Human Expertise in the Age of AI: Playing Against AI-Assisted Opponents*

Ala Avoyan† Ivan Dedyukhin‡

October 31, 2026

Abstract

Most of what we know about how AI changes economic decisions concerns the person who uses it. However, in many interactions what also matters is what our beliefs are of our opponent using AI assistance. In this paper we ask how people play against a counterpart who has access to AI, and whether the answer depends on the expertize of the opponent. In a hybrid experiment, online participants from four age cohorts play the 11-20 money request game against lab participants who are novices or experts in strategic reasoning, with or without an AI assistant. We truthfully communicate opponent information to the participants. We find that opponent description has no effect on people requests on average, but people think considerably longer when the opponent is an expert with AI. We conclude that they treat AI as something that amplifies expertise rather than replaces it in this context. Baby Boomers and Millennials hold opposite views of AI effect, which is what hides the overall effect leading to the flat averages.

JEL Classification: C72, C91, D91, O33;

Keywords: Generative AI; Expertise; Beliefs about AI; Level-k reasoning; Generations; Lab-online experiment;

*We would like to thank Andrew Schotter for his comments and suggestions. IRB approval for the paper was granted by Indiana University's Institutional Review Board (# 29634). Our experiment was preregistered using the AEA RCT Registry, under ID 0017375.

†Indiana University, aavoyan@iu.edu

‡Indiana University, idedyukh@iu.edu

1 Introduction

When we think about how AI changes economic decisions, we usually think about the person who uses it. Whether a worker with an AI assistant produces more, whether a student with one learns less, whether an expert with one becomes dispensable. However, when we bid in an auction, negotiate a price, or compete for a job, the other party may be using an AI assistant, and we have to act without knowing what the assistant has done for them. What we act on, then, is not what AI does for a person but what we believe it does. Those beliefs are forming now, ahead of the evidence, and they will shape hiring, delegation, and competition whether or not they turn out to be right. In this paper we ask: how do people strategize when a counterpart who has access to AI? Do they treat an AI-assisted counterpart as more sophisticated than an unassisted one? Does the answer depend on who the counterpart is, so that AI is believed to turn a novice into an expert, or to make an expert more formidable still? And do people of different ages, who may hold different views of AI, answer these questions differently?

Two views of what AI does to a person using it are consistent with the evidence so far. On the one hand, AI substitutes for expertise: it supplies the answer, so a novice with an assistant is as capable as an expert without one. Evidence for this view comes from studies in which generative AI raised the productivity of the least experienced workers the most (Brynjolfsson et al., 2025b) and narrowed the gap between weaker and stronger performers (Noy and Zhang, 2023; Dell'Acqua et al., 2023). On the other hand, AI complements expertise: its value depends on the skill of the person using it, so it helps experts more than novices. Evidence for this view comes from studies in which AI advice helped high-performing entrepreneurs and hurt low performers (Otis et al., 2024) and in which the return to expertise was higher, not lower, when a task was done with AI (Hitzig et al., 2026). The truth is probably that both are right in different places, and the evidence will keep arriving task by task. What people believe does not wait for it. An employer who believes that AI substitutes for expertise discounts experience; a bidder who believes that it complements expertise worries more about an expert opponent, not less. Almost all of the existing evidence is about the person who has AI. We know little about what people believe about the person who has it, and ours is, to our knowledge, the first experiment in which the counterpart is a real person whose AI access and expertise are both varied, and the first to ask whether the response differs across age cohorts.

To answer these questions, we conduct a hybrid lab-online experiment built around the 11-20 money request game of Arad and Rubinstein (2012). Two players each request an integer between 11 and 20 points, receive what they request, and earn a bonus of 20 points for requesting exactly one point less than the other player. The game is a natural instrument for our question. Requesting 20 ignores the opponent, 19 best responds to an opponent who requests 20, 18 to one who requests 19, and so on, so a request can be taken as a statement about how many steps

of reasoning a player attributes to the other side.[1] There is no pure-strategy equilibrium and no answer to look up, which means there is nothing an assistant can supply beyond a belief about what the other player will do. The choice remains a judgment about the opponent even for a player who has read the original paper.

Our main subjects are 960 participants recruited on Prolific with quotas across four age brackets spanning Gen Z to the Baby Boomers. Each is truthfully told that she will play once against one lab participant of a stated type: an undergraduate or an economics PhD student with coursework in strategic interaction, with or without access to an AI tool on the decision screen. The four opponent types form a 2×2 design in which we vary the opponent's expertise and the opponent's AI access. The lab sessions are what makes the descriptions true: each online participant is paid against the decision of a real lab opponent of the stated type, and in the AI cells a chat box connected to ChatGPT.

This hybrid design is, as far as we know, new. Broad demographic pool of online subjects play against lab subjects whose tools the experimenter controls, so AI access can be described truthfully. Now we can examine whether AI makes opponents more sophisticated, which should show up as requests falling when the opponent has AI. If AI substitutes for expertise, requests against an AI-assisted novice should look like requests against an unassisted expert, and if it complements expertise, the two should differ. And responses may differ across age cohorts. Beyond requests, we record decision times, elicit confidence in the decision and the chance of winning the bonus following Enke and Graeber (2023).

Our first result is that none of the opponent descriptions change average requests across treatments. The substitution contrast is equally flat: requests against an AI-assisted novice are indistinguishable from requests against an unassisted expert. The invariance is not inattention.

What the descriptions move is effort. Median decision time is 45 percent higher against an expert with AI than against an unassisted expert, and the other three opponents are indistinguishable from one another. The effect is the same in every cohort: a joint test of treatment-by-generation interactions on log time does not reject ($p = 0.94$). The pattern is a complementarity. Learning that a novice has an assistant buys the novice no additional thought, and learning that the opponent is an expert buys her none either; only the conjunction does. In the one margin where our subjects visibly respond, they treat AI as something that amplifies expertise rather than as something that replaces it.

A lot of the story is hidden when generations are pooled. Looking by generation also reconciles

[1] This level-k reading is a simplification. A request of k pays k plus a bonus of 20 with the probability that the other player requests exactly $k + 1$, so against a belief that spreads mass over several requests the best response need not sit one step below the belief's mode. Requesting 20 is itself optimal whenever the opponent is thought to request 20 with probability below $1/20$ and 19 with probability below $1/10$, and so on down the range, and the game's unique mixed-strategy equilibrium puts positive weight on every request from 15 to 20. The level-k interpretation is nonetheless the one the original study was designed to elicit and the one we use to read the data. We characterise best responses to arbitrary beliefs in Section **??**.

our expertise null with a literature that would have predicted an effect. When student subjects are told that their opponent is more sophisticated, their responses reveal more steps of reasoning (see Agranov et al. (2012) and Alaoui and Penta (2016)). That is what Millennials do in this paper, and what Generation X does in attenuated form. The mean effect seems flat because Baby Boomers move the opposite way and Gen Z does not move at all. We are able to see these differences since we have a broad age sample from Prolific, and if both opponents were students we would likely only see one type of effect.

Let us go back to the response time results. Why would subjects think harder and make the same request? The model of endogenous depth of reasoning of Alaoui and Penta (2016) provides one possibility. A player takes one more step of reasoning only when her subjective probability that the opponent stops at that step exceeds the probability that she stops one step earlier by enough to cover the cost of the step. Information that makes an opponent's reasoning deeper but predictable shifts that probability mass down the ladder and moves the best response with it. Information that makes the opponent's reasoning harder to forecast flattens the mass, so that no step beyond the first pays at any cognitive cost. Hence, the optimal request stays at 19 or or goes to 20, which is the belief free response. Under the second kind of information a player can spend a long time trying to form a belief and end up where she started. We find supporting evidence in our data. Response time is highest where choices move the least. Let us be more precise, the participants who request 20 against an AI assisted expert deliberate longest of all, about 50 percent longer than the other participants facing that opponent. Furthermore, request of 19 becomes less common once a novice has an assistant as if the assistant makes the naive anchor less reliable. However, we cannot, with choices and times alone, separate this explanation from two others. Beliefs may not move at all, or they may move in offsetting directions across subjects, as the cohort results suggest. Distinguishing them requires eliciting beliefs directly.

2 Related Literature

We contribute to several strands of literature. The first measures what generative AI does to the person who uses it, and it keeps returning to the question our subjects must answer about someone else: does AI substitute for expertise or complement it? AI assistance helped the least experienced workers most in customer support, writing, and consulting (Brynjolfsson et al., 2025b; Noy and Zhang, 2023; Dell'Acqua et al., 2023), but it helped high performers more among entrepreneurs and in agentic coding (Otis et al., 2024; Hitzig et al., 2026); Autor and Thompson (2025) show that the same technology can do either, depending on which of an occupation's tasks it automates, and Autor et al. (2026) that junior lawyers gained more from an AI drafting assistant while using it but only senior lawyers kept an advantage without it. None of these studies asks what other people believe about the person who was assisted, although such beliefs have consequences: Brynjolfsson et al. (2025a) document employment declines

among early-career workers in AI-exposed occupations, which is what we would expect if employers act on the substitution view, whether or not it is correct. On what an assistant can offer in a problem with no correct answer, Ngangoue et al. (2026) characterize ChatGPT's advice on life's big decisions as naive advice in the sense of Schotter (2023).

The second literature changes the counterpart rather than the decision maker's tools. People play differently against computer programs, and how they play depends on what they are told about the program (March, 2021); they choose lower numbers in a beauty contest against an LLM than against a human (Barak and Costa-Gomes, 2025), and they trust and cooperate less with a partner known to have delegated the decision to ChatGPT (Dvorak et al., 2025). Shen et al. (2025) tell subjects in a trust game and a prisoner's dilemma whether an AI assistant is available to themselves, to their counterpart, or to neither. The paper finds that cooperation does not change on average, but unassisted subjects see assisted counterparts as more competitive. In these games a belief about how sophisticated the counterpart is mixes with a belief about what she wants. On the other hand, the 11-20 game has no social-preference component. The closest paper to ours is Di Cagno and Lin (2025). In a three player beauty contest played by students over several rounds, two members of each group receive a ChatGPT advisor built around level-k reasoning, and the third knows that she faces two advised opponents. We ask a similar question of a different setting: our assistant is a general-purpose model given only the rules of the game, the 11-20 game has no equilibrium for deeper reasoning to converge to, the advised opponent is a novice or an expert, and the subjects who face him come from four age cohorts and play once.

The third literature is level-k reasoning (Nagel, 1995; Stahl and Wilson, 1995; Camerer et al., 2004; Costa-Gomes and Crawford, 2006; Crawford et al., 2013), and in particular the finding that depth depends on what a subject believes about the opponent: subjects play deeper when told that the opponent is a graduate student rather than an undergraduate (Agranov et al., 2012), an expert chess player (Palacios-Huerta and Volij, 2009), or a student from a quantitative field (Alaoui and Penta, 2016). Alaoui and Penta (2016) model depth as a choice, in which a player takes one more step when its benefit, which depends on the stakes and on her beliefs about the opponent, exceeds its cognitive cost, and Alaoui et al. (2020) separate a subject's own reasoning from her beliefs about others. We add AI access to the sophistication label. The 11-20 game of Arad and Rubinstein (2012) is the natural instrument: its level-0 action is natural, each best response undercuts by exactly one point, 19 is a best response to a wide range of naive play, and the bonus does not come at the other player's expense. Later work finds that time pressure moves choices toward equilibrium (Lindner and Sutter, 2013), that noisy introspection fits the data better than pure types (Goeree et al., 2018), that risk aversion can account for high requests (Li and Rong, 2018), that stated level-0 beliefs differ from level-0 actions (Wang et al., 2026), and that financial professionals request less than students (Açıkgöz et al., 2024), which supports describing an opponent population as expert. Language models themselves play the

game differently from humans and are sensitive to how it is described (Gao et al., 2025; Kader and Lee, 2025).

3 Experimental Design

3.1 The Game

We use the 11-20 money request game of Arad and Rubinstein (2012). We describe it to participants as follows: You and another person will each independently request an amount of points. The amount must be an integer between 11 and 20 points (inclusive). Each of you *will receive the amount you request.* Additionally, a player will receive a *20-point reward* if they request *exactly one point less* than the other player. We refer to the 20-point reward as the bonus throughout the paper.

This game is particularly suitable for studying strategic reasoning for several reasons. First, the level-0 type specification is intuitively appealing: choosing 20 is a natural anchor for iterative reasoning. Second, best-responding is straightforward: 19 is the best response to 20, 18 is the best response to 19, and so on. Third, the level-1 strategy (choosing 19) is robust to a wide range of level-0 specifications, including uniform randomization. Fourth, unlike other games that trigger level-k reasoning, this game does not invoke social preferences: winning the bonus does not come at the expense of the other player. Finally, the game lacks both pure-strategy Nash equilibria and dominated strategies, making level-k reasoning a natural approach.

An additional advantage of the game is that knowledge does not yield an obvious "eureka" response. What should a player actually do? Suppose one sits down to play this game. They have read the original paper. They know the mixed strategy equilibrium. They have even seen the empirical distributions from prior experiments and even consulted an AI assistant. Still, when it comes time to submit a request, one faces a fundamental problem: none of this knowledge tells the person what *their opponent* will do and hence, what they should do.

The level-k framework treats the choice of 20 as purely naive, a "level-0" anchor from which sophisticated players iterate away. We argue in Section 4 that this characterization may be incomplete: against an opponent whose reasoning is hard to predict, the guaranteed payoff of 20 becomes attractive, so a high request can reflect diffuse beliefs rather than naive ones. In Section A.1, we characterize the beliefs under which choosing 20, and each lower request, is a best response.

3.2 Design

We implement a 2×2 design varying the opponent's AI access (none vs. ChatGPT) and the opponent's expertise (undergraduate vs. economics PhD student). Subjects play the basic version of the 11-20 money request game of Arad and Rubinstein (2012) against an opponent

described according to one of these four conditions. Our primary subjects are recruited from the Prolific platform. Their opponents are participants in a controlled laboratory experiment, whose AI access and expertise we vary directly. Table 1 summarizes the design. We refer to the four treatments as Novice, Novice+AI, Expert, and Expert+AI, where the label describes the opponent a subject faces.

Table 1. Treatment Summary

Treatment	Opponent	Opponent AI access	Prolific	Lab
Novice	Undergraduate	None	219	30
Novice+AI	Undergraduate	ChatGPT	219	29
Expert	Economics PhD student	None	214	13
Expert+AI	Economics PhD student	ChatGPT	219	11

Notes: Prolific columns report the analysis sample after the exclusions described in Section 3.4; 960 Prolific participants completed the study. Lab columns report all laboratory participants. Each laboratory session implemented a single treatment.

3.2.1 Prolific participants

Each Prolific participant is truthfully informed that they will play against one randomly selected opponent from a specific treatment group. Participants know their opponent's group identity but not their specific decision, and each participant plays once. The opponent is described as "an undergraduate student" in the Novice treatments and as "a PhD student in economics.[2]" in the Expert treatments. In the AI treatments, participants are additionally informed that their opponent "will have access to an AI tool on their decision screen."[3] Participants are also told that their opponent will make their decision in a laboratory session in the week of December 15 and that their payment will depend on both decisions.

For each treatment, we set recruitment quotas of 60 participants in each of four age brackets, chosen to approximate the four generational cohorts: Baby Boomers, Gen X, Millennials, and Gen Z.[4]

Before playing, participants answered five comprehension questions about the rules and the

[2]We added the following sentence: As part of their first-year training, Economics PhD students complete a required sequence of courses covering theories of strategic interactions.

[3]The Prolific participants were told that their match will have access to a ChatGPT-based AI tool on their decision screen. They were also shown a screenshot of lab participants decision screen which had a chatbox embedded in it.

[4]Generation was self-reported and also coded from age using the Pew Research Center birth-year definitions (Gen Z 1997-2012, Millennials 1981-1996, Gen X 1965-1980, Baby Boomers 1946-1964), evaluated at the time of the study: Gen Z 28 and younger, Millennials 29-44, Gen X 45-60, Baby Boomers 61 and older. Because Prolific reports age rather than birth year, each boundary is approximate by at most one year, and four participants aged 80 and 81 are grouped with the Baby Boomers. The two classifications agree for 94 percent of participants. We use the age-based coding throughout and report self-reported generation as a robustness check. The recruitment brackets placed 29-year-olds in the youngest group, so under the Pew definitions the analysis sample is not balanced across cohorts: it contains 172 Gen Z, 267 Millennial, 221 Gen X, and 211 Baby Boomer participants, and treatment-by-generation cells range from 39 to 75.

payoffs and could not proceed until each was answered correctly; we record the number of incorrect attempts. We also record the time spent on the request screen, which we use as our measure of deliberation. After submitting their request, participants answered two cognitive uncertainty questions on sliders from 0 to 100: how certain they were about their decision, and how certain they were that their request would result in receiving the bonus. The order of the two questions was randomized. Participants then completed three incentivized Raven matrix questions and a short questionnaire that asked whether they had used an AI tool during the task, how often they use AI tools at work or in their studies and in their personal lives, their generation, and, in the second recruitment batch only, their education. Participants had no AI assistance of their own. Only 8 of the 871 participants in the analysis sample reported using an AI tool during the task, and 7 percent ever left the browser tab during the study, for a median of 16 seconds; the average request of those who did (17.37) is indistinguishable from the average request of those who did not (17.24).

3.2.2 Lab participants

Laboratory sessions were conducted at the Interdisciplinary Experimental Laboratory (IELAB) at Indiana University. Lab participants were 59 undergraduates and 24 economics PhD students, and they played the same game under the same conditions that were described to their Prolific opponents. Each session implemented a single treatment: undergraduates and PhD students participated in separate sessions, with and without AI access. In all sessions, participants had no access to websites, AI tools other than the one we provided, or their phones.

In the AI treatments, each participant's decision screen included a chat window connected to the `gpt-5.1` model served through OpenAI's API, which was the only resource available beyond the experimental interface. The assistant was initialized with the game description and a general instruction to assist the participant. Participants could send as many or as few messages as they wished, including none. In the treatments without AI, the statement we made to Prolific opponents, that their matched player would have no access to any websites, AI tools, or their phone, was therefore literally true, and in the AI treatments the chat window was the only addition. After their decision, lab participants answered the same cognitive uncertainty questions, the same Raven matrix questions, the same two survey questions about their use of AI at work or study and in personal life, and a free-text question describing the strategy behind their choice.

3.3 Discussion on Design Choices

Hybrid Approach Our experimental design matches Prolific participants directly with lab participants rather than using an all-Prolific or an all-lab sample. This *hybrid approach* offers several advantages for identifying how beliefs about AI's effect on strategic sophistication vary with age, and for measuring them on a population that extends well beyond a student pool.

We recruit the main subjects in our study, those forming beliefs about AI's effect, from Prolific in order to have a sample with sufficiently varied characteristics. Prolific participants vary substantially in age, educational background, and professional experience. This heterogeneity is what we want to observe and control for. We can ask how different people, with different priors about AI, believe it affects strategic reasoning. A purely lab-based sample of undergraduate students would provide insufficient variation in these beliefs.

While we want heterogeneous beliefs, we need homogeneous implementation of the AI treatment. If we matched Prolific participants against other Prolific participants, we would lose control over the environment. Prolific participants could access ChatGPT, Claude, other AI tools, web searches, calculators, or other people, and we could neither prevent nor verify such access. More problematically, we could not credibly communicate to their matched opponents what tools they did or did not have access to.

The laboratory environment solves this problem. Lab participants complete the task in a controlled setting, where we directly observe and restrict their access to external resources. In the AI treatment, we provide exactly one AI tool through a standardized interface; in the control treatment, we ensure zero AI access. Importantly, we can truthfully communicate these conditions to Prolific participants. When we tell a Prolific participant that their matched player will have no access to any websites, AI tools, or their phone, this is verifiably true. This credibility is essential: our measure of beliefs about AI works only if participants trust our characterization of their opponent's access.

Direct Design Our design employs direct matching between Prolific participants and lab participants, rather than, say, having Prolific participants predict games between two lab participants (one with AI, one without). This direct matching structure provides clean identification of beliefs about AI's effect on strategic sophistication without confounding them with other forces.

Consider the alternative, in which Prolific participants predict the outcome of a lab-versus-lab game with asymmetric AI access. If a Prolific participant chooses a lower number when predicting the choice of a subject who has access to AI, this could reflect two distinct beliefs: (1) they believe AI directly enhances strategic thinking, or (2) they believe the non-AI player will play differently because that player believes AI is superior, causing the AI player to counter-adjust. This creates an identification problem in which choices reflect compound beliefs about both AI's direct effects and others' beliefs about AI, a recursion that confounds our measurement.

Our direct matching design eliminates this confound, even if it is logistically more involved. When a Prolific participant plays against a lab participant with AI, their choice reveals their belief about AI's direct effect on strategic reasoning, filtered only through standard best-

response logic. They need not model how two other players would reason about each other's AI-influenced capabilities, nor predict the equilibrium of a third-party strategic interaction. They assess: "How does AI change my opponent's likely choice, and what is my best response?" This framing isolates what we seek to measure: whether people believe AI makes others more strategically sophisticated.

Two features of the game qualify this statement. First, the mapping from beliefs to requests is many-to-one. As we show in Section A.1, 19 is a best response to a wide range of beliefs about naive play and 20 is a best response to any sufficiently diffuse belief, so a moderate change in beliefs need not change the request. Second, and for the same reason, a request alone cannot distinguish a player who expects a naive opponent from a player who has stopped trying to predict the opponent. The pre-analysis plan therefore designates decision time, the frequency of requests of 20, and the certainty questions as additional measures, and our second study elicits beliefs directly.

3.4 Details on Implementation

The study was approved by Indiana University's Institutional Review Board and preregistered at the AEA RCT Registry under ID 0017375. The experiment was conducted on the Prolific platform on December 12 and 14, 2025, and at IELAB on December 15 and 16. It was implemented using the oTree platform (Chen et al., 2016), and lab participants were recruited using the ORSEE recruitment system (Greiner, 2015).

The pre-analysis plan specified 100 participants per treatment, 25 in each generation cell. Before launch, we increased the quotas to 60 participants per treatment-by-generation cell, implying a target of 960. Recruitment proceeded in two batches, on December 12 and 14, before any laboratory session had taken place; 960 participants completed the study. Participants who required more than three incorrect attempts across the comprehension questions were excluded from the analysis.

The natural minimal effect of interest in the 11-20 game is one level of reasoning, corresponding to a one-point shift in requests. The pooled standard deviation of requests in our sample is 2.13, close to the values in s (Arad and Rubinstein, 2012). Against this benchmark, the realized design yields minimum detectable effects (80 percent power, $\alpha = 0.05$, two-sided) of 0.41 points for the pooled AI contrast, 0.55 to 0.60 points for pairwise treatment contrasts, 0.81 points for the AI by expertise interaction, and 1.1 to 1.3 points for generation level comparisons. We accordingly report all null results with equivalence bounds.

At the end of each lab session, each lab participant was matched with one randomly selected Prolific participant from the corresponding treatment, and payments were determined according to the rules of the game. After the completion of the lab experiment, each Prolific participant was matched with one randomly selected lab participant from the corresponding treatment,

with replacement. Each Prolific participant received a participation fee of $2.5 and had a 25 percent chance of being selected for a bonus payment, which included earnings from the game and from the Raven matrix questions. Selected participants were paid on December 19. Prolific participants spent about eight and a half minutes. Lab participants received a $5 show-up fee and the average lab payment was $17.25. The average bonus among the Prolific participants selected for payment was $6.1.

4 Main hypotheses

Our primary outcome variable is R_{ij}: Amount requested by Prolific participant i when facing an opponent from treatment $j \in \{\text{Novice}, \text{Novice+AI}, \text{Expert}, \text{Expert+AI}\}$. We state three main hypotheses below.

Hypothesis 1 The average requests are lower when participants face opponents aided by AI.

Finding support for hypothesis 1 and seeing the average lower requests against opponents with AI can indicate beliefs that opponents are more strategically sophisticated. At the same time, our analysis of the game suggests that there can also be an alternate story. If there is enough mass on the opponent choosing lower numbers or there is high diffusion over the beliefs of opponents' choices, making higher requests is an optimal choice. There is a possibility of what we call "giving up." Giving up would likely mean subjects request 20 or close to it.[5] For this reason, the pre-analysis plan also specifies that we examine the frequency of requests of 20, response times, and the answers to the certainty questions, which can help reveal the motives behind particular requests.

Hypothesis 2 The average requests when facing AI-assisted novice opponents (undergraduates + AI) are higher than those when facing experts without AI (PhD students).

Finding support for Hypothesis 2 would suggest that people do not believe AI assistance can elevate novice strategic behavior to levels comparable to unassisted experts.

Hypothesis 3 Treatment effects vary by participant generation.

Prolific participants were sampled to ensure equal representation from four generational cohorts: Gen Z, Millennials, Gen X, and Baby Boomers. We estimate interactions between treatment and generation indicators to test whether responsiveness to AI access of others differs systematically across age groups. We remain agnostic about the direction of heterogeneity but consider it plausible given documented generational differences in AI use and trust.

[5]The analysis of endogenous depth of reasoning a la Alaoui and Penta (2016) also predicts high requests, 19 or 20, against diffuse beliefs of others' levels of reasoning. Hence, higher requests alone might not be sufficient to reveal beliefs about the opponent's strategic capabilities with and without AI assistance.

Table 2. Requests by Treatment

Treatment	n	Mean	SD	Median	Mean (excl. 20)	$P(19)$	$P(20)$
Novice	219	17.33	2.03	18	17.11	0.274	0.078
Novice+AI	219	17.30	2.04	18	16.93	0.192	0.119
Expert	214	17.20	2.21	18	16.86	0.248	0.107
Expert+AI	219	17.17	2.25	18	16.94	0.269	0.078
All	871	17.25	2.13	18	16.96	0.246	0.095

Notes: Mean (excl. 20) is the primary outcome, the average request excluding requests of 20. $P(19)$ and $P(20)$ denote the fraction of participants requesting 19 and 20, respectively. Treatments differ in the described opponent: an undergraduate student (Novice) or an economics PhD student (Expert), with or without access to an AI tool on their decision screen.

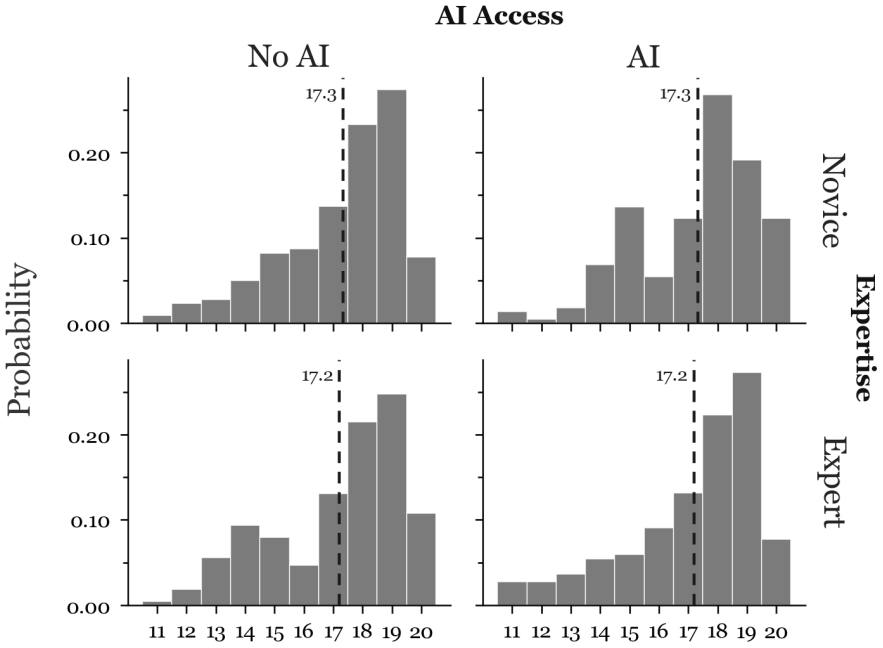


Figure 1. Request distributions by treatments (vertical line shows average)

5 Results

We begin by looking at aggregate levels of requests. The pooled distribution (mean 17.25, standard deviation 2.13) is similar to the one documented in Arad and Rubinstein (2012). Table 2 presents requests by treatment, and Figure 1 displays the full distributions. As we see in the Table 2, the average requests in all 4 treatments are similar and around 17.15. The median is 18 in all four treatments. We cannot reject the hypothesis that requests are drawn from the same distribution across the four treatments. All Bonferroni adjusted pairwise comparisons yield $p = 1$.

Pooling over expertise, requests against AI-assisted opponents are statistically indistinguishable from requests against unassisted opponents, and the same conclusion holds within expert

opponents and within novice opponents. Given our sample size and the data, the 90 percent confidence interval rules out a reduction larger than 0.27 points. The within expert and within-novice bounds are 0.38 and 0.36 points.

Result 1. Average requests against AI-assisted opponents and against unassisted opponents are statistically indistinguishable.

We then move to compare average requests against an AI assisted novice and against an unassisted expert. If latter requests are lower, it would indicate that subjects do not regard AI assistance as a full substitute for expertise. We find that the requests against the two opponents are statistically identical.[6]

Result 2. Requests against an AI assisted novice are statistically indistinguishable from requests against an unassisted expert (17.30 versus 17.20).

5.1 Deliberation times

The invariance of choices does not appear to reflect inattention to treatment assignment. Decision times are broadly similar in the Novice, Novice+AI, and Expert treatments, but are longer in Expert+AI. Consistent with this pattern, Figure 2 shows that the decision-time distribution in Expert+AI first-order stochastically dominates the distributions in the other treatments.

The regression results in Table 3 reinforce this pattern. Neither AI access nor expert status alone is associated with a statistically discernible change in decision time. By contrast, their interaction is positive and statistically significant in both the OLS and median regressions. Longer deliberation is therefore concentrated in the treatment that combines an expert opponent with AI assistance; AI access for novices alone is not associated with additional decision time.

This pattern is also present across cohorts. In each cohort, the increase in median decision time is larger when AI assistance is assigned to the expert than when it is assigned to the novice. We find no evidence that the treatment interaction varies across cohorts.

Result 3. Facing an AI-assisted expert increases median decision time by about 23 percent.

The additional deliberation does not show up in what participants say about their own uncertainty. The two cognitive uncertainty questions, certainty about the decision and certainty about receiving the bonus, do not differ across treatments (KW $p = 0.69$ and $p = 0.78$); decision certainty averages 64 on the 0-100 scale against an AI-assisted expert and 64 against an unassisted novice. Within treatments, certainty tracks the request rather than the opponent: participants who request 20 or 19 report decision certainty near 70, those who request 18 report

[6]The Kolmogorov-Smirnov [KS] test does not reject equality of the full distributions ($p = 0.70$).

Table 3. Decision time by treatment

	OLS: log time		Median: time (s)	
	(1)	(2)	(3)	(4)
AI	0.061	0.059	1.0	3.0
	(0.072)	(0.071)	(3.5)	(3.2)
Expert	−0.002	−0.005	−0.5	2.5
	(0.076)	(0.075)	(3.5)	(3.3)
AI × Expert	0.230**	0.233**	14.9***	14.2***
	(0.100)	(0.099)	(4.9)	(4.6)
Constant	3.582***	3.700***	35.2***	41.2***
	(0.054)	(0.068)	(2.4)	(3.1)
Generation FE	No	Yes	No	Yes
Observations	873	873	873	873

Note: Columns (1)–(2): OLS; columns (3)–(4): median regression. Robust (HC1) standard errors in parentheses; kernel-based for columns (3)–(4). $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

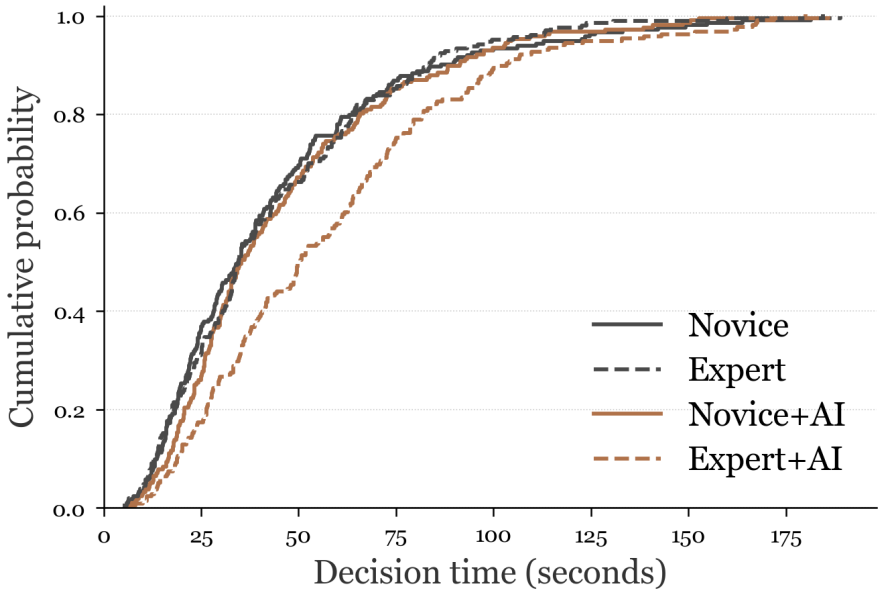


Figure 2. Decision Time by Opponent Type (≤ 200 seconds subsample)

Notes: Decision time is measured as the time participants spent on the page displaying the opponent-type description and making their decision. The Kruskal–Wallis test rejects equality across the four treatment groups, $\chi^2(3) = 27.47$, $p < 0.001$ ($N = 873$).

63, and those who request 17 or less report 59, so the deeper a participant reasons, the less sure she says she is. The bonus-certainty question turns out to be uninformative for our purposes.

5.2 Generations

We turn our attention to the core interest, which is whether sentiments towards AI assistance vary by generation. Firstly, the in aggregate, a joint test of the nine treatment by generation interaction terms rejects the equality claim ($F = 2.49$, $p = 0.008$).[7]

The heterogeneity is concentrated among treatments against expert opponents. Within the expert treatments, the AI by generation interaction is significant ($p = 0.001$). Within the novice treatments, there is no effect ($p = 0.93$). The three way test of whether the expert-side and novice-side patterns differ from each other rejects at $p = 0.048$. The pattern is also stable where it matters: the within-expert interaction is significant in each recruitment batch separately ($p = 0.041$ and $p = 0.024$)[8]

Table 4. Generational Responses to AI Among Expert Opponents

Panel A. AI minus no AI, expert opponents

Generation	Difference	SE	p (t test)	p (MWU)
Gen Z	−0.381	0.484	0.433	0.554
Millennial	+1.016	0.374	0.007	0.045
Gen X	+0.244	0.419	0.562	0.493
Baby Boomer	−1.250	0.436	0.005	0.012

Panel B. Deviation of expert cells from the own-generation novice baseline

Generation	d_{Exp}	p (MWU)	$d_{\text{Exp+AI}}$	p (MWU)
Gen Z	+0.567	0.183	+0.186	0.471
Millennial	−0.860	0.035	+0.156	0.816
Gen X	−0.520	0.147	−0.276	0.525
Baby Boomer	+0.726	0.015	−0.524	0.406

Notes: Generations follow the Pew Research Center birth-year definitions applied to age at the time of the study (Gen Z ≤ 28, Millennials 29-44, Gen X 45-60, Baby Boomers 61 and older; four participants aged 80-81 are included with the Baby Boomers). Panel A reports the difference in average requests between Expert+AI and Expert within each generation (44 to 62 participants per cell; novice cells have 39 to 75). Panel B reports the deviation of each expert cell from the same generation's pooled novice cells. Cell-level estimates are descriptive; minimum detectable effects at these cell sizes are 1.1 to 1.3 points.

Table 4 decomposes the AI contrast among participants facing expert opponents by generation

[7] An ordered logit specification yields $p = 0.031$, and the conclusion is unchanged under self-reported generation ($p = 0.009$), under age cutoffs that place 29-year-olds with Gen Z ($p = 0.027$), and on all 960 completed observations ($p = 0.049$).

[8] Recruitment proceeded in two batches (December 12 and 14, 2025); generational composition is balanced across them. The average treatment-by-generation interaction is unaffected by allowing all treatment and generation effects to differ freely across batches ($p = 0.008$).

and compares responses in each expert condition with the same-generation novice baseline. The clearest differences emerge among Millennials and Baby Boomers, and the two groups respond in opposite directions. Millennials make fewer requests against an unassisted expert than against novices, whereas Baby Boomers make more. With AI assistance, the estimates for both groups move toward their respective novice baselines. Generation X shows a weaker pattern in the same direction as Millennials, while Gen Z shows no clear response to the opponent descriptions.

Across all generations, the Expert+AI estimates are closer to the novice baseline than the Expert estimates, and none is statistically distinguishable from that baseline. These cell-level comparisons are descriptive, however, and the sample sizes limit precision; the results are consistent with convergence toward novice responses but do not establish equivalence. The generational variation appears specific to choices: as Section **??** shows, every cohort deliberates longer when facing an AI-assisted expert.

The cohort split also places our expertise null in the literature. When student subjects are told that their opponent is more sophisticated, they play deeper (Agranov et al., 2012; Alaoui and Penta, 2016). That is what Millennials do here, and what Generation X does in attenuated form. The mean does not move because Baby Boomers move the other way and Generation Z does not move at all. The canonical opponent-label effect replicates in one cohort, reverses in another, and is absent in a third.

Result 4. Treatment effects differ across generations ($p = 0.008$), and the heterogeneity is concentrated among expert opponents ($p = 0.001$). Baby Boomers request more against unassisted experts and less against AI-assisted experts. Millennials request less against unassisted experts, and the difference disappears when the expert has AI. Every cohort treats an AI-assisted expert like a novice.

5.3 Movement in the tails

While average requests are invariant, the tails of the distribution move in directions consistent with the opponent descriptions. Against expert opponents, the share of requests at or below 14, choices that pay only against an opponent reasoning several steps deep, rises from 0.107 to 0.159 ($p = 0.031$). Against AI-assisted novices, the share choosing 19, the best response to a naive opponent who requests 20, falls from 0.274 to 0.192 ($p = 0.055$), as if the AI tool makes the naive anchor less reliable. The probability of requesting 20 displays a similar crossover (interaction coefficient -0.829 in a logit, $p = 0.077$). None of the pairwise distribution tests rejects (simulated χ^2, $p = 0.107$ for Novice+AI versus Novice and $p = 0.218$ for Expert+AI versus Expert), so we read these margins as suggestive: small groups of subjects appear to shift their beliefs in opposite directions that offset in the mean.

Result 5. Against AI-assisted novices, the level-1 reply of 19 becomes less frequent (0.192 versus 0.274, $p = 0.055$). Against expert opponents, deep requests of 14 or below become more frequent (0.159 versus 0.107, $p = 0.031$).

6 Concluding discussion

The treatments move deliberation but not average choices. Three accounts are consistent with this pattern, and the data speak to each.

The first is that beliefs do not move: participants read the opponent description, think about it, and conclude that it does not change what the opponent will do. The lab arm gives this account some support, because the assistant did not in fact change what lab participants requested, and a Prolific participant who expected as much would be right. But it sits uneasily with the cohort results, in which the same description moves Baby Boomers and Millennials by more than a point in opposite directions, and with the tails.

The second is that beliefs move in offsetting directions. Section 5.3 shows the level-1 reply of 19 becoming rarer against an AI-assisted novice while deep requests become more common against experts, and Section 5.2 shows cohorts moving in opposite directions. Under this account the average is flat because the population is mixed, not because individuals are unmoved.

The third is that beliefs move but do not cross the threshold at which the best response changes. In the 11-20 game 19 is a best response to a wide range of beliefs about naive play, and 20 is a best response to any sufficiently diffuse belief (Section A.1). The model of endogenous depth of reasoning of Alaoui and Penta (2016) makes this precise. A player takes one more step of reasoning when the subjective probability that the opponent stops at that step exceeds the probability that she stops one step earlier by enough to cover the cost of the step. Information that makes the opponent's reasoning deeper but predictable shifts probability mass down the ladder and moves the best response with it. Information that makes the opponent's reasoning harder to forecast flattens the mass, so that no step beyond the first pays at any cognitive cost, and the optimal request stays at 19 or retreats to the belief-free 20. Under the second kind of information a player can spend a long time trying to form a belief and end up where she started. The data have that signature: deliberation rises most where choices move least, the participants who request 20 against an AI-assisted expert deliberate longest of all, and the participants who deliberate longest, those facing an AI-assisted expert, are no more certain of their choice than anyone else. A related mechanism would produce the same pattern through a different channel. Ngangoue et al. (2026) find that AI advice homogenizes choices; a participant who expects the expert to follow the assistant, and expects the assistant to say what it says to everyone, has reason to expect the expert to play like everyone else, which is how every cohort treats the AI-assisted expert in Table 4.

References

Açıkgöz, B., Dubois, D., Duchêne, S., Nguyen-Huu, A., and Willinger, M. (2024). Depth of reasoning in the 11–20 game differs between financial professionals and students. A lab-in-the-field experiment. *Economics Letters*, 239. VERIFY article number.

Agranov, M., Potamites, E., Schotter, A., and Tergiman, C. (2012). Beliefs and endogenous cognitive levels: An experimental study. *Games and Economic Behavior*, 75(2):449–463.

Alaoui, L., Janezic, K. A., and Penta, A. (2020). Reasoning about others' reasoning. *Journal of Economic Theory*, 189:105091. VERIFY article number.

Alaoui, L. and Penta, A. (2016). Endogenous depth of reasoning. *The Review of Economic Studies*, 83(4):1297–1333.

Arad, A. and Rubinstein, A. (2012). The 11–20 money request game: A level-k reasoning study. *American Economic Review*, 102(7):3561–3573.

Autor, D., Rodchenko, T., Martin, J., Iscenko, Z., Strand, S., Pearl, D., and Ferere, M. (2026). Does ai assistance enhance or erode expertise? evidence from a three-month field experiment in patent drafting. Technical report, National Bureau of Economic Research.

Autor, D. and Thompson, N. (2025). Expertise. *Journal of the European Economic Association*, 23(4):1203–1271.

Barak, D. and Costa-Gomes, M. (2025). Humans expect rationality and cooperation from LLM opponents in strategic games. arXiv:2505.11011.

Brynjolfsson, E., Chandar, B., and Chen, R. (2025a). Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence. Working paper, Stanford Digital Economy Lab. VERIFY.

Brynjolfsson, E., Li, D., and Raymond, L. (2025b). Generative AI at work. *Quarterly Journal of Economics*, 140(2):889–942. VERIFY pages.

Camerer, C. F., Ho, T.-H., and Chong, J.-K. (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics*, 119(3):861–898.

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Costa-Gomes, M. A. and Crawford, V. P. (2006). Cognition and behavior in two-person guessing games: An experimental study. *American Economic Review*, 96(5):1737–1768.

Crawford, V. P., Costa-Gomes, M. A., and Iriberri, N. (2013). Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *Journal of Economic Literature*, 51(1):5–62.

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S.,

Krayer, L., Candelon, F., and Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Working Paper 24-013, Harvard Business School.

Di Cagno, D. and Lin, L. (2025). How do individuals interact with an ai advisor in strategic reasoning? an experimental study in beauty contest. *Journal of Economic Behavior & Organization*, 237:107159.

Dvorak, F., Stumpf, R., Fehrler, S., and Fischbacher, U. (2025). Adverse reactions to the use of large language models in social interactions. *PNAS Nexus*, 4(4):pgaf112.

Enke, B. and Graeber, T. (2023). Cognitive uncertainty. *The Quarterly Journal of Economics*, 138(4):2021–2067.

Gao, Y., Lee, D., Burtch, G., and Fazelpour, S. (2025). Take caution in using llms as human surrogates. *Proceedings of the National Academy of Sciences*, 122(24):e2501660122.

Goeree, J. K., Louis, P., and Zhang, J. (2018). Noisy introspection in the 11–20 game. *The Economic Journal*, 128(611):1509–1530.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with orsee. *Journal of the Economic Science Association*, 1(1):114–125.

Hitzig, Z., Massenkoff, M., Lyubich, E., Heller, R., and McCrory, P. (2026). Agentic coding and persistent returns to expertise. *Anthropic Research*.

Kader, G. and Lee, D. (2025). The emergence of strategic reasoning of large language models. arXiv:2412.13013.

Li, K. K. and Rong, K. (2018). Choices in the 11–20 game: The role of risk aversion. *Games*, 9(3):53.

Lindner, F. and Sutter, M. (2013). Level-k reasoning and time pressure in the 11–20 money request game. *Economics Letters*, 120(3):542–545.

March, C. (2021). Strategic interactions between humans and artificial intelligence: Lessons from experiments with computer players. *Journal of Economic Psychology*, 87:102426. VERIFY article number.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economic Review*, 85(5):1313–1326.

Ngangoue, K., Schotter, A., and Wang, B. (2026). Life's big decisions with AI advice. CESifo Working Paper 12515, CESifo. SSRN 6354640; original version February 2026, this version August 2026.

Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

Otis, N. G., Clarke, R. P., Delecourt, S., Holtz, D., and Koning, R. (2024). The uneven impact of

generative AI on entrepreneurial performance. SSRN Working Paper 4671369. VERIFY SSRN number.

Palacios-Huerta, I. and Volij, O. (2009). Field centipedes. *American Economic Review*, 99(4):1619–1635.

Schotter, A. (2023). Placeholder – the technical / naive / wise advice typology cited by ngangoue, schotter and wang (2026). TODO: supply the reference.

Shen, C., Karpus, J., Kosch, T., Fernandes, D. P. B., Mello, B., Welsch, R., and Villa, S. (2025). The impact of asymmetric AI assistance on decision-making in social dilemmas: A study on human augmentation in economic games. In *Proceedings of the Augmented Humans International Conference 2025 (AHs '25)*, pages 187–198. Association for Computing Machinery.

Stahl, D. O. and Wilson, P. W. (1995). On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1):218–254.

Wang, Z., Ahmed, I., Unjitwattana, P., Xie, E. Y., Fong, M.-J., and Lin, P.-H. (2026). Revisiting the intra-team communication method to elicit level-k reasoning in beauty contests and 11-20 games. *arXiv preprint arXiv:2608.18069*.

A Appendices

A.1 Additional Analyses

Let q_r denote the probability that the opponent requests amount r, where $r \in \{11, 12, ..., 20\}$ and $\sum_{r=11}^{20} q_r = 1$. Expected payoff in this game can be written as $R + 20 \cdot q_{R+1}$, where R is the requested amount and q_{R+1} is the probability the opponent requests $R + 1$.

We ask what are the beliefs about other's play that make choosing 20 optimal? We solve for the beliefs about others choices and we get that choosing 20 is optimal if: $q_r \leq \frac{21-r}{20}, \forall\, r \in \{12, ..., 20\}$ and $\sum_{r=11}^{20} q_r = 1$. Note, these conditions are absolute bounds on probabilities, not in differences as for any other strategy but 20. Let us elaborate. Choosing 19 is optimal if the belief about any action r does not exceed the belief about action 20 by more than $5(20 - r)\%$: $q_r - q_{20} \leq \frac{20-r}{20}, \forall\, r \in \{12, ..., 19\}$ and $q_{20} \geq 5\%$.[9]

Note, we are not looking for an equilibrium here. We are looking at beliefs about other's requests that make choosing a certain strategy a best response. See Figure 3 for visualizing the restriction on the distribution and 3 example distributions under which choosing 20 is optimal.

The expected payoff $R + 20 \cdot q_{R+1}$ decomposes into a guaranteed component R and a bonus component $20 \cdot q_{R+1}$. Each unit decrease in R sacrifices one guaranteed point in exchange for targeting a different opponent action for the bonus. The optimal choice depends on where the player believes the opponent's probability mass is concentrated.

When facing an AI-equipped opponent, the unpredictability of the opponent's reasoning process may lead to diffuse beliefs across the strategy space. As the subjective probability q_{R+1} shrinks uniformly across all R, the guaranteed component dominates, making higher requests more attractive. This yields a testable prediction: subjects facing AI-equipped opponents should choose higher values of R on average.

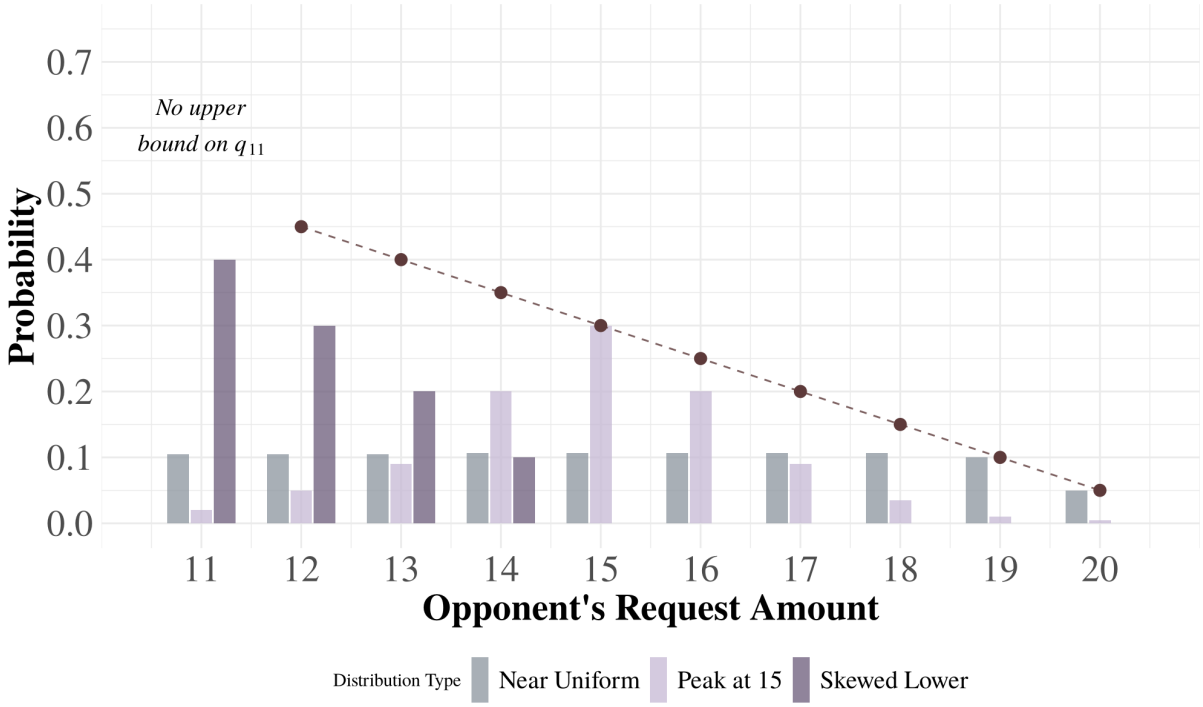
However, higher choices of R are observationally consistent with two opposing beliefs about the opponent. A subject might choose 19 because they believe the AI-equipped opponent will naively choose 20, expecting a high probability of winning the bonus. Alternatively, a subject might choose 20 because they find the AI-equipped opponent's reasoning entirely unpredictable, expecting a low probability of winning the bonus. Both behaviors manifest as high requests, but the underlying beliefs are opposed.

[9] Footnote 2 in Arad and Rubinstein (2012) states that "strategy 19 remains a best response to any belief that does not assign to any action $20 - i$ a probability that is higher by $5i$ percent than the probability assigned to the strategy 20."– which is equivalent to our condition above for 19 being the best response.

Table 5. Distribution of Choices Across Studies (%)

Choice	11	12	13	14	15	16	17	18	19	20	Optimal
Arad and Rubinstein (2012)	4	0	3	6	1	6	32	30	12	6	17
Goeree et al. (2018) (11-20 game)	1	0	1	3	4	4	15	40	22	10	17
Goeree et al. (2018) (1-10 game)	2	1	1	1	1	3	8	36	35	11	18
Li and Rong (2018) (11-20 game)	0	0	0	7	1	4	19	36	25	4	

Figure 3. Beliefs that make 20 optimal



Treatment	11	12	13	14	15	16	17	18	19	20
NoAI	1	1	0	1	2	1	3	3	4	4
AI	0	0	1	1	6	1	2	5	4	0

Table 6. Requests by Treatment

A.2 Endogenous Depth of Reasoning

We apply the framework of Alaoui and Penta (2016) to generate predictions about how beliefs regarding opponent AI access affect strategic reasoning.

A.2.1 Setup

Consider the 11-20 game $G = (A_i, u_i)_{i=1,2}$ where $A_i = \{11, 12, \ldots, 20\}$ and payoffs are as described before. Let $BR_i : A_j \to A_i$ denote player i's best response function.

A player's reasoning is represented by a sequence of action profiles $\{(a_1^k, a_2^k)\}_{k\in\mathbb{N}}$ such that $a_i^{k+1} = BR_i(a_j^k)$ for each $k = 0, 1, \ldots$. The anchor $a^0 = (a_1^0, a_2^0)$ represents what players would choose by default, without strategic reasoning.

In the 11-20 game, the natural anchor is $a_i^0 = 20$ for both players. The path of reasoning then follows undercutting the other player by just 1.

A.2.2 Marginal Benefit of Reasoning

Let p_j denote player i's subjective probability that the opponent stops reasoning at step j (and thus plays $a_j^j = 20 - j$), with $p_{-1} \equiv 0$.

If player i stops at step k, they play $20 - k$. They win the bonus if the opponent plays exactly one more than their request, i.e., $(20 - k) + 1 = 21 - k$. The opponent plays $21 - k$ when they stop at step $k - 1$. Thus, the expected payoff from stopping at step k is:

$$\pi(k) = (20 - k) + 20 \cdot p_{k-1}$$

where the first term is the guaranteed payoff from requesting $20 - k$ and the second term is the expected bonus.

The marginal benefit of advancing from step k to step $k + 1$ is:

$$\begin{aligned} MB(k) &= \pi(k+1) - \pi(k) \\ &= (20 - k - 1) + 20 \cdot p_k - (20 - k) - 20 \cdot p_{k-1} \\ &= -1 + 20 \cdot (p_k - p_{k-1}) \end{aligned}$$

Let $c \geq 0$ denote the cognitive cost per step of reasoning. A player advances from step k to

step $k+1$ if and only if:

$$\begin{aligned} MB(k) &> c \\ -1 + 20 \cdot (p_k - p_{k-1}) &> c \\ p_k - p_{k-1} &> \frac{1+c}{20} \end{aligned}$$

Equilibrium Reasoning Depth. The player continues reasoning as long as the subjective probability that the opponent is at step k sufficiently exceeds the probability that the opponent is at step $k-1$. The player stops at step k^* and plays $a_i^{k^*} = 20 - k^*$, where k^* is the largest k such that $MB(j) > c$ for all $j < k$. Equivalently, reasoning stops at the first step where:

$$p_{k^*} - p_{k^*-1} \leq \frac{1+c}{20}$$

Peaked beliefs Suppose a player believes a human opponent is most likely to stop at step 1 or 2:

$$p_0 = 0.15, \quad p_1 = 0.40, \quad p_2 = 0.30, \quad p_3 = 0.10, \quad p_4 = 0.05$$

The marginal benefits are:

$$\begin{aligned} MB(0) &= -1 + 20 \cdot (p_0 - p_{-1}) = -1 + 20 \cdot 0.15 = 2 \\ MB(1) &= -1 + 20 \cdot (p_1 - p_0) = -1 + 20 \cdot 0.25 = 4 \\ MB(2) &= -1 + 20 \cdot (p_2 - p_1) = -1 + 20 \cdot (-0.10) = -3 \end{aligned}$$

For moderate cognitive costs (e.g., $c = 1$), we have $MB(0) = 2 > 1$ and $MB(1) = 4 > 1$, so the player advances through steps 0 and 1. Since $MB(2) = -3 < 1$, reasoning stops at step 2. The player chooses $a_i^2 = 18$.

Diffuse beliefs Suppose a player believes an AI-equipped opponent could stop at any step with equal probability:

$$p_j = 0.1 \quad \forall j \in \{0, 1, \ldots, 9\}$$

The marginal benefit of the first step is:

$$MB(0) = -1 + 20 \cdot (p_0 - p_{-1}) = -1 + 20 \cdot 0.1 = 1$$

For all subsequent steps $k \geq 1$:

$$MB(k) = -1 + 20 \cdot (p_k - p_{k-1}) = -1 + 20 \cdot 0 = -1 < 0$$

The optimal stopping point depends on cognitive costs. For $c < 1$, the player advances to step 1 and plays $a_i^1 = 19$. For $c \geq 1$, the player does not reason beyond the anchor and plays $a_i^0 = 20$.

Diffuse beliefs (AI-equipped opponent). Suppose a player believes an AI-equipped opponent could stop at any step with equal probability:

$$p_j = 0.1 \quad \forall j \in \{0, 1, \ldots, 9\}$$

The marginal benefit of the first step is:

$$MB(0) = -1 + 20 \cdot (p_0 - p_{-1}) = -1 + 20 \cdot 0.1 = 1$$

For all subsequent steps $k \geq 1$:

$$MB(k) = -1 + 20 \cdot (p_k - p_{k-1}) = -1 + 20 \cdot 0 = -1 < 0$$

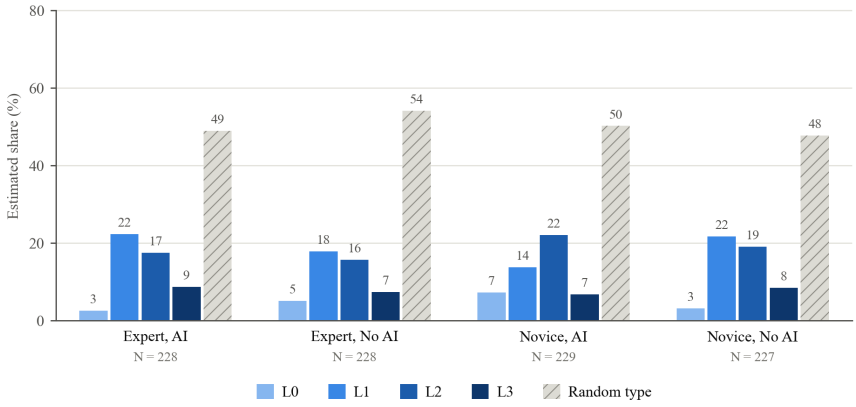
The optimal stopping point depends on cognitive costs:

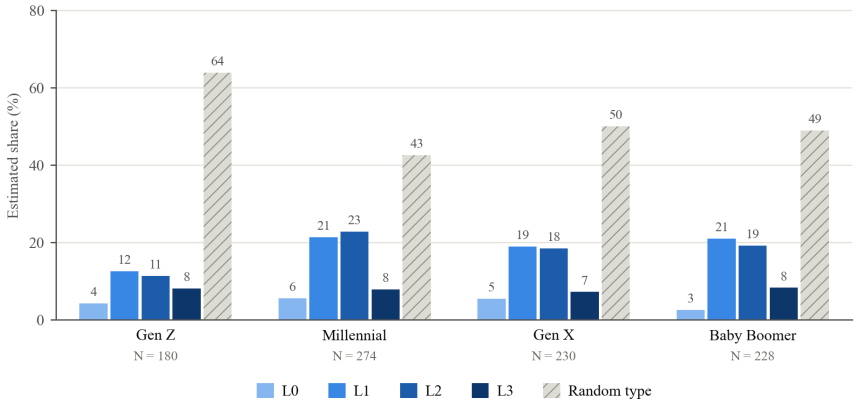
$$a_i^* = \begin{cases} 20 & \text{if } c > 1 \\ \{19, 20\} & \text{if } c = 1 \\ 19 & \text{if } c < 1 \end{cases}$$

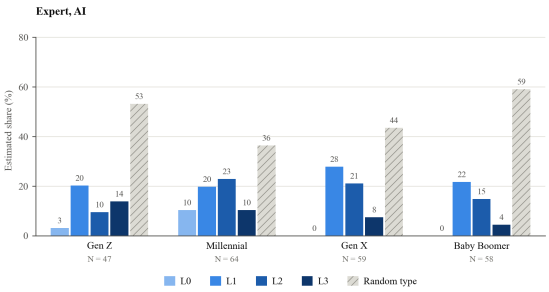
Under diffuse beliefs, reasoning beyond step 1 is never profitable regardless of cognitive costs. This contrasts sharply with peaked beliefs, where low cognitive costs support reasoning to step 2 or beyond.

A.3 Level-k MLE fit

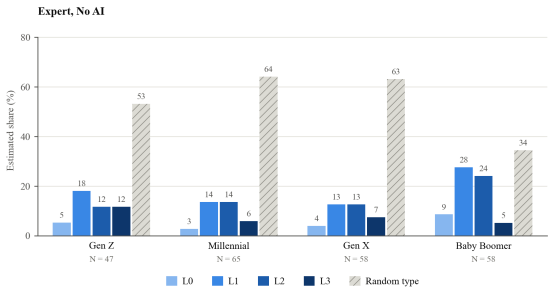
We adopt the maximum-likelihood estimation (MLE) procedure following Arad and Rubinstein (2012). Specifically, we consider five agent types: Level-0 through Level-3, as well as a uniform-random type. We estimate the population shares of each type that maximize the likelihood of the observed aggregate data generated in the experiment.



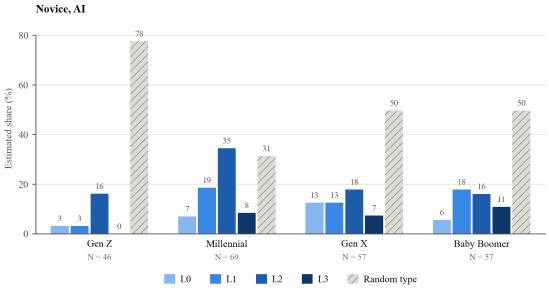




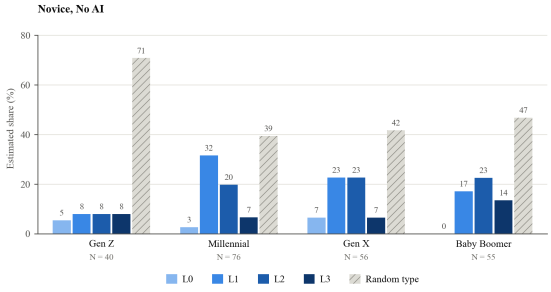
(a) Expert, AI



(b) Expert, No AI

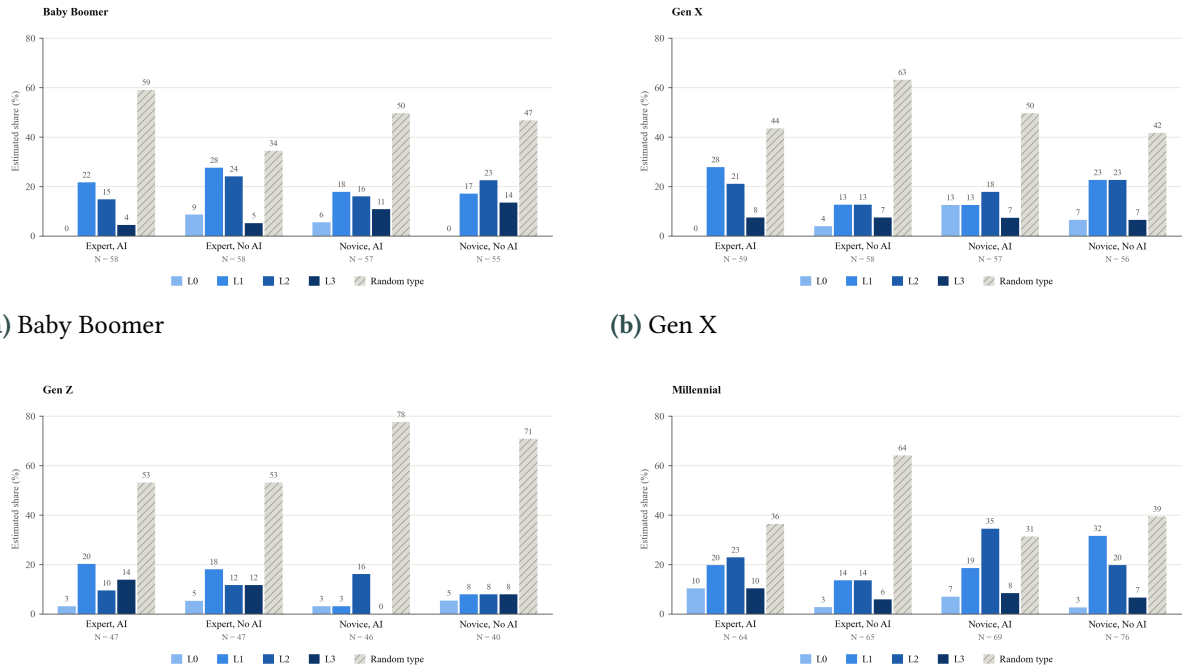


(c) Novice, AI



(d) Novice, No AI

Figure 4. Level-k type shares by generation, per treatment.



(a) Baby Boomer

(b) Gen X

(c) Gen Z

(d) Millennial

Figure 5. Level-k type shares by treatment, per generation.

B Graphs and Tables

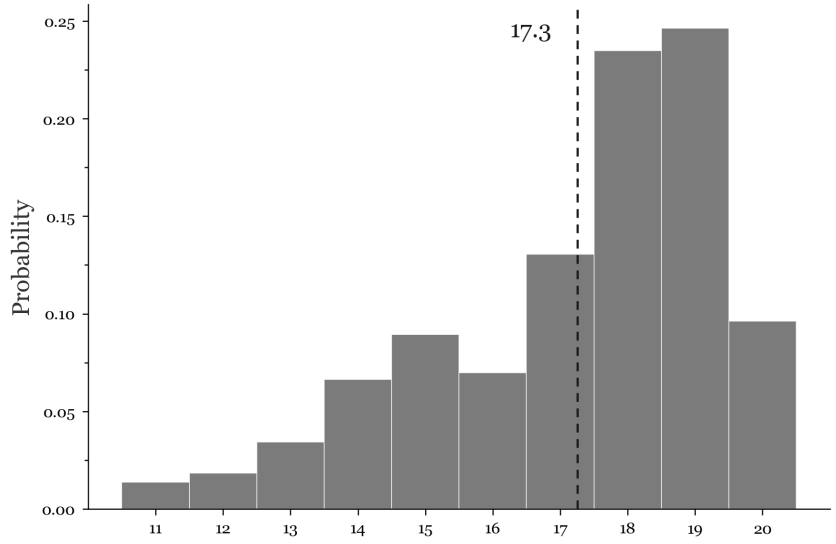


Figure 6. All request distribution (vertical shows average)