Human Responses to AI Oversight: Evidence from Centre Court

David Almog*, Romain Gauriot†, Lionel Page‡, and Daniel Martin§

June 10, 2026

Abstract

With the increasing capabilities of AI, firms now have the ability to use AI systems to audit and overrule worker decisions, so a central question is whether such oversight would change how humans make decisions. We study how umpires changed their decision-making after the introduction of AI review in professional tennis, a high-stakes setting with objectively correct answers. We find that on close calls umpires reduced their overall mistake rate but shifted the composition of their mistakes. Umpires became more likely to call balls in, substituting more disruptive errors with less disruptive ones. We provide suggestive within-match evidence consistent with public correction as a mechanism: umpires adjust subsequent close calls in the direction that avoids repeating the type of error previously overturned by Hawk-Eye. Our results suggest that AI oversight systems can alter worker behavior and should be evaluated by error composition, not just aggregate accuracy.

*Harvard Business School, dalmog@hbs.edu.
†Deakin University, romain.gauriot@deakin.edu.au.
‡University of Queensland, lionel.page@uq.edu.au.
§University of California, Santa Barbara, danielmartin@ucsb.edu.

1 Introduction

Many organizations now use artificial intelligence (AI) systems to advise workers.[1] A more interventionist use of AI systems is to audit human decisions and block or reverse likely mistakes, as in algorithmic screening of human pricing interventions (Huelden, Jascisens, Roemheld, and Werner 2024) and proposed applications in law and medicine (Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan 2018; Rambachan 2024; Raghu, Blumer, Corrado, Kleinberg, Obermeyer, and Mullainathan 2019). Using AI to overrule human errors appears to be a straightforward improvement for firm performance and societal welfare. However, assessing the full impact of an AI system that audits and reverses human decisions requires understanding whether its presence alters human decision-making. While a large amount of research has focused on how humans respond when assisted by AI, very little is known about how humans respond when their decisions might be overruled by AI. Because being overruled can carry career concerns (e.g., lost reputation or firm sanctions) and career benefits (e.g., avoiding revenue losses) or bring about psychological costs (e.g., shame and embarrassment of being overruled) and psychological benefits (e.g., relief at having their mistakes fixed), individuals might consciously or unconsciously alter their decision-making under AI oversight.

We use the introduction of AI review in umpiring at top tennis tournaments to study human responses to AI oversight.[2] Due to the very small distances involved, camera imaging alone is not sufficient to determine whether a ball lands in or out of bounds, so Hawk-Eye leverages state-of-the-art machine learning algorithms to use the ball's trajectory and spin to reconstruct the landing point.[3] Starting in 2006, players were given the option to challenge a line call, and if the call contradicted the highly accurate predictions of Hawk-Eye, the umpire's decision was overturned, or in some cases, the point was replayed. While this form of AI oversight is initiated by humans, the AI system performs the audit and decides whether to overturn the human decision without any further human input.

Because we have Hawk-Eye data from immediately before the introduction of AI review, we are able to make a direct comparison before and after the introduction of AI review.[4]

[1] For example, see Hoffman, Kahn, and D. Li (2018), Kawaguchi (2020), Agarwal, Moehring, Rajpurkar, and Salz (2023), Grimon and Mills (2025), and Balakrishnan, Ferreira, and Tong (2026).

[2] We refer to any process that uses an AI system to audit and reverse decisions as "AI oversight". AI systems that perform assessments of human output can be used to overturn human decisions, but they can also be used independently as a tool for assessing worker ability or performance (e.g., Avery, Leibbrandt, and Vecci 2023; Almog, Lippman, and Martin 2026).

[3] For similar reasons, Hawk-Eye technology is used across 25 top sports, where camera imaging alone can be insufficient to provide a definitive assessment.

[4] As shown in Table C.1, we have data from one tournament without Hawk-Eye review that was held after its introduction, which we treat as "before" data.

Although this kind of before-and-after comparison can be vulnerable to time-varying confounds, many important factors stayed the same over the period we study: there were no substantial changes to the training and performance assessments of umpires, the pool of umpires, or the positioning and instructions to line judges.[5] Additionally, for "close" calls and the "closest" calls (balls landing within 100 mm or 20 mm of the line, respectively), we do not find a significant difference between the pre and post periods in the locations where balls bounced, in the distances of those bounces from the line, or in whether the balls landed in or out. However, to help mitigate some of the remaining confounding concerns, we include a number of control variables in our regression analysis, such as tournament, match, and point-level factors.

Based on this analysis, we uncover evidence that human decision-making changes under AI oversight. We find that after the introduction of Hawk-Eye review, umpires lowered their overall mistake rate by 8% (1.1 p.p.) on close calls. However, this aggregate decrease masks two important sources of heterogeneity: the distance of the ball from the line and whether the ball was in or out. For balls just outside of the line (within 20 mm), the mistake rate actually *increased* by 34% (8.5 p.p.). This puzzling result is explained by a change in behavior that occurs with AI oversight. We find that for the closest calls, umpires increased the rate at which they called balls in by 14.5% (6.2 p.p.) after the introduction of Hawk-Eye, which produced a shift from making Type II errors (calling a ball out when in) to Type I errors (calling a ball in when out).

We provide further evidence that this shift reflects a behavioral response to being publicly corrected by AI. In a within-match analysis, we examine whether Hawk-Eye reversals earlier in a match predict how umpires call subsequent close balls. Subsequent calls move in the direction that would avoid repeating the previously corrected error: after an in-to-out reversal, umpires become less likely to call later close balls in, while after an out-to-in reversal, they become more likely to do so. The latter response is larger, consistent with umpires responding especially strongly to the error of incorrectly calling a ball out when it was in. Because challenges and reversals are endogenous, we interpret this evidence cautiously. However, they provide within-match evidence consistent with adaptation to public AI oversight.

The increase in calling balls in and the dynamic pattern within matches fits an institutional asymmetry in the Hawk-Eye review protocol.[6] When an incorrect in call is overturned, the correct outcome can usually be implemented directly. When an incorrect out call is overturned, however, the point has already been stopped, so the umpire must decide whether

[5] These institutional details were verified by a leading tennis official at the time of Hawk-Eye's introduction.

[6] In Appendix D, we provide a model of umpire decision-making that shows how asymmetric override costs can shift attention and decision thresholds. We then structurally estimate these costs and use the model to generate counterfactual predictions.

to replay the point or award it to the challenger. These discretionary remedies are more visible and contentious, making the latter type of error especially costly under public AI review. Because the ATP Tour did not use Hawk-Eye to assess umpire performance during the period we study, these costs are unlikely to operate only through formal career concerns; they may also include psychological costs from being publicly corrected.[7]

We believe that our results are most likely to generalize to settings where algorithms perform well, reversals are visible, and some types of correction are more disruptive than others, such as with pricing guardrails, fraud review, content moderation escalations, returns/refunds, customer-service overrides, medical triage protocols, compliance review, and perhaps bail or claims processing.[8]

However, we chose this setting not because it closely resembles others, but because it addresses several main challenges in determining whether AI oversight impacts human decision-making. First, to answer this question it is necessary to have observations on decision-making both with and without AI oversight, but firms may be resistant to share such data given the legal and public relations risks involved. For Hawk-Eye, there was a testing period during which the technology was used solely for broadcast and data recording purposes, without a challenge system in place, allowing us to track umpire performance both before and after the formal introduction of Hawk-Eye review in a tournament.[9] Second, to determine the types of errors people make, we need a clear "ground truth" (an objectively correct answer). In some cases, the closest one can get is by aggregating many expert opinions, such as in diagnostic radiology (Agarwal, Moehring, Rajpurkar, and Salz 2023). Hawk-Eye is precise enough that we can treat it as ground truth without meaningful loss. Third, to get a full picture of the types of errors people make, we need to observe or estimate counterfactual outcomes, so as not to run afoul of incomplete data due to "selective labels" (Rambachan 2024). Fortunately, the tennis setting does not suffer from this issue. Fourth, it is necessary to control or otherwise account for the private information of decision-makers, as this may introduce confounds in analysis. In our setting, there are few sources of private information that would be helpful in assessing close calls. In addition, it is worth noting that this setting offers a binary decision-making problem, as the task is just to match a binary action (call in or out) to a binary state (ball bounces in or out), which greatly simplifies our analyses.

The economics literature on AI and human decision-making, which we review in Section 5, has focused almost exclusively on the paradigm in which humans are provided with AI

[7]This illustrates how insights from behavioral economics can help in understanding human-AI interactions Camerer 2019.

[8]In Appendix B, we provide a more detailed assessment of the external validity (EV) of this setting using the approach proposed in List (2020).

[9]Our officiating source noted the following about the broadcast period: "Even though these replays are not shown on the big screen, they inform the media and TV commentators, which can lead to widespread comments about the umpire being wrong."

recommendations. Our primary contribution is to help take this literature in a new direction by considering a paradigm where AI systems are used to overrule human mistakes. One reason that AI oversight might be less studied is that giving humans final decision rights is more socially acceptable than handing these rights over to AI. In this regard, our study also provides additional perspective on the debate surrounding when formal decision-making authority should be given to humans versus AI (Ide and Talamas 2025; Athey, Bryan, and Gans 2020).

Despite the legal, ethical, and regulatory issues involved, firms might use AI to overrule human mistakes because it has the potential to improve decision quality. In fact, AI oversight appears to meaningfully lower aggregate mistake rates in our setting. Before the introduction of Hawk-Eye review, points were incorrectly awarded 13.9% of the time for close calls, but after the introduction of Hawk-Eye review, points were incorrectly awarded 9% of the time. Of this 4.9 p.p. mistake reduction, 28.6% is due to improvements in umpire performance and the remaining 71.4% is due to AI overruling incorrect human decisions.[10] On the other hand, it has been found that AI recommendations can fail to improve choices in a range of settings. One reason that has been proposed for this is the biased beliefs of human decision makers (Agarwal, Moehring, Rajpurkar, and Salz 2023; Caplin, Deming, S. Li, Martin, Marx, Weidmann, and Ye 2025; Kovach, Martin, and Tserenjigmid 2026). Given the incentives to adopt AI oversight, especially relative to AI recommendations, when and how to allow AI oversight is nearly certain to be an issue that firms, policymakers, regulators, and legal authorities are forced to grapple with in the coming years.

The rest of the paper proceeds as follows. In Section 2, we provide additional details on tennis umpiring, Hawk-Eye review, and the data sources we leverage in our analysis. In Section 3, we examine overall mistake rates, changes in the composition of errors, and within-match evidence on the mechanisms behind these changes. In Section 4, we explore additional sources of heterogeneity. In Section 5, we discuss related literature on human-AI interaction and behavioral economics. Finally, we conclude with a brief summary and discussion of possible future directions in Section 6.

2 Setting and Data

In March 2006, at the Nasdaq-100 Open in Key Biscayne (currently known as the Miami Open), Hawk-Eye review was officially used for the first time at an ATP Tour event.[11] Later

[10] In light of this improvement, it may be unsurprising that Major League Baseball (MLB) has followed the ATP by announcing the adoption of an Automated Ball-Strike (ABS) challenge system beginning in 2026. Like tennis's original Hawk-Eye implementation, the MLB system preserves human officials as the first decision-makers while allowing players to use tracking technology to challenge a limited set of calls.

[11] The ATP Tour is the top tennis tour organized by the Association of Tennis Professionals.

that year, many tennis tournaments that use non-clay surfaces, including the US Open, adopted this new technology by allowing players to challenge calls.[12] Hawk-Eye uses six to ten computer-linked television cameras positioned around the court to collectively create a three-dimensional representation of the ball's trajectory. Hawk-Eye performs with an average error of just 3.6 mm, so just like the ATP Tour we will treat Hawk-Eye readings as if they are the ground truth.[13] While the possibility of error means that Hawk-Eye readings are not perfectly reflective of ground truth, our binning thresholds (20 mm, 100 mm) are far larger than this error rate, so our main qualitative patterns are unlikely to be driven by measurement noise in the ground truth.

2.1 Professional Tennis, Umpires, and Hawk-Eye

Professional tennis is a racket sport that is played either one-on-one (singles) or two-on-two (doubles). In singles tennis, two players compete against each other on opposite sides of the tennis court. The objective is to score points by hitting the ball "in" (within the bounds of the opponent's side of the court). We will concentrate solely on men's singles matches (singles matches where both players are men) because the analysis of other formats is underpowered in our data.[14] The scoring system for tennis is based on a series of points, games, and sets. Players accumulate points to win a game and accumulate games to win a set. Matches are typically played as the best of three or five sets, with each set requiring a player to win at least six games.

A crew of up to ten umpires is involved in officiating a match. The roles typically include one chair umpire who oversees the entire match, making decisions on points, penalties, and overall match control. Additionally, there are usually nine line umpires positioned around the court, each responsible for specific lines, with the sole duty to determine whether the ball bounced in or out when it is relatively close to their respective line.[15] The chair umpire has the authority to overrule the decisions of the line umpires if necessary, so drawing inferences on the performance of an individual umpire is complicated, especially since we do not have data on whether the chair umpire overruled an individual line umpire. Thus, in this paper, we evaluate the performance of the umpire crew as a whole. An additional issue is that we do not have data on the chair umpire or umpiring crew of a particular match, so we cannot perform a within-umpire/crew assessment of the impact of AI review. However, even if we

[12] Players have a fixed number of challenges that they can make per match, but successful challenges do not count against this limit. Players were not allowed to challenge chair umpire decisions on non-clay courts before this system was put in place.

[13] For perspective, the standard diameter of a tennis ball is 67 mm.

[14] If more data were available on the other formats, it might be of interest to examine whether the impacts of AI oversight vary by gender or team composition.

[15] Figure C.2 provides a visual representation of the positioning of each umpire.

had such data, it is likely that umpire-specific assessments would be underpowered given that we focus on close calls.

International chair umpires are certified with Gold, Silver, or Bronze badges, while line umpires are graded according to national or other systems. Certification and grading methods remained unchanged during the study period, and Hawk-Eye metrics were not incorporated into these evaluations for the first couple of years. A leading tennis official who was involved in testing and introducing Hawk-Eye provided valuable institutional insights that will be used throughout the paper.

The Hawk-Eye review protocol works by endowing players with 2-3 challenges per set.[16] If a challenge is successful, players do not lose a challenge opportunity. When a player challenges a call, a computerized path and the final landing location of the ball are displayed on a large screen in the stadium for the umpire, players, and the crowd to observe the outcome of the challenge. The public nature of the challenge process adds excitement for the spectators, but simultaneously adds pressure to the umpires, as their decisions are being publicly scrutinized.

An important component of the review system implementation is the asymmetric resolution of incorrect calls, which varies depending on whether the challenged call was initially in or out. If a ball is initially ruled in by the umpires and the challenge successfully overturns the call, then the point ends with the challenger winning the point and the correct outcome being enforced. However, when a ball is initially ruled out by the umpires, but the review shows otherwise, enforcing the correct outcome is not always possible because the point was unnecessarily stopped. In this case, the umpire has to make an arbitrary decision on whether the opponent of the challenger had a real chance to return the ball. If the answer is yes, the point has to be replayed from scratch. This situation can be perceived as detrimental due to the inability to implement the correct outcome (continuing the point from where it stopped) and because it forces umpires to make arbitrary decisions that, in multiple instances, have been shown to infuriate one of the players involved.

2.2 Data

While this paper is not the first to utilize tennis data for drawing inferences on decision-making, we have assembled a novel and rich data set that, for the first time, enables us to comprehensively assess the performance of professional tennis umpires. To achieve this goal, we utilized three distinct data sources. Our primary data set, the *Hawk-Eye Base* data set, encompasses precise information on points and, crucially, the location of every bounce for

[16] In practice, players very rarely exhaust their challenges. This is advantageous for our research question, as umpires are almost always under the threat of being challenged.

over 100 tournaments.[17] Our second data set, the *Challenge* data set, tracks the outcomes of challenges during a sample of tournaments that took place in the first few years following the implementation of Hawk-Eye review in professional tennis. In the period of time where challenges are used, the first data set only permits drawing conclusions on players' behavior because, when observing a correct call, it is not possible to disentangle whether the umpire made the correct call initially or if a Hawk-Eye review corrected the umpire's incorrect call. In contrast, the second data set allows precise identification of incorrect calls, as every won challenge is, by definition, an admission of the umpire's mistake. Nonetheless, this second data set is incomplete because it only includes points that players decided to challenge, and this selection is susceptible to players' perceptual and behavioral biases. Merging the first two data sets was challenging due to systematic inconsistencies in the Challenge data set (e.g., the score was often flipped across players and sometimes missing). To overcome this limitation, we turned to a third data source: a manual review of points by replaying match videos. This allowed us to identify the ground truth for challenges in a subset of matches, enabling us to determine an effective approach for merging the first two data sets.[18] In the end, we were able to merge 2,038 out of the 2,108 challenges registered for those 28 tournaments (a 97% merge rate).

The consolidated data set produced by merging the Hawk-Eye Base data set and the Challenge data set comprises a total of 698 matches across 35 distinct tournaments, and summary statistics for this data set can be found in Table 1.[19] This data set includes 109 matches from seven tournaments that were played before Hawk-Eye review and 589 matches from 28 tournaments after Hawk-Eye review was active. The pre- and post-Hawk-Eye samples are very similar along the main point-level dimensions most relevant for identification, especially the shares of bounces within 20 mm and 100 mm of the line. The differences in tournament and round composition motivate the inclusion of match-level controls in the analyses below. Figure C.1 shows that the distribution of players' strokes was highly similar before and after the introduction of Hawk-Eye review, both in proximity to the line and in speed.

We will now offer a more comprehensive description of the composition and role of each of the three data sources.

[17] We excluded from the analysis the 15 clay court tournaments. We elaborate on this decision in the Appendix A.1.

[18] The video auditing process was also instrumental in validating the best rules for determining when a call was made incorrectly.

[19] Table C.1 provides information on the month, category, court type, and the number of matches for each tournament.

	Before Hawk-Eye review (PostHK=0)	After Hawk-Eye review (PostHK=1)	Total	Post - Pre [Std. diff.]
A. Sample counts				
Players	71	161	174	
Tournaments	7	28	35	
Matches	109	589	698	
Points	15,435	83,902	99,337	
Bounces	71,630	402,467	474,097	
B. Tournament tier (share of tournaments)				
ATP 250	28.6%	42.9%	40.0%	+14.3 p.p. [0.30]
ATP 500	0.0%	10.7%	8.6%	+10.7 p.p. [0.49]
ATP 1000	71.4%	46.4%	51.4%	-25.0 p.p. [0.53]
C. Match round (share of matches)				
Final	6.4%	4.6%	4.9%	-1.8 p.p. [0.08]
Semifinal	12.8%	9.0%	9.6%	-3.8 p.p. [0.12]
Quarterfinal	17.4%	14.9%	15.3%	-2.5 p.p. [0.07]
Round of 16	12.8%	19.2%	18.2%	+6.3 p.p. [0.17]
Other	50.5%	52.3%	52.0%	+1.8 p.p. [0.04]
D. Bounce composition (share of all bounces)				
< 20 mm from the line	0.78%	0.69%	0.70%	-0.09 p.p. [0.01]
< 100 mm from the line	3.66%	3.53%	3.55%	-0.13 p.p. [0.01]
Serves	29.2%	27.6%	27.8%	-1.64 p.p. [0.04]
Non-serves	70.8%	72.4%	72.2%	+1.64 p.p. [0.04]
E. Average speed in km/h (s.d.)				
Serves	147.7 (24.2)	150.0 (23.1)	149.6 (23.3)	+2.3 [0.10]
Non-serves	81.7 (21.4)	83.3 (20.4)	83.1 (20.5)	+1.6 [0.08]
All	101.0 (37.4)	101.7 (36.5)	101.6 (36.6)	+0.7 [0.02]

Table 1: Summary statistics for the consolidated data set. The last column reports the difference in the sample characteristics before and after Hawk-Eye review. Bracketed values are absolute standardized differences.

2.2.1 Hawk-Eye Base Data Set

The main source of data for this paper is the Hawk-Eye Base data set, which consists of the official Hawk-Eye data for matches played at the international professional level between March 2005 and March 2009. This data set provides a number of pieces of information for each point, including the position of every bounce captured by the Hawk-Eye system during that point, the serving player, the ongoing score, and the point winner. Altogether, this data set includes information on 1.8 million bounces from more than 1,800 men's singles matches, spanning various prestigious tournaments such as Grand Slam tournaments, Masters 1000 tournaments, and International series tournaments. This data set has been used in the past to study important economics questions such as risk management (Ely, Gauriot, and Page 2017), tournament incentives (Gauriot and Page 2019), and mixed strategy play (Gauriot, Page, and Wooders 2023). These papers have focused on the players' perspective, and the Hawk-Eye Base data set is well-suited for this purpose. However, this data set alone is insufficient to analyze umpires' performance, as it does not permit us to identify whether the final call comes from the umpire or via an overturned challenge.

2.2.2 Challenge Data Set

The Challenge data set was originally obtained from the ATP Tour and encompasses all challenges recorded during 35 tournaments across the initial three years following the introduction of Hawk-Eye (2006-2008). It captures comprehensive information on 2,784 challenges, including details about the players involved, the match itself, the specific point in the match when the challenge was made, and the outcome of each challenge. Of the 35 tournaments, we use the 28 that also appear in our Hawk-Eye Base data set. In order to compile this data set, Abramitzky, Einav, Kolkowitz, and Mill (2012) undertook the arduous effort of collecting and compiling umpire match sheets. Abramitzky, Einav, Kolkowitz, and Mill (2012) acknowledge the selection problem involved in only observing the challenged points (around 2.6% of total points). Nevertheless, they are able to draw inferences on the optimality of decision-making from the players' standpoint by relying on the idea that a higher propensity to challenge implies challenging less conservatively. However, without the point-by-point data, it is hard to assess the umpire's performance once Hawk-Eye review was active. By merging the first two data sets, we solve the aforementioned selection problem and gain information on what was originally called by the umpire crew.

2.2.3 Video Auditing

In order to assess the quality of our merging algorithm, we audited full-match video replays using TennisTV, the official ATP streaming service. We identified 43 matches that appeared

in both the Hawk-Eye Base and Challenge data sets, providing us with an assessment of how well a given merging algorithm would work in the 144 challenges witnessed during these matches.

Using variables such as match, point, distance, and who hit the ball, we developed a merging algorithm with a 99.3% merge rate, for which only 4.9% of the video-audited challenges were merged to an incorrect point. The algorithm consists of eight iterations, starting with the most stringent merging rule and gradually relaxing criteria for challenges that persist without merging.[20]

As an additional benefit, auditing match replays enabled us to test the effectiveness of the four criteria we implemented to identify incorrect calls (two criteria for each type of mistake). We deem an in call to be incorrect when a player's stroke is recorded as bouncing out in the Hawk-Eye Base data set, yet they still win the point, or if the data set records at least three more strokes after an out bounce. We deem an out call to be incorrect if all the strokes of a point are recorded as in, and the player delivering the final hit loses, or if there is a second serve after the first one was recorded as in the Hawk-Eye Base data set.[21]

3 Empirical Results

In this section, we use our consolidated data set to study the impact of AI oversight on umpiring decision-making in professional tennis. We begin by examining changes in the overall mistake rate for all hits. We then highlight the influence of distance from the line and whether a ball was in or out.

3.1 Overall Mistake Rate

Before Hawk-Eye review was introduced, the umpire mistake rate was only 0.61% of all ball bounces. However, if we look at the 2,622 (3.6%) bounces that fell within 100 mm of the line, the mistake rate jumps to 13.89%. Looking at the 556 (0.78%) bounces that fell within 20 mm of the line, the mistake rate jumps even more to 32.91%. Given our interest in the impact of AI oversight on human mistakes, our primary focus will be on calls within 100 mm or 20 mm of the line.[22]

While the likelihood of a close call (a ball bouncing within 100 mm or 20 mm of the line) did not change substantially after the introduction of Hawk-Eye and the probability of a

[20] A detailed explanation of the merging algorithm is provided in Appendix A.2.

[21] Appendix A.3 documents these criteria further.

[22] We also find that over 93% of the challenges in our sample are for calls within 100 mm of the line.

close call being in or out did not change substantially either,[23] we find that the mistake rate on close calls did change substantially. We estimate the effect of AI oversight on umpires' performance using the following specification:

$$\mathbb{1}(Incorrect\,Call)_{ipm} = \alpha_0 + \alpha_1 \boldsymbol{PostHK_m} + \beta X_p + \gamma Y_m + \epsilon_{ipm} \tag{1}$$

$\mathbb{1}(Incorrect\,Call)$ is an indicator variable equal to 1 if the call i made by the umpire in point p in match m was incorrect. $\boldsymbol{PostHK}$ is an indicator variable equal to 1 if the match m belongs to a tournament with the Hawk-Eye review active. X is a vector of point characteristics: distance fixed effects (by 20 mm bins), speed (whether it is below or above the median), score, game, set and an indicator if the point played is in the tie-break stage.[24] Y is a vector of match characteristics that includes round and tournament tier.[25]

Our main coefficient of interest is α_1, which can be interpreted with OLS regression as the effect in percentage points of AI oversight on the probability that an umpire's call is incorrect. Table 2 reports the results of estimating this equation for calls within 100 mm of the line. In our primary specification, the mistake rate is estimated to decrease by 8% (1.1 p.p.). This decrease in the mistake rate is consistent with a model of costly attention in which the costs of being overruled by the AI outweigh any benefits (see Appendix D for a model of this form).

Importantly, the average effect of Hawk-Eye review is fairly stable across specifications. Adding point and match controls changes the estimate only from -1.4 to -1.1 percentage points, so the main change with added controls is a loss of precision rather than a meaningful loss in magnitude. The noise in these estimates is unsurprising because the single $PostHK$ coefficient averages over sharply heterogeneous effects by distance to the line and by which side of the line the ball lands on. First, an important component of call difficulty is the distance from the line, so the ability to change attention enough to fix a mistake might depend on the distance from the line. Second, it might matter which side of the line the ball is on, and umpires might have different perceptions of the cost of Type I and Type II errors under Hawk-Eye review. As subsequent figures show, mistake rates fell for many close calls after the introduction of Hawk-Eye review but increased for the narrow region just outside the line, so the overall average masks offsetting effects.

[23] See Table 1 for the likelihood of a close call. The probability that a ball was out when it bounced within 100 mm of the line was 45.3% before Hawk-Eye review and 46.8% after, and the corresponding numbers for 20 mm are 48.4% and 48.9%.

[24] A tie-break is a one-off game held to decide the winner of a set when two players are locked at 6-6.

[25] Given that the period without oversight has no 500-tier tournaments, we aggregate the 250 and 500 groups, so we basically control for whether the match is a Masters 1000 tournament or not.

	(1)	(2)	(3)	(4)
	Incorrect call			
PostHK	-0.014**	-0.011	-0.011	-0.011
	(0.007)	(0.007)	(0.007)	(0.007)
Point controls		X	X	X
Match controls			X	X
Cluster level				Match
N	16,812	16,812	16,335	16,335
Baseline mean	.139	.139	.136	.136

Standard errors in parentheses

* $p < .10$, ** $p < .05$, *** $p < .01$

Table 2: OLS regressions of umpire mistakes for balls bouncing within 100 mm of the line. $PostHK = 1$ if the Hawk-Eye review is active. Point controls include distance fixed effects (by 20 mm bins), speed (whether it is below or above the median), score, game, set, and an indicator if the point played is in the tie-break stage; and match controls include round and tournament tier.

3.2 Shift in Type I and Type II Errors

Figure 1 illustrates the relationship between mistake rates and the distance to the line for the 16,812 bounces within 100 mm of the line. In this figure, each dot represents the rate at which umpires made an incorrect call for all balls bouncing within one of ten 20 mm bins. The five bins to the left of the dashed line correspond to balls that landed out of bounds, while the five bins to the right side of that line represent balls that landed inside of the line. For the rest of the paper, negative distances will be used to denote the distance to the line for balls that bounced outside.

The blue dots in the figure were calculated using tournaments before the introduction of Hawk-Eye review. The red dots represent tournaments with Hawk-Eye review. While it is expected that umpires' performance decreases as the ball gets closer to the line, we can also observe that the red line lies mostly under the blue line (this holds true for eight of the ten bins). This mirrors the results in Table 2, which shows that the overall performance of umpires improves after the introduction of AI oversight, and in Table 4a, which shows that this is particularly true for balls bouncing between 100 and 20 mm away from the line. In addition, it is noticeable that the only region of the court where the introduction of Hawk-Eye increased the rate at which umpires make mistakes is the two closest bins on the outside of the line.

To further analyze this change in the type of errors made, we examine how the impact of AI oversight on umpire performance varies with distance to the line. We re-estimate equation 1 (including point and match controls), but now we interact $PostHK$ with the

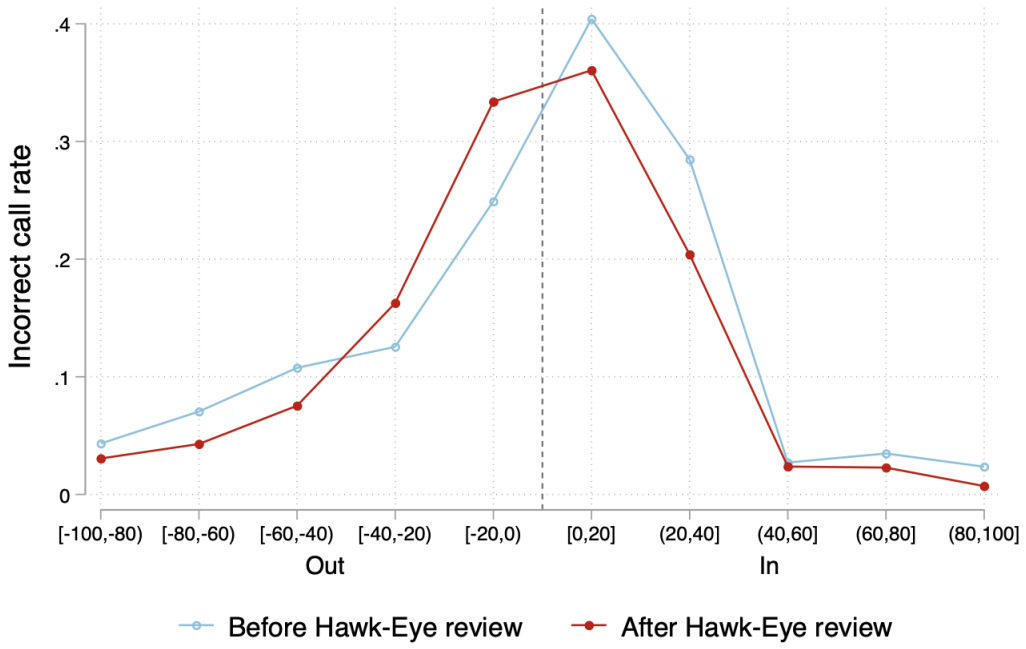


Figure 1: Incorrect call rates by proximity to the line. Each dot is the rate of incorrect calls for a bin of 20 mm. Dots to the left of the dashed line represent bins out of bounds, and those to the right of the dashed line represent bins in bounds.

indicator variables for each of the ten 20 mm distance bins into which the ball bounced. We report the estimated coefficients of these interaction terms graphically in Figure 2. For each bin, we plot the estimated coefficient as a dot and provide the respective 95% confidence interval around the estimated coefficient. The red dashed line, which separates the in and out parts of the court, shows a sharp discontinuity. The first bin to the left of the red dashed line (balls bouncing just out) exhibits the most significant positive increase in incorrect calls, with an estimated coefficient of 8.6 percentage points (significant at the 1% level). From that point onward, the coefficients gradually adjust back down to the average treatment effect, just below zero. In the first two bins to the right of the red dashed line (balls bouncing in), we observe the most substantial decrease in incorrect calls, with estimated coefficients of -4.1 (significant at the 5% level) and -7.7 (significant at the 1% level) percentage points. Similarly, as we continue to move to the next bins in that direction, the estimated coefficients gradually increase to the average treatment effect level. As highlighted by Goldsmith-Pinkham, Hull, and Kolesár (2024), our estimates from Figure 2 may suffer from "contamination bias" due to the regression of multiple treatments while controlling for observables in an additively separable manner. To address this concern, we implement one-treatment-at-a-time regressions, one of the suggested solutions by Goldsmith-Pinkham, Hull, and Kolesár (2024). Figure C.3 shows that our original estimates do not suffer from

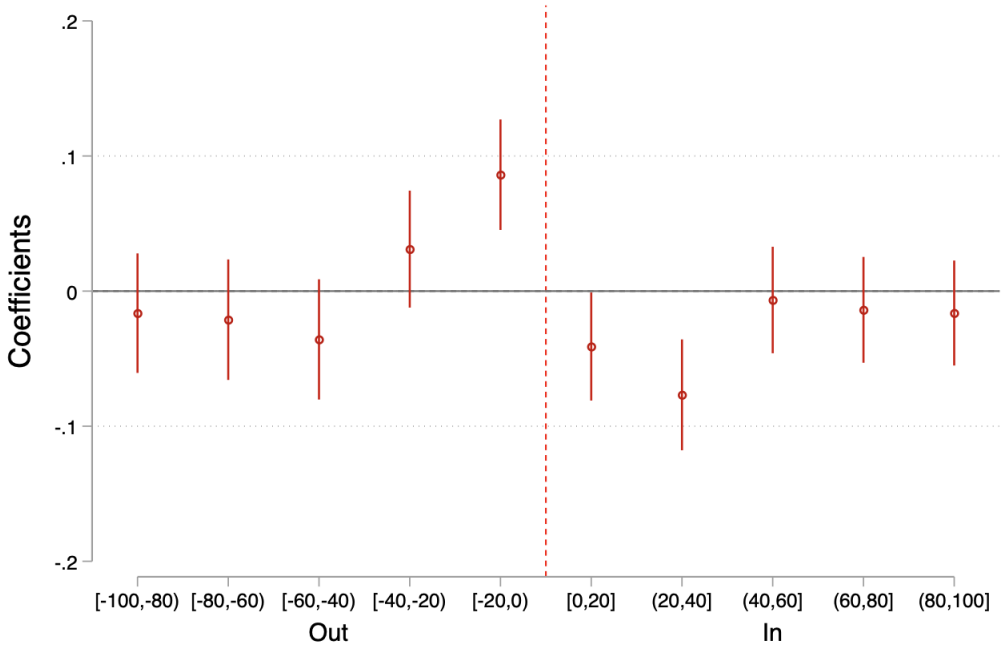


Figure 2: Impact of AI oversight on the incorrect call rate by proximity to the line. Each dot represents the coefficient on the interaction between the distance bin and PostHK, the indicator variable that equals 1 if the Hawk-Eye review system is active. Controls for point and match characteristics are included.

this type of bias.

We provide an event-study version of Figure 2 in Figure C.4. This figure illustrates how the impact of AI oversight on umpire performance progresses over time by separately examining matches in the first half and second half of the period we study. The first group comprises tournaments from 2006 and the first semester of 2007, while the second group includes tournaments from the second semester of 2007 and those from 2008. Figure C.4 suggests that the shift in umpires' error types took over a year to fully manifest. We chose this division into two time sub-periods to yield a similar number of observations in both groups, and the results are robust to breaking the tournaments into more time sub-periods, though the effects become more noisily estimated.

The key shift in errors occurs for the closest calls. What drives this shift? Figure 3 plots regression-adjusted call-in rates for balls landing within 20 mm of the line, regardless of which side of the line they actually bounced on. The figure estimates separate linear trends before and after the introduction of Hawk-Eye review, controlling for point, tournament, and shot characteristics. It reveals a shift toward more inside calls after the introduction of AI review:

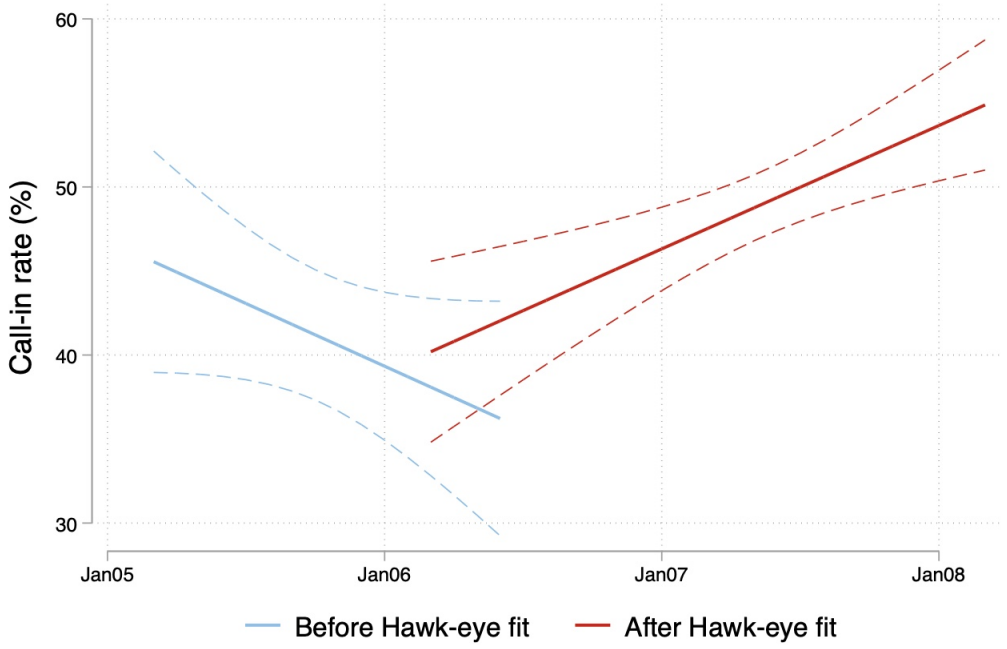


Figure 3: Call-in rates before and after the introduction of Hawk-Eye review for balls landing within 20 mm of the line, regardless of which side of the line they actually bounced on. The solid lines report fitted values from a linear probability model in which the probability of calling a ball in is allowed to vary linearly with month, separately before and after Hawk-Eye review. The dashed lines indicate 95% confidence intervals. The regression controls for game, set, point, tiebreak, round, tournament tier, and shot speed. Standard errors are clustered at the match level.

from 42.8% to 49.0%.[26] In addition, the increase appears gradual. Before Hawk-Eye review, the fitted monthly trend in call-in rates was negative, declining by 0.62 percentage points per month. After Hawk-Eye review, the trend reversed: call-in rates increased by approximately 0.61 percentage points per month ($p = 0.001$). The difference between the pre- and post-Hawk-Eye slopes is 1.23 percentage points per month ($p = 0.002$), indicating a statistically significant change in trend after the introduction of AI review. The next subsection provides within-match evidence consistent with one mechanism behind this gradual shift: umpires appear to adjust their subsequent calls after being publicly corrected by Hawk-Eye.

3.3 Possible Mechanisms

Several mechanisms could, in principle, contribute to the behavioral responses we document following the introduction of Hawk-Eye review. One possibility is institutional guidance

[26]The fact that balls landing close to the line are more often called out is consistent with evidence from the visual perception literature on moving objects Whitney, Wurnitsch, Hontiveros, and Louie (2008).

or supervisory pressure to adopt a particular default for close calls. However, the leading official we contacted, who played a key role in the deployment of Hawk-Eye at the ATP Tour, confirmed that umpires were not explicitly instructed to make more in calls for close calls. A second possibility is strategic adaptation driven by asymmetric challenge exposure. In our setting, challenge rates are similar across error types for close calls (Type I: 41.5%, Type II: 44.9%), suggesting that differential challenge likelihood is unlikely to be the primary driver of the shift in the types of errors we observe. Finally, one might think that the effects reflect information revelation alone, with Hawk-Eye providing new information about umpire performance. Yet even prior to the stadium-based review regime, incorrect calls were already visible to outside observers through broadcast coverage. The key change with Hawk-Eye is the public, on-the-spot overturning of on-court decisions, which may itself generate reputational or psychological costs. Taken together, these considerations point away from explanations based solely on directives, sharply asymmetric review exposure, or information revelation alone.

Another possibility is that the on-the-spot overturning of on-court decisions may have made the psychological costs of being overturned stronger or more salient. To provide evidence more closely tied to this channel, we examine whether public reversals are followed by changes in subsequent line calling within the same match. We focus on balls landing within 20 mm of the line and estimate whether the probability that a close ball is initially called in varies with the number of earlier Hawk-Eye reversals in the match. We distinguish between two types of reversals: in-to-out reversals, in which an initial in call is shown by Hawk-Eye to have been out, and out-to-in reversals, in which an initial out call is shown by Hawk-Eye to have been in.

We regress whether the ball was initially called in on the number of in-to-out reversals and out-to-in reversals using match fixed effects to hold fixed the match-specific officiating environment and help separate short-run responses to public correction from slower-moving changes in call tendencies across tournaments. Figure 4 shows the estimated coefficients on the number of in-to-out reversals and out-to-in reversals. These coefficients indicate that subsequent calls move in the direction that would avoid repeating the previous publicly corrected error. Following prior in-to-out reversals, umpires become less likely to call later close balls in. Following prior out-to-in reversals, they become more likely to do so. The latter response appears larger: after one prior reversal, the increase in call-in propensity after an out-to-in reversal is larger than the decrease after an in-to-out reversal. This asymmetry is consistent with the idea that umpires respond especially strongly to the error that becomes most costly and salient under Hawk-Eye review: incorrectly calling a ball out when it was in.

Because challenges and reversals are endogenous, these estimates should be read with caution and not as a stand-alone causal test. Still, this pattern suggests a possible mechanism

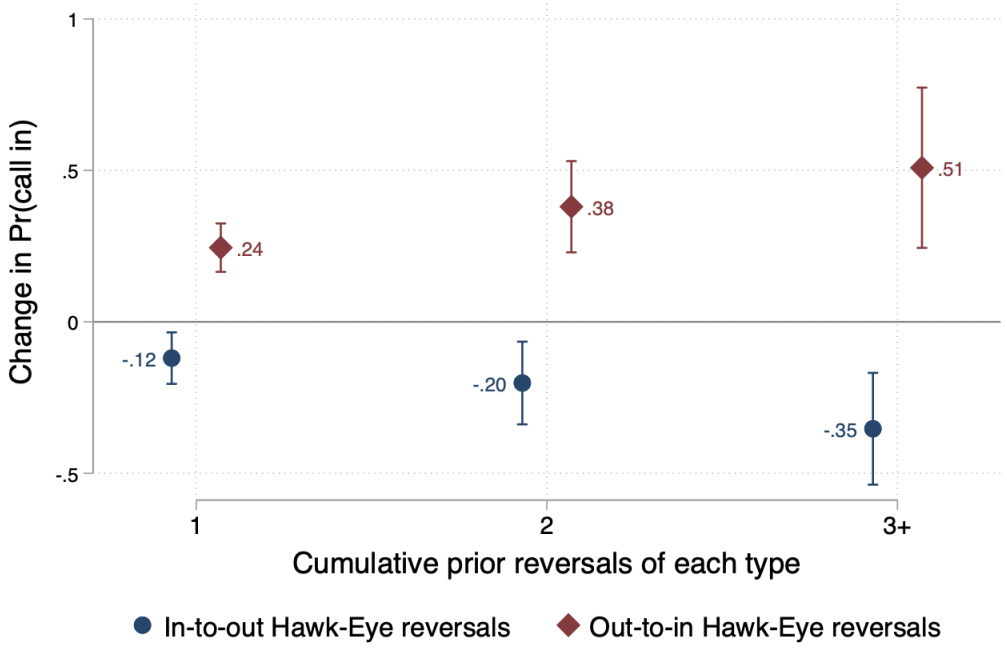


Figure 4: The figure reports coefficients from a linear probability model estimated on post-Hawk-Eye calls for balls landing within 20 mm of the line. The dependent variable equals one if the ball is initially called in. The omitted category for each series is zero prior Hawk-Eye reversals of that type (in-to-out or out-to-in) within the same match. Cumulative prior reversals are capped at three, with the third category pooling three or more reversals. The specification includes match fixed effects and point-level controls. Standard errors are clustered at the match level. Bars denote 95% confidence intervals.

for the increase in the rate of calling balls in for the closest calls that is shown in Figure 3. While a general drift towards calling balls in could explain a higher average call-in rate, it would not naturally predict opposite within-match adjustments after different types of Hawk-Eye reversals. The direction and asymmetry of the responses are consistent with public reversals becoming costly under AI oversight, and especially so when the original call incorrectly stopped play by calling a ball out.

4 Examining Heterogeneity

In this section, we study three main sources of heterogeneity: skill, stakes, and type of perceptual task. Practically, this relates to differences by tournament stage (finals or earlier stage), tournament tier (higher or lower tier), and shot type (serve and non-serve).

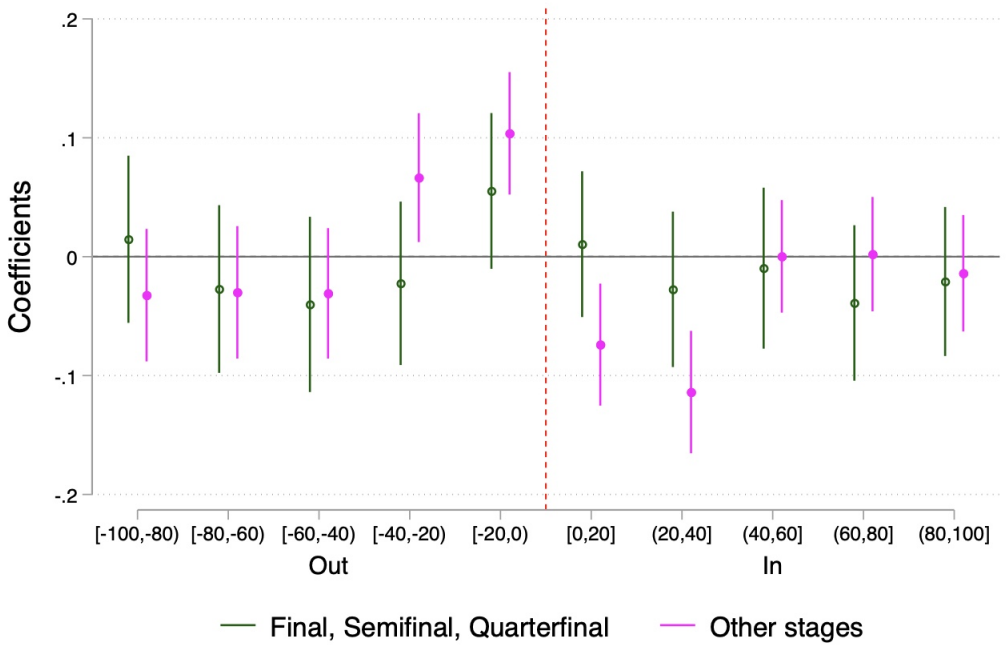


Figure 5: Impact of AI oversight on the incorrect call rate by proximity to the line (by tournament stage). Matches are grouped into those at the final, semifinal, and quarterfinal stages and those at all other (earlier) stages of a tournament and then analyzed separately. Each dot represents the coefficient on the interaction between the distance bin and PostHK, the indicator variable that equals 1 if the Hawk-Eye review system is active.

4.1 Tournament Stage and Tier

Umpires are rewarded for good performance with assignments to advanced-stage matches, as the pool of officials needed becomes smaller and organizers can be more selective as the tournament progresses. These selection dynamics within tournaments are likely to create variation in umpiring skill across tournament stages.

To examine whether the impact of AI oversight differs by tournament stage, we split our sample into two groups: the first group includes Final, Semifinal, and Quarterfinal matches, where we expect to find the most skilled umpires, while the second group contains the remaining matches. In Figure 5, we examine the AI oversight effect separately on these two groups of matches. The change in the type of errors in bins closer to the line is more pronounced and only significant for matches in the early stages, as highlighted by the magenta markers. The estimated effects for advanced-stage matches (green markers) are not statistically significant for any bin.

In a study on radiologists, Chan, Gentzkow, and Yu (2022) found that aversion to false negatives tends to be negatively related to radiologist skill. Specifically, lower-skilled radiol-

ogists are more likely to falsely diagnose a healthy patient with pneumonia (false positive) because this error is less costly than missing a diagnosis. Our findings align with those of Chan, Gentzkow, and Yu (2022), as less-skilled umpires are more likely to avoid the more costly type of error (calling a ball out when it is actually in).

Later-stage matches also involve higher stakes, raising the possibility that the observed effects reflect stakes rather than differences in umpire skill. We assess the impacts of stakes separately by comparing higher-tier tournaments, such as Masters 1000 events, with lower-tier tournaments, such as ATP 500 and 250 events. As our officiating expert noted, umpire tournament assignment decisions depend on factors beyond tier, including geography, budget, availability, and logistical constraints, and there is substantial overlap in officials assigned across tournament levels. Thus, the officiating pool is broadly comparable across tiers.

Figure C.5 shows that the patterns we observe between early- and later-stage matches do not appear when comparing higher- and lower-tier tournaments. If stakes were driving the heterogeneity in Figure 5, we might expect similar patterns across tournament tiers, but we do not. Taken together, these results suggest that differences across tournament stages are not driven by differences in match stakes.

4.2 Serves and Non-Serves

There is value in distinguishing between serves and non-serves,[27] as we can think of them as distinct tasks for the umpires. The primary differences, from an officiating standpoint, between serves and subsequent hits lie in speed and anticipation. Serves may be considered more challenging due to their significantly higher speed: the average speed for a serve in our sample is 150 km/h, compared to 83 km/h for non-serves.[28] However, serves have the advantage of anticipation: umpires know the ball is about to be served and will most likely bounce in a very specific region of the court (the serving box). This appears to be important: when comparing the slowest quartile of serves (132 km/h) with the fastest quartile of non-serves (97 km/h), we find that the error rates for serves are lower, despite still containing faster strokes. The average incorrect rate during the period without Hawk-Eye review for the first quartile in serves and the fourth quartile in non-serves is 11.6% versus 15.6% for balls within 100 mm of the line, and 25.7% versus 40% within 20 mm of the line. Therefore, even though serves are much faster than non-serves, there is compelling evidence to support that the anticipation advantage benefits umpires during serves.

In Figure C.7, we break down the incorrect call rates by serves and non-serves. This figure shows that the shift in types of errors is present in both serves and non-serves. For

[27] The serve starts off every point in tennis, with players alternating serving each game.

[28] Figure C.6 shows the speed distribution for both types.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Incorrect call (Serves)				Incorrect call (Non-serves)			
PostHK	-0.006	-0.002	0.000	0.000	-0.024**	-0.021**	-0.023**	-0.023**
	(0.010)	(0.009)	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)	(0.011)
Point controls		X	X	X		X	X	X
Match controls			X	X			X	X
Cluster level				Match				Match
N	9,253	9,253	8,990	8,990	7,559	7,559	7,345	7,345
Baseline mean	.139	.139	.137	.137	.138	.138	.136	.136

Standard errors in parentheses

* $p < .10$, ** $p < .05$, *** $p < .01$

Table 3: OLS regressions of umpire mistakes for balls bouncing within 100 mm of the line separately for serves and non-serves. $PostHK = 1$ if the Hawk-Eye review is active. Point controls include distance fixed effects (by 20 mm bins), speed (whether it is below or above the median), score, game, set, and an indicator if the point played is in the tie-break stage; and match controls include round and tournament tier.

serves, the dominant effect is the shift in errors. For non-serves, the dominant effect is a decrease in mistakes. This suggests that non-serves may have a higher potential for wholesale improvement, which could be due to their more manageable speed and the opportunity they provide to rectify mistakes caused by distraction or sparse attention.

Table 3 reports the results from estimating equation 1 separately for serves and non-serves that bounced within 100 mm of the line. In our main specification, we do not find evidence that AI oversight impacted umpire performance during serves. However, we do find evidence that AI oversight corresponds to a 2.3 percentage point decrease in incorrect calls for non-serves. This corresponds to a 17% reduction compared to the baseline mean level of 0.136.

If we restrict attention to calls in which the ball bounces between 20 and 100 mm from the line, Table 4a shows that there is a statistically significant reduction in mistakes for both serves and non-serves. However, we find something different if we look at the closest calls. In Table 4b, we present the estimates for equation 1 when restricting the sample to bounces within 20 mm. This subset of calls is arguably the most complicated for umpires, and improving performance further would entail excessive cognitive costs. For these calls, the introduction of Hawk-Eye resulted in a 7.3 percentage point increase in incorrect calls for serves, representing a 22.9% increment from the baseline. While this result focuses on a very specific subset of the sample (serves within 20 mm), it provides evidence of the types of situations that can lead AI oversight to backfire.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Incorrect call (Serves)				Incorrect call (Non-serves)			
PostHK	-0.020**	-0.018**	-0.017**	-0.017**	-0.012	-0.017**	-0.020***	-0.020**
	(0.008)	(0.008)	(0.008)	(0.009)	(0.009)	(0.009)	(0.009)	(0.011)
Point controls		X	X	X		X	X	X
Match controls			X	X			X	X
Cluster level				Match				Match
N	7,499	7,499	7,238	7,238	6,047	6,047	5,785	5,785
Baseline mean	.092	.092	.091	.091	.082	.082	.080	.080

Standard errors in parentheses

* $p < .10$, ** $p < .05$, *** $p < .01$

(a) For balls bouncing 20-100 mm away from the line.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Incorrect call (Serves)				Incorrect call (Non-serves)			
PostHK	0.063**	0.063**	0.073**	0.073**	-0.035	-0.037	-0.031	-0.031
	(0.031)	(0.031)	(0.032)	(0.030)	(0.031)	(0.031)	(0.033)	(0.034)
Point controls		X	X	X		X	X	X
Match controls			X	X			X	X
Cluster level				Match				Match
N	1,804	1,804	1,752	1,752	1,512	1,512	1,470	1,470
Baseline mean	.325	.325	.319	.319	.333	.333	.325	.325

Standard errors in parentheses

* $p < .10$, ** $p < .05$, *** $p < .01$

(b) For balls bouncing within 20 mm of the line.

Table 4: OLS regressions of umpire mistakes (separately for serves and non-serves). $PostHK = 1$ if the Hawk-Eye review system is active; point controls include distance fixed effects (by 20 mm bins), speed (whether it is below or above the median), score, game, set, and an indicator if the point played is in the tie-break stage; and match controls include round and tournament tier.

5 Related Literature

Our primary contribution is to a growing literature on human-AI interaction in workplaces, and a secondary contribution is to the behavioral economics literature. We discuss our relationship to those two literatures in this section.

5.1 Human and AI Interaction

In the coming years, firms will have the option to use AI systems to audit and overrule worker decisions across a wide array of settings. This is due to the confluence of two forces. First, machine learning algorithms are becoming increasingly good at predicting human mistakes. This includes mistakes made by bail judges (Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan 2018), radiologists (Rajpurkar, Irvin, Zhu, Yang, Mehta, Duan, Ding, Bagul, Ball, Langlotz, et al. 2017; Topol 2019), and workforce professionals who have hiring and promotion responsibilities (Chalfin, Danieli, Hillis, Jelveh, Luca, Jens, and Mullainathan 2016). Second, there has been a substantial drop in the cost of monitoring worker behavior, brought about by the rise in digitization.[29]

For a concrete example of AI oversight, consider Zalando, a leading e-commerce company specializing in fashion retail. Zalando allows category managers to suggest discounts based on their private information about fashion trends. Huelden, Jascisens, Roemheld, and Werner (2024) show that while human interventions appear promising, the negative impact of poor interventions (stemming from overconfidence and other biases) offsets the benefits of good interventions (based on private information). However, they show that using algorithmic tools to predict and block undesirable human interventions has the potential to recoup these gains. In fact, recent papers in the economics literature have shown that there are large potential gains from using AI to overrule human mistakes in other high-stakes settings, including law (Rambachan 2024) and medicine (Raghu, Blumer, Corrado, Kleinberg, Obermeyer, and Mullainathan 2019).

While AI is increasingly good at prediction, humans are kept in the decision process (kept "in the loop") because of social reasons (tradition, comfort, fairness, etc.), due to labor market concerns (union power, contractual responsibilities, inequality considerations, etc.), because of the costs of implementing AI systems, and because humans can sometimes perform better by incorporating additional information, understanding context, handling edge cases, and adapting to changing circumstances. One form of joint AI and human decision-making that has been actively studied is when AI provides recommendations to human decision-makers. Grimon and Mills (2025) show that Child Protective Services work-

[29]The strong impact of these complementary forces (digitization and AI) is demonstrated by Raymond (2024) in the real estate market.

ers are able to reduce child hospitalizations when provided with an algorithmic risk score. However, unexpectedly, the gains did not seem to come from following the algorithmic predictions but rather from reallocating their attention to different features of the allegations. Raymond (2024) finds that the availability of algorithmic tools in the housing market generates responses from non-adopter investors as well, shifting their focus to market sectors where human investors have comparative advantages over algorithms. Cho (2023) uses a quasi-experimental sports setting to examine how baseball umpires are affected when they work with AI assistance, focusing on their skill development. Kreitmeir and Raschky (2024) use a ban on ChatGPT in Italy to show that high-productivity programmers have worse output in the presence of AI assistance.[30]

Making decisions in environments with algorithmic recommendations elevates the need for humans to exercise proper discretion, which can lower the effectiveness of these recommendations. The findings from Hoffman, Kahn, and D. Li (2018) highlight challenges in exercising discretion in hiring, as managers are observed overruling recommendations due to personal biases rather than private information motives. Relatedly, Kawaguchi (2020) reports field experimental evidence from a retail business, showing that workers' adherence to algorithmic recommendations varies systematically with contextual factors, including sales volatility and worker experience. Exploring how humans adopt algorithmic recommendations, Agarwal, Moehring, Rajpurkar, and Salz (2023) find that providing radiologists with access to AI predictions does not, on average, result in improved performance. Another noteworthy finding from their study is that radiologists take significantly more time to reach a decision when AI information is provided.

Because giving AI final decision rights is potentially problematic, why would firms not simply provide workers with AI recommendations instead? The results above highlight two potential reasons why. First, humans may struggle to exercise appropriate discretion, sometimes deferring to AI when they should rely on their own judgment and sometimes overriding it when it is superior. For example, AI guidance can be systematically ignored due to overconfidence (Caplin, Deming, S. Li, Martin, Marx, Weidmann, and Ye 2025), and decision makers may apply an overly constant weight to algorithmic recommendations even when the informativeness of their private information varies substantially (Balakrishnan, Ferreira, and Tong 2026), leading to predictable over- and under-reliance. Second, in settings where a timely final decision is required, the very process of interpreting and integrating recommendations can create operational delay.

[30]One potential advantage of our setting for studying human-AI interaction is that, like the baseball setting, human line judges in tennis do not have any clear source of private information. We thank Ashesh Rambachan for raising this point.

5.2 Behavioral Economics

With AI oversight, discretion, a prominent force in algorithmic recommendation systems, no longer plays a major role. However, AI oversight potentially introduces a different set of behavioral forces, such as shame, pride, embarrassment, stress, and concerns about what being corrected signals about competence. Because of this connection, we contribute to the literature on social image and peer effects (Bursztyn and Jensen 2017). In much of this literature, people change behavior because their actions are observable to a relevant audience. In our setting, the salient observable object is not just the original action, but the public correction of that action: Hawk-Eye can reveal that the umpire crew made a mistake and can do so in front of players, spectators, and broadcast viewers. Our results provide suggestive evidence that the cost of being corrected in public can overpower the potential incentive to free ride on technology.[31]

Related field evidence underscores the importance of observability in human-AI collaboration. In a recommendation-based setting, Almog (2025) finds that making workers' reliance on AI visible to an evaluator reduces adoption and worsens performance, consistent with concerns that high reliance on AI is interpreted as a negative signal of competence. Our setting complements this evidence by studying an audit-and-override system rather than a recommendation system. In both cases, behavior changes not only because AI changes the information available to the decision-maker, but also because the interaction with AI changes what others can infer about the decision-maker's ability, effort, or judgment.

A revealing feature of our empirical context is that, with the introduction of AI oversight, one type of mistake became more controversial and led to a larger backlash, including complaints from players and fans. This introduced a change in the costs of different error types, which led to a decrease in the number of controversial mistakes at the expense of an increase in the number of less embarrassing ones. It is worth noting that Hawk-Eye decisions were available for television broadcasts both before and after the introduction of AI review, so this aspect of embarrassment does not change over the period we study. Relatedly, Margolin, Reimer, and Schaupp (2024) study the introduction of the Video Assistant Referee (VAR) in German professional soccer, a system of real-time post-decision feedback that allows initially wrong calls to be reversed. They show that immediate review changes referees' effort and decision-making, but the effects are heterogeneous: experienced referees improve final decisions without harming initial decisions, whereas inexperienced referees make worse initial decisions and do not improve final decisions. Their setting differs from ours, however, because review is conducted by a human video official, supported by computer-aided analysis, rather than by a high-accuracy AI system that audits and overrules decisions.

[31] The structural estimation in Appendix D also adds to existing methods for quantifying the welfare effects of shame and pride (e.g., Butera, Metcalfe, Morrison, and Taubinsky 2022).

6 Conclusion

This paper studies how workers respond when an algorithm can publicly overturn their decisions. Using the introduction of Hawk-Eye review for close line calls in professional tennis, we document that AI oversight changes not only average accuracy, but also the composition of mistakes. On average, oversight improves performance on close calls: for bounces within 100 mm of the line, the incorrect-call rate falls by about 1.1 percentage points (roughly 8% relative to baseline). Yet the response is not uniform across states. For the closest decisions, umpires shift toward calling balls in. This shift reduces Type II errors (calling out when in) but increases Type I errors (calling in when out), generating a clear backfire region: for balls just outside the line, mistake rates rise sharply even as they fall for balls just inside.

The main goal of this paper is to leverage a unique empirical setting to study what the adoption of AI oversight in tennis can reveal about human behavior. We do not aim to suggest an optimal practice for this application, as it is obvious that, in terms of accuracy alone, line umpires in tennis are strictly dominated by Hawk-Eye. Before the introduction of Hawk-Eye review, points were incorrectly awarded 13.9% of the time for close calls. After the introduction of Hawk-Eye review, points were incorrectly awarded 9% of the time. A system where Hawk-Eye was the decision-maker would eliminate such mistakes almost entirely. However, the ATP did not adopt Electronic Line Calling Live across the Tour until 2025. This provides an important lesson about labor displacement: if it took 20 years and a pandemic for the ATP Tour to replace line umpires with a superior (and almost perfect) technology, it is hard to imagine AI replacing doctors or judges anytime soon. Collaboration between AI and humans will be a topic of first-order relevance for years to come, and we believe this paper provides valuable information to guide the design of better AI interventions.

For organizations deploying AI oversight (audit-and-override systems), three implications follow. First, accuracy metrics that collapse across error types can be misleading. Oversight may reduce overall error while shifting errors toward states where the organization or customers bear higher costs. Effective governance therefore requires monitoring error *composition* and cost-weighted performance, not only average accuracy. Second, the design of the override protocol is central. When correction is more visible, more disruptive, or more contentious for one error type, workers will rationally adjust thresholds toward the "safer" mistake. Designing symmetric remedies and reducing discretionary, high-friction corrections can mitigate undesirable substitution. Third, oversight is most likely to distort behavior when tasks operate near human perceptual limits or when further accuracy improvements are cognitively expensive. In such settings, complementary changes (training, workflow redesign, better escalation rules, or selectively automating edge cases) may dominate simply adding review.

More broadly, the tennis setting illustrates a practical reality: even near-perfect technologies may coexist with human decision makers for long periods, making the design of hybrid oversight regimes a first-order managerial problem. Future experimental work with greater control over AI and human costs can build on these findings by varying key implementation features, such as observability of overrides, symmetry of remedies, and whether overrides affect evaluations and careers, to map when AI oversight shifts the frontier inward versus when it primarily reshapes behavior through incentives and social costs.

References

Abramitzky, Ran, Liran Einav, Shimon Kolkowitz, and Roy Mill (2012). "On the Optimality of Line Call Challenges in Professional Tennis". *International Economic Review* 53.3, pp. 939–964.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz (2023). "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology". *Working Paper 31422, National Bureau of Economic Research*.

Almog, David (2025). "Barriers to AI Adoption: Image Concerns at Work". *arXiv: 2511.18582*.

Almog, David, Lucas Lippman, and Daniel Martin (2026). "When an AI Judges Your Work: The Hidden Costs of Algorithmic Assessment". *arXiv: 2603.02076*.

Athey, Susan, Kevin Bryan, and Joshua Gans (2020). "The Allocation of Decision Authority to Human and Artificial Intelligence". *AEA Papers and Proceedings* 110, pp. 80–84.

Avery, Mallory, Andreas Leibbrandt, and Joseph Vecci (2023). "Does Artificial Intelligence Help or Hurt Gender Diversity? Evidence from Two Field Experiments on Recruitment in Tech". *Working paper*.

Balakrishnan, Maya, Kris Johnson Ferreira, and Jordan Tong (2026). "Human-Algorithm Collaboration with Private Information: Naïve Advice-Weighting Behavior and Mitigation". *Management Science* 72.1.

Bhattacharya, Vivek and Greg Howard (2022). "Rational Inattention in the Infield". *American Economic Journal: Microeconomics* 14.4, pp. 348–93.

Brown, Zach Y and Jihye Jeon (2024). "Endogenous information and simplifying insurance choice". *Econometrica* 92.3, pp. 881–911.

Bursztyn, Leonardo and Robert Jensen (2017). "Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure". *Annual Review of Economics* 9.1, pp. 131–53.

Butera, Luigi, Robert D. Metcalfe, William Morrison, and Dmitry Taubinsky (2022). "Measuring the Welfare Effects of Shame and Pride". *American Economic Review* 112.1, pp. 122–168.

Camerer, Colin F. (2019). "Artificial Intelligence and Behavioral Economics". In: *The Economics of Artificial Intelligence: An Agenda*. Ed. by Ajay Agrawal, Joshua Gans, and Avi Goldfarb. Chicago: University of Chicago Press, pp. 587–608.

Caplin, Andrew and Mark Dean (2013). *Behavioral Implications of Rational Inattention with Shannon Entropy*. Tech. rep. National Bureau of Economic Research.

Caplin, Andrew, David Deming, Shangwen Li, Daniel Martin, Philip Marx, Ben Weidmann, and Kadachi Ye (2025). "The ABC's of Who Benefits from Working with AI: Ability, Beliefs, and Calibration". *Management Science*.

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Ludwig Jens, and Sendhil Mullainathan (2016). "Productivity and Selection of Human Capital with Machine Learning". *American Economic Review: Papers & Proceedings* 106.5, pp. 124–127.

Chan, David C., Matthew Gentzkow, and Chuan Yu (2022). "Selection with Variation in Diagnostic Skill: Evidence from Radiologists". *The Quarterly Journal of Economics* 137.2, pp. 729–783.

Cho, Sungwoo (2023). "The Effect of Robot Assistance on Skills". *Working paper*.

Ely, Jeffrey, Romain Gauriot, and Lionel Page (2017). "Do Agents Maximise? Risk Taking on First and Second Serves in Tennis". *Journal of Economic Psychology* 53, pp. 135–142.

Gauriot, Romain and Lionel Page (2019). "Does Success Breed Success? a Quasi-Experiment on Strategic Momentum in Dynamic Contests". *The Economic Journal* 129.624, pp. 3107–3136.

Gauriot, Romain, Lionel Page, and John Wooders (2023). "Expertise, Gender, and Equilibrium Play". *Quantitative Economics* 14.3, pp. 981–1020.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár (2024). "Contamination Bias in Linear Regressions". *American Economic Review* 114.12, pp. 4015–51.

Grimon, Marie-Pascale and Christopher Mills (2025). "Better Together?: A Field Experiment on Human-Algorithm Interaction in Child Protection". *Working paper*.

Hoffman, Mitchell, Lisa Kahn, and Danielle Li (2018). "Discretion in Hiring". *The Quarterly Journal of Economics* 133.2, pp. 765–800.

Huelden, Tobias, Vitalijs Jascisens, Lars Roemheld, and Tobias Werner (2024). "Human-Machine Interactions in Pricing: Evidence from Two Large-Scale Field Experiments". *Working paper*.

Ide, Enrique and Eduard Talamas (2025). "Artificial Intelligence in the Knowledge Economy". *Journal of Political Economy* 133.12, pp. 3713–4112.

Kawaguchi, Kohei (2020). "When Will Workers Follow an Algorithm? A Field Experiment with a Retail Business". *Management Science* 67.3, pp. 1670–1695.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2018). "Human Decisions and Machine Predictions". *The Quarterly Journal of Economics* 133.1, pp. 237–293.

Kovach, Matthew, Daniel Martin, and Gerelt Tserenjigmid (2026). "Learning from an Unknown DGP: Experimental Evidence on Belief Updating with AI Recommendations". *Working paper*.

Kreitmeir, David and Paul A Raschky (2024). "The Heterogeneous Productivity Effects of Generative AI". *arXiv preprint arXiv:2403.01964*.

List, John A. (2020). "Non est Disputandum de Generalizability? A Glimpse into The External Validity Trial". *NBER Working Paper No. 27535*.

Margolin, Maximilian, Marko Reimer, and Daniel Schaupp (2024). "The Effects of Real-Time Feedback on Effort and Performance: Evidence from a Natural Quasi-Experiment". *Management Science*.

Matějka, Filip and Alisdair McKay (2015). "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model". *American Economic Review* 105.1, pp. 272–298.

Raghu, Maithra, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan (2019). "The Algorithmic Automation Problem: Prediction, Triage, and Human Effort". *Working paper*.

Rajpurkar, Pranav, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Robyn L. Ball, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng (2017). "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning". *Working paper*.

Rambachan, Ashesh (2024). "Identifying Prediction Mistakes in Observational Data". *The Quarterly Journal of Economics*.

Raymond, Lindsey (2024). "The Market Effects of Algorithms". *Working paper*.

Sims, Christopher A (2003). "Implications of Rational Inattention". *Journal of Monetary Economics* 50.3, pp. 665–690.

Topol, Eric (2019). "High-performance medicine: the convergence of human and artificial intelligence". *Nature Medicine* 25, pp. 44–56.

Whitney, David, Nicole Wurnitsch, Byron Hontiveros, and Elizabeth Louie (2008). "Perceptual Mislocalization of Bouncing Balls by Professional Tennis Referees". *Current Biology* 18, R947–9.

A Data Management

A.1 Exclusion of Clay Tournaments

Clay surfaces present some particular difficulties for the Hawk-Eye system, so Hawk-Eye review has not been adopted in clay tournaments. One issue is that the surface is continuously changing during a match, so the accuracy of Hawk-Eye decreases and requires constant re-calibration to operate at its best.[32] Also, Hawk-Eye review on clay courts might introduce controversies where the ball's mark and Hawk-Eye might disagree. As noted in The Guardian, Lars Graf, one of the ATP's most experienced officials, explained the decision not to use Hawk-Eye on clay: "We decided not to use Hawk-Eye on clay because it might not agree with the mark the umpire is pointing at."[33]

Because Hawk-Eye has not been adopted in clay tournaments, they may appear as an ideal control group for comparing umpire reactions in tournaments with Hawk-Eye review versus those on clay. However, there are several factors that led us to exclude clay tournaments from our analysis. For one, we are unable to determine from our data set which calls were corrected after examining the clay marks. In addition, before the introduction of Hawk-Eye, players were allowed to request the chair umpire to review the ball bounce marks on the surface, and changing the call based on that is permitted, so a form of review system was available.

A.2 Merging Algorithm

Our merging algorithm was designed to identify the highest number of the 144 challenges (derived from the 43 matches with video replays) within the Hawk-Eye Base data set, while keeping the number of false positives (challenges matched into the wrong point) as low as possible. We used standard variables, such as set, game, score, distance difference, player hitting the ball, and whether it was a tie-break, in order to minimize overfitting.

Our final merging algorithm comprises eight iterations, where we start using stricter criteria and gradually relax them in subsequent iterations. Some iterations, like 5 and 6, focus on very specific situations of the match (tie-breaks), and while they do not seem useful in the testing sample, they might prove useful in a bigger sample.

The algorithm merged 143 out of the 144 challenges, achieving a merge rate of 99.3%. Out of these, 136 were merged to the correct point, as validated through video auditing replays, resulting in an accuracy rate of 95.1%. It would be hard to improve this efficacy as

[32] https://www.perfect-tennis.com/why-there-is-no-hawk-eye-on-clay/

[33] https://www.theguardian.com/lifeandstyle/2009/jun/27/tennis-hawk-eye

Iteration	Correct Merge	False Positives	Efficacy	Challenges Left
1	98	2	98%	44
2	25	2	93%	17
3	5	0	100%	12
4	1	0	100%	11
5	0	0	–	11
6	0	0	–	11
7	5	0	100%	6
8	2	3	40%	1

Table A.1: Performance for each iteration of the merging algorithm. Out of the 144 challenges audited, only one remained unmerged.

there are many challenges with missing variables like set and game. We will now elaborate on the specific criteria used in each of the eight iterations, followed by Table A.1, which summarizes the number of challenges successfully merged (and false positives) achieved in each iteration.

Merging criteria for each iteration:

Iteration 1. Same set, game, score, and player (with distance difference <35 mm).

Iteration 2. Same set, game, and player (with distance difference <15 mm).

Iteration 3. Same set, score, and player (with distance difference <15 mm).

Iteration 4. Same game, score, and player (with distance difference <10 mm).

Iteration 5. For tie-breaks: Same game and score (with distance difference <35 mm).

Iteration 6. For tie-breaks: Same game. If multiple, pick closest in distance (with distance difference <15 mm).

Iteration 7. Same set, game and player. If multiple, pick closest point in score and then in distance (with distance difference <35 mm).

Iteration 8. Same set and player. If multiple, pick closest point in distance (with distance difference <35 mm).

A.3 Identification of Incorrect Calls

The Hawk-Eye Base data set contains the precise location of every ball bounce and identifies the winner of each point. However, it lacks direct information on the umpires' calls. Therefore, we are tasked with identifying the original umpire calls, which we do through the application of four criteria. The first two determine points where the umpire incorrectly called a ball in, and the final two identify the other type of incorrect call (umpire calling a ball out incorrectly). In the period involving challenges, we first restore the umpire's original calls following a successful challenge. Then, we assess whether any of the four criteria are satisfied. We are confident that these four criteria comprehensively identify potential mistakes in our data set.

Criterion 1. A player's stroke is out and wins the point.

We selected strokes where the first player to hit a ball out in a given point also won the point. We identified 55 points that satisfy this criterion, and 54 of them were confirmed by the video replays to be mistakes, resulting in a 98% accuracy rate.

Criterion 2. The point continued after an out.

To complement Criterion 1, we must also identify points where a player hit a ball out, and despite the error, the point continued, but the same player ultimately lost the point. To achieve this, we examine instances where, after a player hits the ball out, there are at least three more strokes registered. Out of the 25 points satisfying this criterion, 24 were umpire mistakes, yielding an accuracy rate of 96%.

We also tested a more relaxed criterion that identifies points recording at least two more strokes after an out. However, this criterion proved to be too lenient, encompassing multiple cases where the point was correctly called, and players continued hitting the ball afterward. When the criterion is relaxed to require at least two more strokes, the accuracy rate drops to 52%. This adjustment results in the identification of one new mistake at the expense of 24 incorrectly identified instances. This is why we adhere to the 3+ rule.

Criterion 3. All the strokes of a point are in, and the player hitting last loses.

We identify strokes where the last player to hit the ball in did not win the point. Out of the 25 points identified in the auditing matches, 23 were confirmed to be mistakes, resulting in a 92% accuracy rate.

Criterion 4. Observing a second serve after the first serve was in.

This criterion helps identify two types of instances where there is no winner in a point as a direct consequence of an umpire mistake. First, it recognizes those first serves that bounced in but were incorrectly ruled out, resulting in no winner for the point, which then moves to a second serve. Second, it detects points that lasted multiple strokes but had to be replayed

because the umpire stopped the point by incorrectly calling a ball out. The accuracy rate for this criterion is 83.7%, with 36 of the 43 selected points identified as umpire mistakes. Some inaccurately selected points by this criterion include lets (serves that hit the top of the net and the point is replayed) and other unusual events (e.g., a server touching the line while serving, and the serve not counting). The occurrence of these events, and our ability to identify them, should not change with the introduction of Hawk-Eye.

For this criterion, we limit the analysis to strokes within 40 mm of the line, as beyond this threshold the criterion becomes less accurate. For balls bouncing between 40-100 mm away from the line, this criterion has a 30% accuracy rate, as only seven of the 23 points audited were actual umpire mistakes.

B External Validity

To provide a more complete assessment of external validity, we build around the four transparency conditions given as an *onus probandi* in List (2020).

B.1 Selection

This condition concerns the representativeness of the studied group compared to the underlying population, in particular in "observables that might impact preferences, beliefs, or individual constraints." Professional tennis umpires, especially at the top tennis tournaments we study, have long experience and are highly skilled, so our study is more likely to generalize to other specialist populations, like expert technicians. However, these tennis umpires are not required to undertake years of formal education to be qualified for their roles, unlike doctors and judges, so to the extent that this matters for AI oversight, we might be understating the effect for such populations, which could have stronger concerns about AI oversight.

B.2 Attrition

This condition asks whether there are motivational or incentive differences between subject and target groups to maintain compliance. As noted in the introduction, one of the leading tennis officials at the time of Hawk-Eye's introduction confirmed that many important factors remained consistent over the period we study: there were no substantial changes to the training and performance assessments of umpires, the pool of umpires, or the positioning and instructions given to line judges. While this makes it easier to isolate the impact of AI oversight, it could impact the external validity of our results. Whether our results are seen in other settings could depend on whether AI oversight (or any institutional feature that accompanies the introduction of AI oversight) produces attrition in that setting. Such differential attrition could introduce confounds that attenuate the AI oversight effect. For example, those who have the highest cost of being overruled by AI might be the ones that are most likely to attrit, leading to less impact on those that remain.

B.3 Naturalness

This condition addresses the artificiality of a setting, such as in the choice task. We view the task in our study as having parallels to doctors making diagnosis decisions based on images that are hard to read. However, because it is more of a perceptual task than about general intelligence, this task might not capture ego-relevant concerns, perhaps understating

the effect in a general population. A potentially interesting follow-up study is to investigate how humans respond to AI oversight in ego-relevant tasks.

Another aspect of our setting is that sports officiating is a highly visible activity. There is a live audience of spectators, along with a television-viewing audience. Thus, our results are more likely to extend to contexts where the outcome and overruling mechanism are observed by an audience that the decision-maker would care about. For example, the decision-maker's audience could be their customers, co-workers, or the general public. We view this aspect of our setting as having parallels to many professional decisions that occur under public scrutiny – for example, doctors make diagnostic decisions in front of patients and colleagues, judges determine bail outcomes in court with prosecutors, clerks, and officers present, and HR decisions are often made in team settings. However, having a larger audience might induce additional human responses that do not apply to more private or less intimate settings.

A final factor is that AI is nearly perfect at the task we study. As a result, our results are more likely to extend to settings where AI is highly accurate *when used for corrections*. For example, while fully autonomous self-driving cars have not yet taken over the streets, many vehicles already include lane-keeping assistance systems that intervene when a driver drifts. These systems, though imperfect, operate with a high degree of confidence in the moments they are engaged, similar to other decision-support technologies that are used in professional settings. It remains to be seen if our results apply to settings where the AI is less accurate *when used for corrections*. Being overruled by a less precise technology may influence the behavioral forces we study in different ways, as humans can shift blame or defer shame. A potentially interesting follow-up question is whether human reactions to AI oversight differ by the ability of the AI to overrule human decisions correctly.

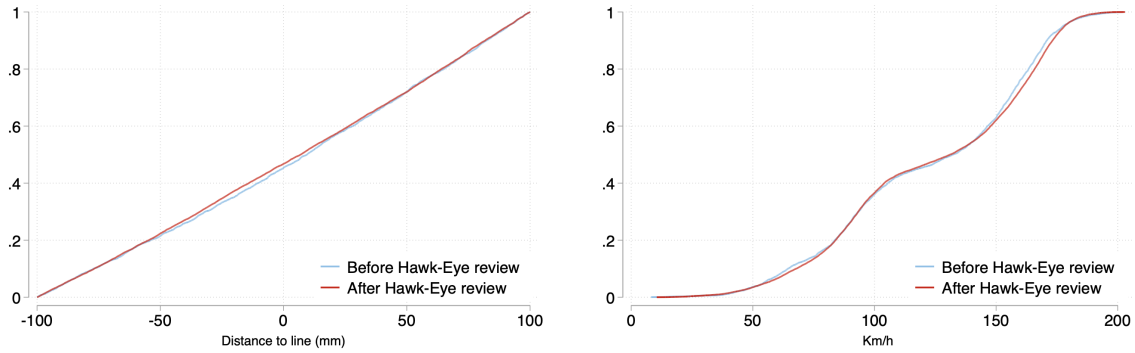
B.4 Scaling

This condition concerns whether the effects persist if the intervention is scaled. We have already addressed issues related to horizontal scaling (moving to other populations), but here we address issues of vertical scaling (scaled up to a larger population). For cost reasons, only the top tournaments implemented AI oversight, but all umpires at this level faced this technology, so all of the umpires' professional peers faced the same AI oversight. Therefore, our study sheds light on this level of scale. One concern could be that the widespread introduction of AI oversight might dampen any psychological costs of AI oversight. However, we see that the response to AI oversight gets larger with time, so the exposure and visibility that comes with increases in scale might actually increase the effect.

C Additional Tables and Figures

Tournament	Start Month	Category	Court Type	Number of Matches			
				2005	2006	2007	2008
Marseille	Feb	International (250)	Hard				25
Rotterdam	Feb	International Gold (500)	Hard			15	26
Dubai	Mar	International Gold (500)	Hard				19
Las Vegas	Mar	International (250)	Hard				19
Indian Wells	Mar	Masters (1000)	Hard	17		27	
Miami	Mar	Masters (1000)	Hard	15	18	26	
Queens	Jun	International (250)	Grass	13	19	18	
UCLA	Jul	International (250)	Hard			23	
Indianapolis	Jul	International (250)	Hard		7	23	
Washington	Jul	International (250)	Hard			21	
Montreal	Aug	Masters (1000)	Hard	20		27	
Toronto	Aug	Masters (1000)	Hard		25		
Cincinnati	Aug	Masters (1000)	Hard	11	19	24	
New Haven	Aug	International (250)	Hard			18	
Beijing	Sep	International (250)	Hard			20	
Kremlin Cup	Oct	International (250)	Carpet06 Hard07		8	15	
Madrid	Oct	Masters (1000)	Hard		30	30	
Basel	Oct	International (250)	Hard			12	
Paris Masters	Oct	Masters (1000)	Carpet06 Hard07		31	33	
Shanghai	Nov	Masters (1000)	Carpet05 Hard06-07	14	15	15	

Table C.1: Information on the 35 tournaments in the consolidated data set. The numbers in each cell indicate the matches available in our data set. The color represents the group of tournaments, with blue indicating tournaments without Hawk-Eye review and red indicating tournaments with Hawk-Eye review.



(a) Distance to line CDF.

(b) Speed CDF.

Figure C.1

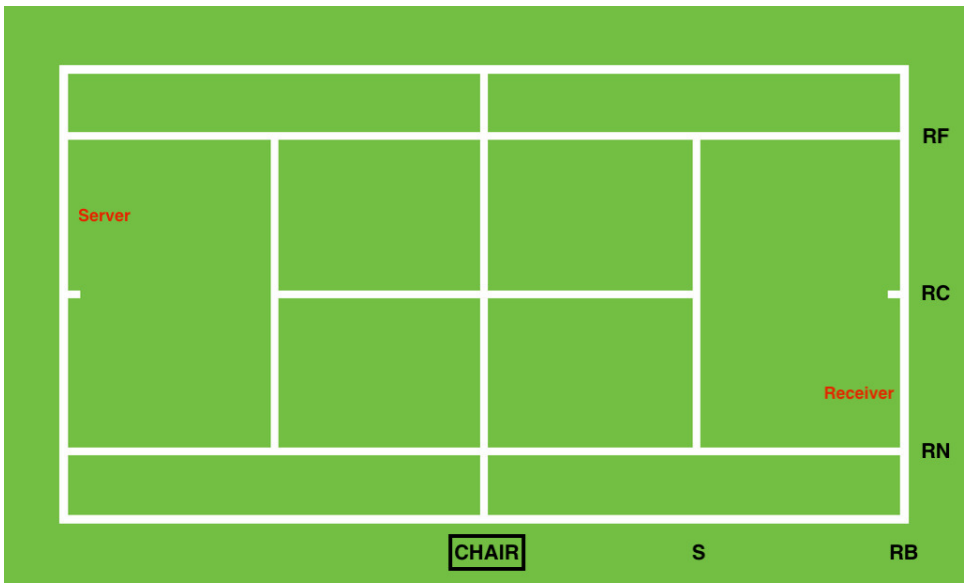


Figure C.2: The location of the umpiring crew: S = the serve-line umpire, RB = the right baseline umpire, RN = the right near long-line umpire, RC = the right center-line umpire, RF = the right far-side long-line umpire. The left side of the court has corresponding line umpires, except for the serve-line umpire, who moves to the left side when the right side player has the serve.

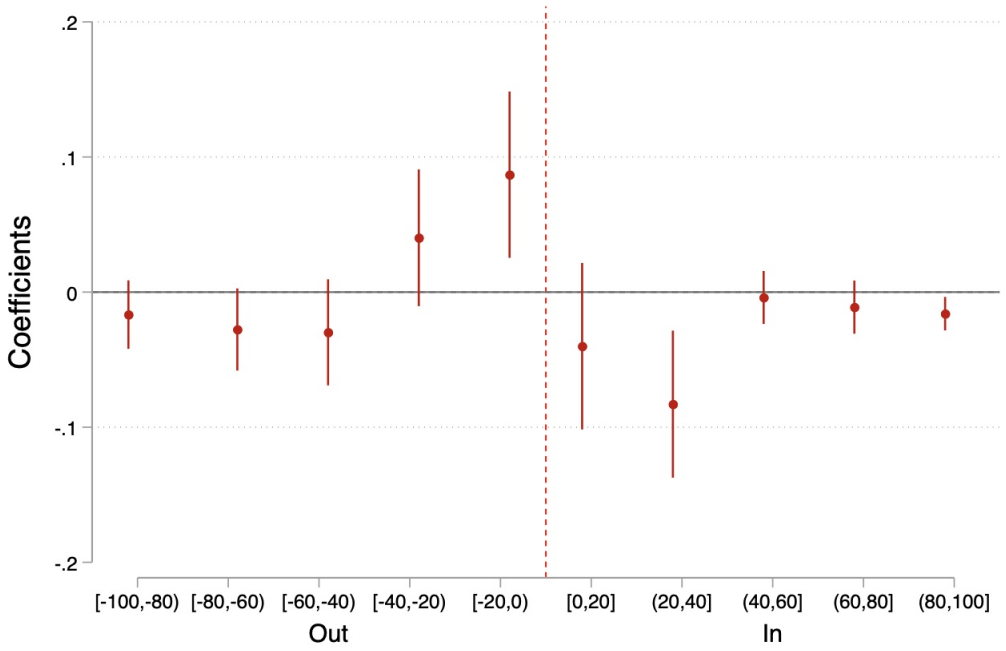


Figure C.3: An analog of Figure 2, using a one-treatment-at-a-time approach.

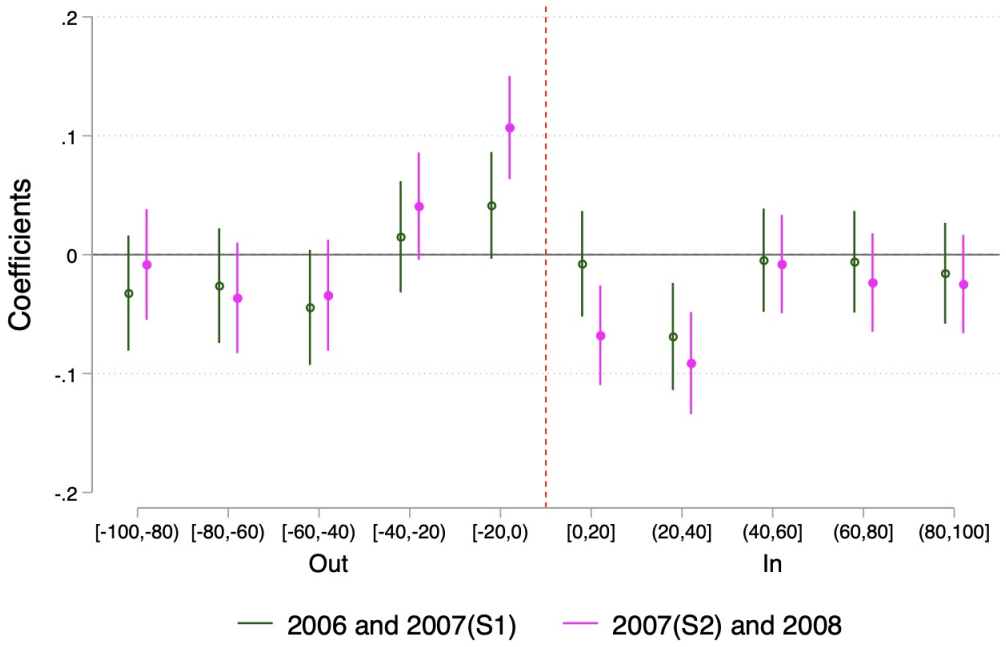


Figure C.4: Impact of AI oversight on the incorrect call rate by proximity to the line (by time sub-period). Matches are grouped into those with Hawk-Eye review in all of 2006 and the first semester of 2007 and those with Hawk-Eye review in the second semester of 2007 and all of 2008. Each dot represents the coefficient on the interaction between the distance bin and PostHK, the indicator variable that equals 1 if the Hawk-Eye review system is active.

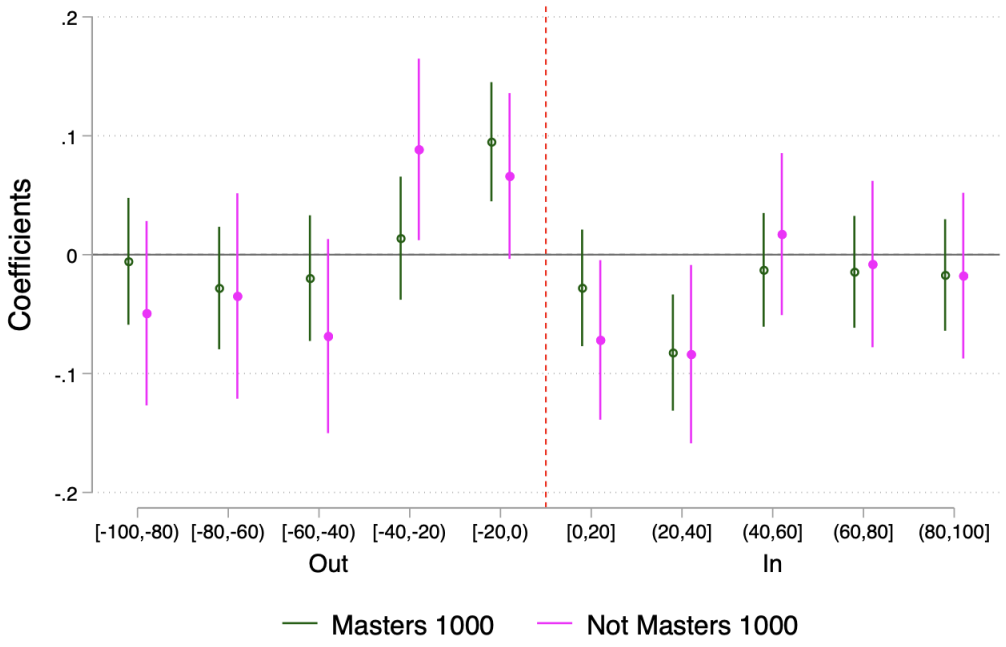


Figure C.5: Heterogeneous Impact of AI Oversight on Umpires' Performance Across 20 mm Distance Bins by Tournament Category. Matches are grouped into higher-ranked tournaments (Masters 1000) and lower-ranked tournaments (International 500 and 250). Each dot represents the coefficient on the interaction between the distance bin and PostHK, the indicator variable that equals 1 if the Hawk-Eye review system is active, analyzed separately by tournament category.

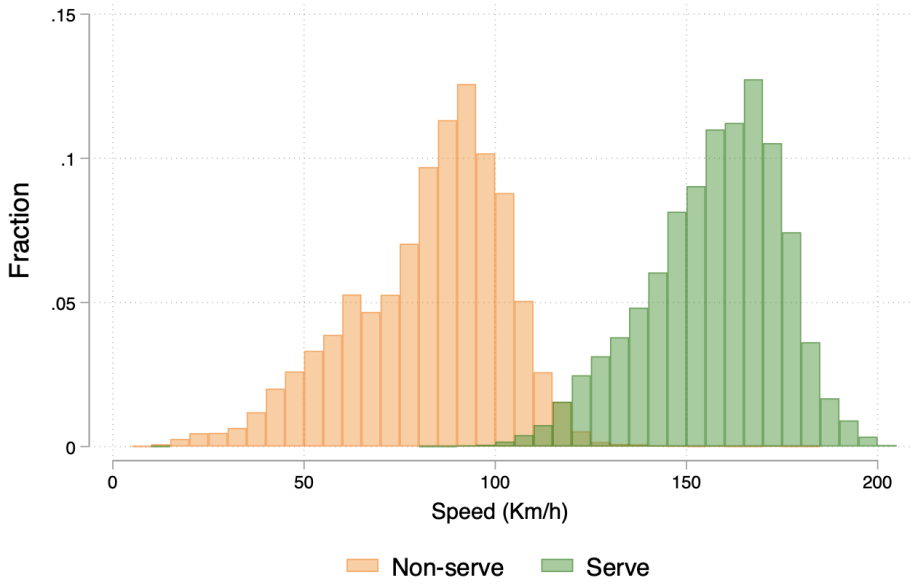
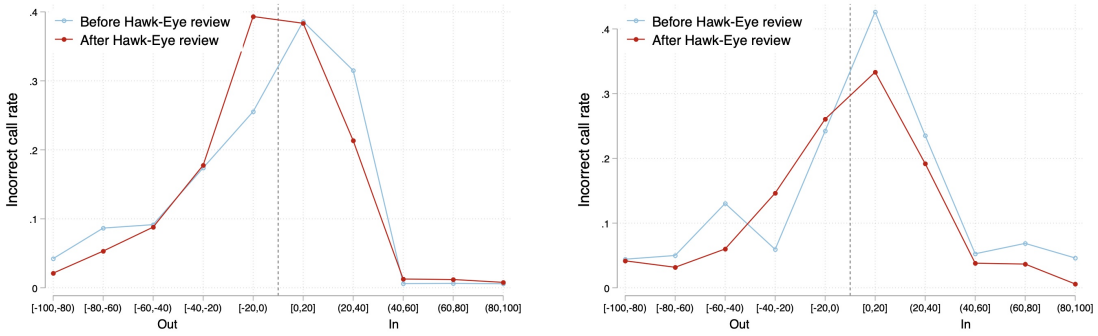


Figure C.6: Speed distribution for all the balls that bounced within 100 mm of the line, separated by non-serves and serves.



(a) Serves.

(b) Non-serves.

Figure C.7: Incorrect call rates by proximity to the line (separately for serves and non-serves). Each dot is the rate of incorrect calls for a bin of 20 mm, and dots to the left of the dashed line represent bins out of bounds, and those to the right of the dashed line represent bins in bounds.

D Model of Rationally-Inattentive Umpires

In this section, we introduce a model of attention-constrained umpires that shows asymmetric override costs can shift attention and thresholds.[34] Here, we apply the rational inattention theory introduced by Sims (2003) and characterized by Matějka and McKay (2015) and Caplin and Dean (2013), which we refer to as the "Shannon model." Bhattacharya and Howard (2022) use this model to study attention constraints in baseball, while Brown and Jeon (2024) use it to study consumers choosing among complex health insurance plans. We structurally estimate the costs of being overruled by AI using this model and then use these estimates to generate counterfactual predictions.

D.1 Model Setup

In our model, the set of states is $\omega \in \{\omega_I, \omega_O\}$ for the ball being in (ω_I) or out (ω_O), and the set of actions is $a \in \{a_I, a_O\}$ for calling the ball in (a_I) or out (a_O). The probability an incorrect call is challenged is η_I when the ball is in, and η_O when the ball is out, and both are equal to zero in tournaments without AI review. When the umpire's call is *not challenged*, we assume that they receive a normalized utility of 0 when correct and -1 when incorrect. When the umpire's call is *challenged*, we also assume that they receive a normalized utility of 0 when correct (when their call is upheld). But when incorrect (their call is overturned), we assume that they receive c_I when the ball is in and c_O when it is out. Thus, before taking into account attentional costs, gross expected utility U is

$$U(a,\omega) = \begin{array}{c} \\ a_I \\ a_O \end{array} \begin{array}{c} \begin{array}{cc} \omega_I & \omega_O \end{array} \\ \begin{pmatrix} 0 & -1+\eta_O\,(1+c_O) \\ -1+\eta_I\,(1+c_I) & 0 \end{pmatrix} \end{array}$$

We interpret c_I and c_O as the umpire's net payoff from an overturned mistaken call, measured relative to the normalized payoff of 0 from being correct and -1 from being incorrect. For shorthand, we refer to the utility loss embedded in these parameters as the *AI oversight penalty*. These parameters include three parts: the utility loss from having made a mistake (which we assume to be -1), the utility boost from having the correct call be implemented (due to altruism, relief, and so on), and the utility loss from being overruled by AI (due to shame, embarrassment, subsequent arguments, and so on).

We expect that c_I and c_O are less than or equal to 0 because, for the parameters to be positive, the umpire would need to prefer having an incorrect call overturned to being correct

[34] While it could be interesting to consider the dynamic effects of being overruled by AI, we start by considering a static model.

in the first place. When c_I and c_O are equal to zero (when there is no AI oversight penalty), then the umpire receives a utility of 0 after their call is overturned. When c_I and c_O are equal to -1, then the umpire receives a utility of -1 after their call is overturned. This is as if any utility gain from having the correct call implemented is perfectly offset by utility loss from having the AI overrule them. When c_I and c_O are less than -1, the utility loss from being overruled by AI dominates any utility gain from having the correct call implemented.

D.2 Model of Attention

As is standard in models of attention, we assume that the umpire starts out with prior belief μ, where $\mu(\omega)$ is the probability of state $\omega \in \{\omega_I, \omega_O\}$. After receiving a noisy mental signal of whether the ball is in or out, the umpire forms posterior $\gamma \in \Gamma$, where $\gamma(\omega)$ is the probability of state $\omega \in \{\omega_I, \omega_O\}$. Given a posterior belief γ, the umpire decides what call to make (mixing between actions is allowed), and we assume the umpire maximizes expected utility when making a choice to take each action at that posterior.

Under the assumption that the umpire updates their prior belief using Bayes' rule, their attention can be represented by a Bayes-consistent information structure π, which stochastically generates posterior beliefs. Specifically, π is a function that maps a state ω from a set of states Ω into $\Delta(\Gamma)$, the set of probability distributions over Γ that have finite support, so that $\pi : \Omega \to \Delta(\Gamma)$. Let Π denote the set of all such functions, $\pi(\gamma)$ be the unconditional probability of posterior $\gamma \in \Gamma$, $\pi(\gamma|\omega)$ be the probability of posterior γ given state ω, and $\Gamma(\pi) \subset \Gamma$ denote the support of a given π. We limit the set of information structures to those in $\Pi(\mu) \subset \Pi$ that generate correct posteriors for a given prior belief μ, so that

$$\Pi(\mu) = \left\{ \pi \in \Pi | \forall \gamma \in \Gamma(\pi), \forall \omega \in \Omega, \gamma(\omega) = \frac{\mu(\omega)\pi(\gamma|\omega)}{\sum_{\omega' \in \Omega} \mu(\omega')\pi(\gamma|\omega')} \right\}.$$

Next, we assume information structures are chosen, which we interpret as the umpire choosing how much attention to pay and to what aspects of the game to pay attention. In the standard Shannon model, each information structure $\pi \in \Pi(\mu)$ has a cost that scales linearly with the mutual information between posteriors and the prior:

$$K(\mu, \pi) = \kappa \left(\left[\sum_{\gamma \in \Gamma(\pi)} \pi(\gamma) \sum_{\omega \in \Omega} \gamma(\omega) \ln \gamma(\omega) \right] - \sum_{\omega \in \Omega} \mu(\omega) \ln \mu(\omega) \right),$$

where $\kappa \in \mathbb{R}_{++}$ is the marginal cost of attention. We assume that attention costs are increasing in the difficulty of judging where a ball landed and that this perceptual difficulty is not directly affected by the introduction of Hawk-Eye review.

We extend the standard Shannon model by allowing the marginal cost of attention to differ across states:

$$K(\mu,\pi) = \kappa_I \left(\left[\sum_{\gamma \in \Gamma(\pi)} \pi(\gamma)\gamma(\omega_I) \ln \gamma(\omega_I) \right] - \mu(\omega_I) \ln \mu(\omega_I) \right) + \kappa_O \left(\left[\sum_{\gamma \in \Gamma(\pi)} \pi(\gamma)\gamma(\omega_O) \ln \gamma(\omega_O) \right] - \mu(\omega_O) \ln \mu(\omega_O) \right),$$

where $\kappa_I, \kappa_O \in \mathbb{R}_{++}$. The standard Shannon model is the special case in which $\kappa_I = \kappa_O$.

Whitney, Wurnitsch, Hontiveros, and Louie (2008) use data from Wimbledon 2007 to show that umpires call more tennis balls as being out when they are actually in than in when they are actually out. Their results are consistent with psychometric experiments showing that moving objects generate a perceptual bias and are perceived as shifted in the direction of their motion. Allowing attention costs to differ across states provides a parsimonious way to accommodate such asymmetric perceptual errors.

For the two-state, two-action problem, we focus on the full-support case in which the optimal information structure induces one posterior associated with each action. Let γ^{a_I} denote the posterior at which the umpire calls the ball in and γ^{a_O} denote the posterior at which the umpire calls the ball out.

With state-dependent attention costs, the analogue of the Invariant Likelihood Ratio condition of Caplin and Dean (2013) contains an additional term. This term is the difference in posterior-average marginal attention costs across the two action-contingent posteriors. In the two-state case, it is

$$(\kappa_I - \kappa_O)\left[\gamma^{a_I}(\omega_I) - \gamma^{a_O}(\omega_I)\right].$$

The generalized ILR conditions are therefore

$$\frac{\gamma^{a_I}(\omega_I)}{\gamma^{a_O}(\omega_I)} = \exp\left\{ \frac{U(a_I,\omega_I) - U(a_O,\omega_I) + (\kappa_I - \kappa_O)\left[\gamma^{a_I}(\omega_I) - \gamma^{a_O}(\omega_I)\right]}{\kappa_I} \right\},$$

and

$$\frac{\gamma^{a_O}(\omega_O)}{\gamma^{a_I}(\omega_O)} = \frac{1-\gamma^{a_O}(\omega_I)}{1-\gamma^{a_I}(\omega_I)} = \exp\left\{ \frac{U(a_O,\omega_O) - U(a_I,\omega_O) - (\kappa_I - \kappa_O)\left[\gamma^{a_I}(\omega_I) - \gamma^{a_O}(\omega_I)\right]}{\kappa_O} \right\}.$$

Thus the usual ILR equations are recovered when $\kappa_I = \kappa_O$, since the additional term then equals zero.

Under the normalization of utility made previously, the utility gain from making the correct call rather than the incorrect call in matches without AI review is one. In matches with AI review, the utility difference is instead

$$U(a_I,\omega_I) - U(a_O,\omega_I) = 1 - \eta_I(1+c_I),$$

and

$$U(a_O, \omega_O) - U(a_I, \omega_O) = 1 - \eta_O(1 + c_O),$$

where c_I and c_O are the AI oversight penalties and η_I, η_O capture the relevant exposure to AI review.

For matches without AI review, let γ^{a_I} and γ^{a_O} denote the posteriors associated with calling the ball in and out. Using the no-review utility normalization above, the generalized ILR conditions imply

$$\kappa_I \left(\ln \frac{\gamma^{a_I}(\omega_I)}{\gamma^{a_O}(\omega_I)} - [\gamma^{a_I}(\omega_I) - \gamma^{a_O}(\omega_I)] \right) + \kappa_O [\gamma^{a_I}(\omega_I) - \gamma^{a_O}(\omega_I)] = 1,$$

and

$$\kappa_I [\gamma^{a_I}(\omega_I) - \gamma^{a_O}(\omega_I)] + \kappa_O \left(\ln \frac{\gamma^{a_O}(\omega_O)}{\gamma^{a_I}(\omega_O)} - [\gamma^{a_I}(\omega_I) - \gamma^{a_O}(\omega_I)] \right) = 1.$$

Provided that

$$\ln \frac{\gamma^{a_I}(\omega_I)}{\gamma^{a_O}(\omega_I)} \ln \frac{\gamma^{a_O}(\omega_O)}{\gamma^{a_I}(\omega_O)} - [\gamma^{a_I}(\omega_I) - \gamma^{a_O}(\omega_I)] \left(\ln \frac{\gamma^{a_I}(\omega_I)}{\gamma^{a_O}(\omega_I)} + \ln \frac{\gamma^{a_O}(\omega_O)}{\gamma^{a_I}(\omega_O)} \right) \neq 0,$$

the solution in terms of posteriors is

$$\kappa_I = \frac{\ln \frac{\gamma^{a_O}(\omega_O)}{\gamma^{a_I}(\omega_O)} - 2 [\gamma^{a_I}(\omega_I) - \gamma^{a_O}(\omega_I)]}{\ln \frac{\gamma^{a_I}(\omega_I)}{\gamma^{a_O}(\omega_I)} \ln \frac{\gamma^{a_O}(\omega_O)}{\gamma^{a_I}(\omega_O)} - [\gamma^{a_I}(\omega_I) - \gamma^{a_O}(\omega_I)] \left(\ln \frac{\gamma^{a_I}(\omega_I)}{\gamma^{a_O}(\omega_I)} + \ln \frac{\gamma^{a_O}(\omega_O)}{\gamma^{a_I}(\omega_O)} \right)}$$

and

$$\kappa_O = \frac{\ln \frac{\gamma^{a_I}(\omega_I)}{\gamma^{a_O}(\omega_I)} - 2 [\gamma^{a_I}(\omega_I) - \gamma^{a_O}(\omega_I)]}{\ln \frac{\gamma^{a_I}(\omega_I)}{\gamma^{a_O}(\omega_I)} \ln \frac{\gamma^{a_O}(\omega_O)}{\gamma^{a_I}(\omega_O)} - [\gamma^{a_I}(\omega_I) - \gamma^{a_O}(\omega_I)] \left(\ln \frac{\gamma^{a_I}(\omega_I)}{\gamma^{a_O}(\omega_I)} + \ln \frac{\gamma^{a_O}(\omega_O)}{\gamma^{a_I}(\omega_O)} \right)}.$$

Thus, in the full-support, nondegenerate case, the no-review posteriors uniquely determine κ_I and κ_O when the implied values are positive.

Let γ^{a_I} and γ^{a_O} now be the corresponding posteriors for matches with AI review, and assume $\eta_I, \eta_O > 0$. Using the review utility specification above, the generalized ILR conditions imply

$$c_I = \frac{1 - \kappa_I (\ln \gamma^{a_I}(\omega_I) - \ln \gamma^{a_O}(\omega_I)) + (\kappa_I - \kappa_O) [\gamma^{a_I}(\omega_I) - \gamma^{a_O}(\omega_I)]}{\eta_I} - 1,$$

and

$$c_O = \frac{1 - \kappa_O (\ln \gamma^{a_O}(\omega_O) - \ln \gamma^{a_I}(\omega_O)) - (\kappa_I - \kappa_O) [\gamma^{a_I}(\omega_I) - \gamma^{a_O}(\omega_I)]}{\eta_O} - 1.$$

Thus, in the full-support, nondegenerate case, the review posteriors uniquely determine c_I and c_O when the implied values of κ_I and κ_O are positive.

D.3 Structural Estimates

To structurally estimate the AI oversight penalty, we undertake two steps. First, we estimate the marginal costs of attention when there is no AI review. When there is no AI review, these marginal costs are jointly determined by the optimal posteriors $\gamma^{a_I}(\omega_I)$ and $\gamma^{a_O}(\omega_O)$. Because there is a single posterior for each action in our model, the optimal posteriors are equal to the *revealed posterior*, which is $\bar{\gamma}^a(\omega)$, the probability of each state conditional on the action taken. Our non-parametric estimate of the *true* revealed posterior is the *observed* probability of each state conditional on the action taken.

As shown in Table D.1, our estimate of the revealed posterior for the ball being in when the umpire calls the ball in is 0.8921, and our estimate of the revealed posterior for the ball being out when the umpire calls the ball out is 0.8276, for all calls within 100 mm of the line before the introduction of AI oversight. Therefore,

$$\bar{\gamma}^{a_O}(\omega_I) = 1 - \bar{\gamma}^{a_O}(\omega_O) = 0.1724$$

and

$$\bar{\gamma}^{a_I}(\omega_O) = 1 - \bar{\gamma}^{a_I}(\omega_I) = 0.1079.$$

Using the generalized ILR conditions for matches without AI review gives

$$\kappa_I = 0.855 \qquad \text{and} \qquad \kappa_O = 0.292.$$

Thus, under the state-dependent attention-cost specification, the estimated marginal cost of attention is higher for the state in which the ball is in than for the state in which the ball is out, which is qualitatively consistent with findings from the psychometric literature.

Parameter	Recovery equation	Estimate
$\bar{\gamma}^{a_I}(\omega_I)$	N/A	0.892
$\bar{\gamma}^{a_O}(\omega_O)$	N/A	0.828
κ_I	Generalized ILR	0.855
		(0.083)
κ_O	Generalized ILR	0.292
		(0.048)

Bootstrap standard errors in parentheses.

Table D.1: Estimated revealed posteriors and state-dependent costs of attention before the introduction of AI oversight.

As a second step, we use these estimated costs of attention, the observed challenge rates, and the revealed posteriors from matches with AI review to estimate the AI oversight

penalties. In this part of the estimation, $\bar{\gamma}^{a_I}$ and $\bar{\gamma}^{a_O}$ refer to the revealed posteriors in matches with AI review.

As shown in Table D.2, the observed challenge rates are 0.449 when balls were in and 0.415 when balls were out. Given the values of κ_I and κ_O determined without AI review, the rationalizing AI oversight penalties are

$$c_I = -1.425 \qquad \text{and} \qquad c_O = -1.019.$$

Parameter	Recovery equation	Estimate
η_I	N/A	0.449
η_O	N/A	0.415
$\bar{\gamma}^{a_I}(\omega_I)$, AI review matches	N/A	0.884
$\bar{\gamma}^{a_O}(\omega_O)$, AI review matches	N/A	0.866
c_I	Generalized ILR	-1.425
		(0.151)
c_O	Generalized ILR	-1.019
		(0.077)

Bootstrap standard errors in parentheses.

Table D.2: Estimated challenge rates, revealed posteriors, and AI oversight penalties in the Hawk-Eye review period.

It is important to highlight that the AI penalty, in our case, can be seen as a lower bound on the costs of being overruled. Under the scenario in which umpires do not care at all about the final outcome being correctly implemented, the incremental utility loss from being overruled, relative to making an unchallenged incorrect call, is the negative of the utility change from -1 to the AI oversight penalty. For example, when $c_I = -1.425$, being overruled changes utility by -0.425 relative to an unchallenged incorrect call, corresponding to a positive incremental utility loss of 0.425. Otherwise, the cost of being overruled is larger. These results are consistent with anecdotal evidence regarding the increased controversy surrounding this type of error.

These estimates suggest that the introduction of AI oversight reduced the utility of both types of overturned errors relative to the no-review incorrect-call payoff of -1. When the ball is in and the umpire incorrectly calls the ball out, the estimated utility of the overturned call changes from -1 without Hawk-Eye review to -1.425 after its introduction. When the ball is out and the umpire incorrectly calls the ball in, the estimated utility changes from -1

without Hawk-Eye review to -1.019 after its introduction.[35] The larger estimated penalty is for overturned calls when the ball is in: umpires care approximately 42.5% more about these errors relative to the no-review incorrect-call payoff, compared with approximately 1.9% more for overturned calls when the ball is out. Also, under this attention-cost model, the cost of AI oversight dominates any utility gain from having the correct call implemented for both types of errors.

D.4 Counterfactual Predictions

The structural model can also be used to generate counterfactual predictions for oversight regimes that are not observed. These counterfactuals hold the distribution of close calls fixed, use the estimated attention costs from Table D.1, and solve the generalized ILR conditions after changing the review exposure parameters (η_I, η_O) or the oversight penalties (c_I, c_O). Table D.3 reports the predicted umpire behavior for balls within 100 mm of the line.

Counterfactual	Call in	Type I error	Type II error	Human error
Actual public challenge review	53.1%	13.2%	11.8%	12.5%
Neutral AI penalty, $c_I = c_O = -1$	50.0%	11.5%	16.2%	14.0%
Lower challenge exposure, $\eta_I = \eta_O = 0.25$	51.8%	12.5%	13.5%	13.1%
Automatic public review, $\eta_I = \eta_O = 1$	55.6%	14.6%	8.3%	11.3%

Table D.3: Model-implied counterfactual initial-call behavior for balls within 100 mm of the line. Type I error is the probability of calling a ball in conditional on the ball being out. Type II error is the probability of calling a ball out conditional on the ball being in. Human error is the unconditional initial-call error rate.

These counterfactuals generate predictions that the reduced-form estimates alone cannot answer. The first possibility we consider is what would happen if the challenge rate stayed the same, but the AI review was changed to somehow reduce the utility impact on umpires. For this, we look at a neutral AI penalty in which an overturned error has the same payoff as an unchallenged incorrect call. This would substantially reduce the tendency to call balls in, but it would also weaken the incentive to avoid Type II errors. Our model predicts that the call-in rate would fall from 53.1% to 50.0%, Type I errors would fall from 13.2% to 11.5%, and Type II errors would rise from 11.8% to 16.2%. Overall, the human-error rate is predicted to increase by 1.5 p.p.

The second possibility we consider is the impact of reducing the challenge rate, perhaps by reducing the number of challenges available to players. Reducing the challenge exposure to 25% in both states lowers the expected cost of being overturned, moving the predicted

[35] The bootstrap standard errors were obtained from drawing 1000 random samples with replacement, each containing the same number of observations as the original dataset.

call-in rate closer to 50% and increasing Type II errors relative to the actual public challenge regime. Asymmetric challenge probabilities would push thresholds in the direction of the error type that is less likely to be challenged. Overall, the model again predicts the human-error rate would increase, but to a lower extent than neutralizing the AI penalty.

Finally, we consider automatic review of human calls. Automatic public review of every human close call predicts an even higher call-in rate and a lower initial human-error rate before review. It is important to note that this differs from full automation, which removes the human call from the payoff problem entirely. In that case, the behavioral substitution disappears and final close-call error is governed by the accuracy of Hawk-Eye rather than by human attention or the costs of being overruled.