.44, 0.47]	0.45 [0.44, 0.47]
Observations	1,344	1,344
Participants	224	224

Table 2: Bayesian multilevel models of decision latency. Outcome is log-transformed decision page time (seconds). Reference levels and formatting same as Table 1.

Note. All coefficients are on the log-seconds scale. Predicted decision times reported in the text are back-transformed marginal predictions from this model, averaging over the observed covariate distribution within each condition. The estimate $b = 0.63 [0.12, 1.16]$ for the association between decision latency and AI following within the accountability condition is from a supplementary model fit within the accountability subsample only; it is not reproduced here for space.

expert-aligned cases. Moreover, within the accountability condition, longer decision times predicted *higher* AI following ($b = 0.63 [0.12, 1.16]$; see note in Table 2).

The variance structure of the models further reveals a dissociation: reliance was primarily case-driven ($SD_{case} = 1.06$ vs. $SD_{pid} = 0.51$), while latency was primarily person-driven ($SD_{pid} = 0.53$ vs. $SD_{case} = 0.21$), indicating that what people decide depends on which case they face, while how long they take reflects stable individual differences.

How They Reasoned: Justification Patterns in the Accountability Condition

Content coding of the 312 written justifications revealed that 98.1% were framed in purely case-based terms, with no reference to the AI recommendation. Reasoning quality did not credibly predict appropriate reliance ($b = 0.20 [−0.32, 0.74]$) or expert-aligned decisions ($b = 0.20 [−0.31, 0.70]$). However, two reasoning dimensions showed negative associations with expert-aligned decisions. Moral framing was negatively associated with both appropriate reliance ($b = −0.62 [−1.26, −0.02]$) and expert alignment ($b = −0.62 [−1.26, 0.03]$); the association with appropriate reliance was credible but the association with expert alignment was not. Case specificity showed a negative association with both appropriate reliance ($b = −1.07 [−2.15, −0.06]$) and expert alignment ($b = −1.07 [−2.11, −0.07]$), with both associations credible.

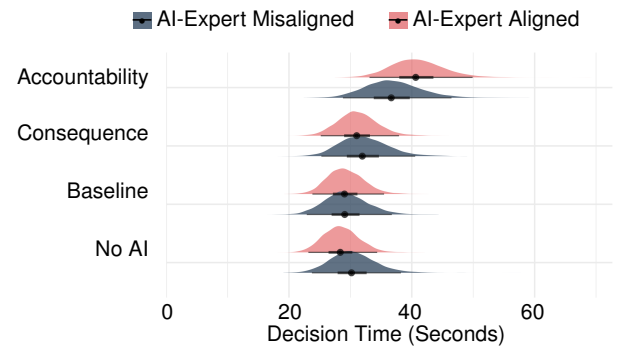


Figure 3: Predicted decision latency (seconds) by condition and AI-expert alignment. Red distributions represent AI-expert aligned cases; blue distributions represent AI-expert not aligned cases.

Moral Framing by Reliance Type

Percentage of justifications coded as containing moral framing

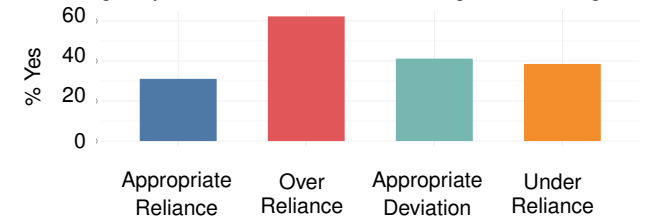


Figure 4: Percentage of participant justifications coded as containing moral framing, broken down by reliance type. Over-reliance cases showed the highest rate of moral framing (approximately 63%), followed by Appropriate Deviation (approximately 41%) and Under-reliance (approximately 39%). Appropriate Reliance cases had the lowest rate (approximately 31%), suggesting that participants were more likely to invoke moral reasoning when their decisions diverged from what AI recommended in a harmful direction.

This pattern is reinforced by how reasoning dimensions are distributed across reliance types (Figure 4). Over-reliance justifications had the highest rates of both moral framing (63%) and case specificity (42.6%) among all four reliance types, indicating that more elaborate and emotionally engaged reasoning was most prevalent in decisions that aligned with AI but diverged from expert majority.

No AI Condition: Revision Decisions Following AI Exposure

Participants in the No AI condition frequently revised their initial independent decisions to align with the AI recommendation after exposure, and did so at comparable rates whether the AI was expert-aligned or not. Among participants whose initial decisions differed from the AI recommendation ($n = 164$ disagreement instances from 58 participants), the switching rate was 42.9% when the AI recommendation was not expert-aligned and 28.7% when it was

Rate of revision towards the AI recommendation
Among participants who initially disagreed with the AI ($n = 164$)

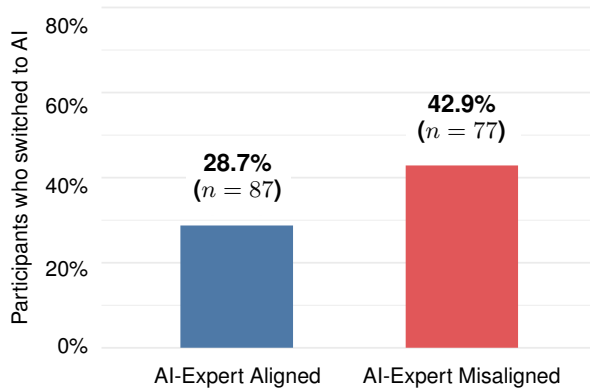


Figure 5: Switching rate by AI-expert alignment among control condition participants who initially disagreed with AI ($n = 164$ disagreement instances). Participants switched toward an AI recommendation that was not expert-aligned at nearly twice the rate of switching toward an AI-expert aligned one (42.9% versus 28.7%); the model's posterior interval for this contrast spanned zero.

expert-aligned (Figure 5). That is, participants were more likely to revise away from expert-aligned initial judgments than toward them. The model coefficient for AI-expert alignment on switching was in the expected direction ($b = -0.77$ [$-1.91, 0.30$]), but the posterior interval spanned zero.

However, this aggregate result obscures a sharper finding in the expert-consensus analysis. On unanimous (3-0) cases, where rater consensus was highest, the probability of switching toward an AI recommendation that was not expert-aligned credibly exceeded the probability of switching when the AI recommendation was expert-aligned (contrast = 0.31 [0.03, 0.61]; $P(\text{not aligned} > \text{aligned}) = 0.985$). On split (2-1) cases the direction was similar but the interval spanned zero. Uncalibrated switching was therefore concentrated on higher-consensus cases.

Revision decisions were fast, with a mean revision page time of approximately 12 seconds across all disagreement instances. The net result was a slight deterioration in aggregate decision quality: final decisions after the revision process (58.8% [53.6, 64.0]) fell below initial unaided rates (61.0% [55.9, 66.1]).

Discussion

This study tested whether consequence framing, written justification, or sequential AI disclosure changed how lay participants used an imperfect recommendation in a controlled child welfare analogue task. AI recommendations shaped judgment across most conditions, and requiring more deliberation did not produce more discriminating reliance. These are behavioral findings, not tests of a unique cognitive mechanism. Anchoring, perceived authority, attentional reframing, affective responses, and participants' doubts about their initial reviews are possible interpretations; the design cannot

determine which account caused the observed patterns.

Accountability Does Not Protect Against Automation Bias

The accountability requirement increased decision time but did not produce a reliance profile that differed from baseline. More effort, higher-quality written reasoning, and greater moral engagement did not increase deviation from AI when the recommendation differed from expert majority judgment. Moreover, 98.1% of justifications did not mention AI. This pattern could reflect post-hoc rationalization or implicit authority deference (Logg, Minson, and Moore 2019; Nisbett and Wilson 1977), but it could also arise from participants treating the recommendation as one unmentioned input. The experiment did not directly measure these explanations.

Deliberation and Deference Are Positively Associated

Without AI exposure, participants spent longer on more ambiguous cases; with AI visible, this case-related variation was attenuated in the baseline and consequence conditions. Within the accountability condition, longer decision times were associated with higher AI reliance. This association is inconsistent with a simple claim that additional deliberation necessarily reduces heuristic reliance (Evans and Stanovich 2013; Kahneman 2011), but it does not show that deliberation caused deference. Selective accessibility is one possible account (Mussweiler and Strack 1999).

Consequence Framing Preserves Independent Judgment Without Increasing Deliberation

Consequence participants had the highest rate of alignment with expert-majority judgment, without increased decision time or a credible increase in AI following when AI and experts disagreed. This parallels evidence that increasing the perceived cost of errors can reduce over-reliance (Vasconcelos et al. 2023). Attentional reframing and risk-as-feelings accounts both offer plausible interpretations (Loewenstein et al. 2001; Lerner et al. 2015), but the present study establishes only that the condition's effect did not depend on longer deliberation or more elaborate written moral reasoning.

Elaborated Moral Reasoning Was Associated With Over-Reliance

Moral framing and case specificity were negatively associated with expert alignment, and over-reliance justifications had among the highest reasoning-quality scores. These associations show that visible markers of careful reasoning can coexist with AI deference; they do not establish that moral reasoning or case-specific reasoning caused the decisions.

Sequential Disclosure Reveals the Limits of Pre-Commitment

Among participants who initially disagreed with AI, switching occurred in 42.9% of instances when AI differed from the expert majority and 28.7% when it aligned; the posterior interval for this aggregate association spanned zero. In the expert-consensus analysis, the difference was credible on unanimous cases (contrast = 0.31 [0.03, 0.61]). Thus, prior commitment did not reliably protect an initial judgment from later AI exposure. This pattern may reflect authority deference or anchoring, but it could also reflect participants reconsidering perceived

gaps in their initial review, including differences between narrative information and the model's structured inputs. The design does not distinguish these possibilities.

Implications for Design Written documentation alone should not be treated as evidence that a reviewer reached an independent judgment. Interfaces that delay recommendations until an initial assessment is registered or surface the features driving a prediction merit further testing (Rastogi et al. 2022; Schemmer et al. 2023). Our results do not establish that either design will generalize beyond this task.

Limitations Participants were lay adults recruited online rather than child welfare professionals, so the study does not show how caseworkers would behave. A CPS-retained descriptive check was similar to the primary sample, but the professional-exposure subgroup was small and cannot support inferential claims; professional replication is needed. The expert benchmark had moderate agreement and represents majority judgment, not ground truth. The No AI reference comprises initial decisions from sequential-disclosure participants rather than a fully independent group. Justification coding used AI-generated labels validated on a human-coded subset and identifies correlations rather than causes. Finally, the explainable, unbranded logistic-regression classifier may elicit different reliance from LLM-based, branded, or less interpretable systems.

Conclusion

Accountability increased effort without reducing AI deference, and sequential disclosure did not fully preserve prior independent judgments. Consequence framing produced the highest expert-majority alignment without longer deliberation. Together, the findings caution against equating observable deliberative effort with independent oversight and motivate evaluation of decision framing in professional settings. They do not establish why the interventions produced these patterns or whether the results generalize to practicing caseworkers or other AI systems.

Acknowledgments

This work was partially supported by the National Science Foundation (IIS-2214141, CCF-2403758), Foresight Institute, Amazon, and The Ann W. and Spencer T. Olin-Chancellor's Fellowship.

References

Agudo, U.; Liberal, K. G.; Arrese, M.; and Matute, H. 2024. The impact of AI errors in a human-in-the-loop process. *Cognitive Research: Principles and Implications*, 9(1): 3.

Ahn, E.; Choi, M.; McNellan, C.; and Goldkind, L. 2026. AI governance in social work: A triple mandate-informed accountability model. Working Paper 6469798, SSRN.

Ahn, E.; Morstatter, F.; Waters-Roman, D.; Palmer, L.; and McCroskey, J. 2025. Qualitative exploration of child welfare workers' decision-making experiences and perspectives on fairness. *Journal of Public Child Welfare*, 19(1): 229–252.

Alon-Barkat, S.; and Busuioc, M. 2023. Human–AI interactions in public sector decision making: “Automation bias” and “selective adherence” to algorithmic advice. *Journal of Public Administration Research and Theory*, 33(1): 153–169.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica. May 23. Available at <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

Barocas, S.; and Selbst, A. D. 2016. Big data's disparate impact. *California Law Review*, 104: 671–732.

Berrick, J. D. 2018. *The Impossible Imperative: Navigating the Competing Principles of Child Protection*. Oxford University Press.

Bürkner, P.-C. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1): 1–28.

Busuioc, M. 2021. Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review*, 81: 825–836.

Bučinca, Z.; Malaya, M. B.; and Gajos, K. Z. 2021. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 188.

Carter, L.; and Liu, D. 2025. How was my performance? Exploring the role of anchoring bias in AI-assisted decision making. *International Journal of Information Management*, 82: 102875.

Chouldechova, A.; Benavides-Prado, D.; Fialko, O.; and Vaithianathan, R. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 134–148.

Coskun, M.; and Leidner, D. 2024. Putting a human in the loop: Increasing uptake, but decreasing accuracy of automated decision-making. *PLOS ONE*, 19(2): e0298037.

Dawes, R. M.; Faust, D.; and Meehl, P. E. 1989. Clinical versus actuarial judgment. *Science*, 243(4899): 1668–1674.

European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation). *Official Journal of the European Union*, L 119: 1–88.

Evans, J. S. B. T. 2008. Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59: 255–278.

Evans, J. S. B. T.; and Stanovich, K. E. 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3): 223–241.

Goddard, K.; Roudsari, A.; and Wyatt, J. C. 2012. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1): 121–127.

- Government of Canada. 2021. Directive on Automated Decision-Making. Treasury Board of Canada Secretariat. Available at <https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592>.
- Green, B. 2022. The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45: 105681.
- Green, B.; and Chen, Y. 2021. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 418.
- Haidt, J. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4): 814–834.
- Kahneman, D. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.
- Kahneman, D.; and Tversky, A. 1973. On the psychology of prediction. *Psychological Review*, 80(4): 237–251.
- Kawakami, A.; Sivaraman, V.; Cheng, H.-F.; Stapleton, L.; Cheng, Y.; Qing, D.; Perer, A.; Wu, Z. S.; Zhu, H.; and Holstein, K. 2022. Improving human-AI partnerships in child welfare: Understanding worker practices, challenges, and desires for algorithmic decision support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–18. ACM.
- Kunda, Z. 1990. The case for motivated reasoning. *Psychological Bulletin*, 108(3): 480–498.
- Lebovitz, S.; Lifshitz-Assaf, H.; and Levina, N. 2022. To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science*, 33(1): 126–148.
- Lerner, J. S.; and Keltner, D. 2000. Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition & Emotion*, 14(4): 473–493.
- Lerner, J. S.; and Keltner, D. 2001. Fear, anger, and risk. *Journal of Personality and Social Psychology*, 81(1): 146–159.
- Lerner, J. S.; Li, Y.; Valdesolo, P.; and Kassam, K. S. 2015. Emotion and decision making. *Annual Review of Psychology*, 66: 799–823.
- Lerner, J. S.; and Tetlock, P. E. 1999. Accounting for the effects of accountability. *Psychological Bulletin*, 125(2): 255–275.
- Loewenstein, G.; Weber, E. U.; Hsee, C. K.; and Welch, N. 2001. Risk as feelings. *Psychological Bulletin*, 127(2): 267–286.
- Logg, J. M.; Minson, J. A.; and Moore, D. A. 2019. AI-gorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151: 90–103.
- Lyell, D.; and Coiera, E. 2017. Automation bias and verification complexity: A systematic review. *Journal of the American Medical Informatics Association*, 24(2): 423–431.
- Mosier, K. L.; Skitka, L. J.; Heers, S.; and Burdick, M. 1998. Automation bias: Decision making and performance in high-tech cockpits. *International Journal of Aviation Psychology*, 8(1): 47–63.
- Mussweiler, T.; and Strack, F. 1999. Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology*, 35(2): 136–164.
- Nisbett, R. E.; and Wilson, T. D. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3): 231–259.
- Noddings, N. 1984. *Caring: A Feminine Approach to Ethics and Moral Education*. University of California Press.
- Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.
- Parasuraman, R.; and Manzey, D. H. 2010. Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3): 381–410.
- Parasuraman, R.; and Riley, V. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2): 230–253.
- Pennington, N.; and Hastie, R. 1992. Explaining the evidence: Tests of the story model for juror decision making. *Journal of Personality and Social Psychology*, 62(2): 189–206.
- Rastogi, C.; Zhang, Y.; Wei, D.; Varshney, K. R.; Dhurandhar, A.; and Tomsett, R. 2022. Deciding fast and slow: The role of cognitive biases in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1): 83.
- Saxena, D.; and Guha, S. 2024. Algorithmic harms in child welfare: Uncertainties in practice, organization, and street-level decision-making. *ACM Journal on Responsible Computing*, 1(1): 2.
- Schemmer, M.; Kuehl, N.; Benz, C.; Bartos, A.; and Satzger, G. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 410–422. ACM.
- Shneiderman, B. 2022. *Human-Centered AI*. Oxford: Oxford University Press.
- Skitka, L. J.; Mosier, K. L.; and Burdick, M. 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5): 991–1006.
- Slovic, P.; Finucane, M. L.; Peters, E.; and MacGregor, D. G. 2004. Risk as analysis and risk as feelings: Some thoughts about affect, reason, risk, and rationality. *Risk Analysis*, 24(2): 311–322.
- Stevenson, M. T.; and Doleac, J. L. 2024. Algorithmic risk assessment in the hands of humans. *American Economic Journal: Economic Policy*, 16(4): 382–414.
- Sundar, S. S. 2008. The MAIN model: A heuristic approach to understanding technology effects on credibility. In Metzger, M. J.; and Flanagin, A. J., eds., *Digital Media, Youth, and Credibility*. MIT Press.
- Tetlock, P. E. 1983. Accountability and complexity of thought. *Journal of Personality and Social Psychology*, 45(1): 74–83.

- Tetlock, P. E. 1992. The impact of accountability on judgment and choice: Toward a social contingency model. *Advances in Experimental Social Psychology*, 25: 331–376.
- Toulmin, S. E. 1958. *The Uses of Argument*. Cambridge University Press.
- Tversky, A.; and Kahneman, D. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157): 1124–1131.
- Vaithianathan, R.; Maloney, T.; Putnam-Hornstein, E.; and Jiang, N. 2013. Children in the public benefit system at risk of maltreatment: Identification via predictive modeling. *American Journal of Preventive Medicine*, 45(3): 354–359.
- Vasconcelos, H.; Jörke, M.; Grunde-McLaughlin, M.; Gerstenberg, T.; Bernstein, M. S.; and Krishna, R. 2023. Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 129.