Pretending not to know reveals a powerful capacity for self-simulation

Matan Mazor[1], Chaz Firestone[2], & Ian Phillips[2]

[1] University of Oxford

[2] Johns Hopkins University

Author note

Correspondence concerning this article should be addressed to Matan Mazor, All Souls College, High Street, Oxford OX1 4AL. E-mail: matan.mazor@all-souls.ox.ac.uk

Abstract

Feigning ignorance is crucial in contexts as diverse as diplomacy, warcraft and personal relationships, each demanding strategic concealment of information. To be effective, such 'epistemic pretense' requires us to anticipate how we would behave with different knowledge, and then to act in accord with that counterfactual knowledge state. Decades of research on hindsight bias suggest that people are poor at appreciating how they would behave when ignorant. In stark contrast, here we discover a remarkable capacity to simulate decision-making under a counterfactual knowledge state, by comparing real and 'pretend' play in two large-scale gamified experiments. Subjects saw the full solution to a game (e.g., all ship locations in *Battleship*, or the hidden word in *Hangman*) but then attempted to play as though they didn't have this information. Impressively, they mimicked broad and subtle patterns of ordinary play, completely convincing peers of their ignorance. Nevertheless, computational modeling uncovered traces of 'over-acting' in their decisions, consistent with a schematic simulation of their minds. Opening up a new approach to studying self-simulation, our results reveal intricate metacognitive knowledge about decision-making, drawn from a rich—but simplified—internal model of cognition.

Keywords: pretense; metacognition; theory of mind

Word count: 7030

Pretending not to know reveals a powerful capacity for self-simulation

Introduction

Pretense relies on an ability to simulate and mimic one's own behavior under a counterfactual belief state. For example, in order to successfully deceive your friends into thinking that you were surprised by the birthday party they threw for you, it is not sufficient that you are able to reason about their mental states ("I know that they are planning a surprise party, but they don't know that I know that…") — you also need to convincingly simulate and mimic your hypothetical behavior had you not known about the party ("Where would I look first had I not known? What would I say? How long would it take me to recover from the surprise?"). This is not a trivial challenge: previous research on "hindsight biases" suggests that knowledge about the actual state of the world can interfere with our ability to correctly judge what we would have believed (Fischhoff, 1975, 1977; Roese & Vohs, 2012; Wood, 1978) or perceived (Bernstein & Harley, 2007; Bernstein, Wilson, Pernat, & Meilleur, 2012; Harley, Carlsen, & Loftus, 2004) without this knowledge. Such biases remain potent even when instructing participants to overcome them (Harley et al., 2004; Pohl & Hell, 1996). Moreover, even if pretenders can correctly determine what they would have believed, they must further accurately simulate how they would think and behave in this different belief state.

The reliance of this kind of epistemic pretense on self-simulation makes it a promising tool for revealing the structure and content of people's internal models of their own minds. When directly asked, participants are able to provide relatively accurate descriptions of their own decision-making (Morris, Carlson, Kober, & Crockett, 2023) and perception (Levin & Angelone, 2008; Mazor, Siegel, & Tenenbaum, 2023). Pretending not to know opens a new window into the structure and content of this metacognitive knowledge, with two important advantages. First, by not relying on explicit reports, pretense has the potential to reveal implicit self-knowledge – that is, structured knowledge about the self that is not reportable. And second, data obtained from pretense experiments can be analyzed and modeled using the same tools employed by cognitive scientists to study non-pretense behavior, affording a direct and finer-grained comparison between pretend and genuine decision-making.

To this end, we examine pretense in a game setting. Using an online version of the games Battleship and Hangman (in which players seek to uncover the locations of enemy ships or the identity of a word), participants played a 'non-pretend' (normal) version of the game, as well as a 'pretend' version where they were given complete information about the hidden ships / target word but were instructed to behave as if they didn't have this information. Overall, we find a highly impressive capacity for pretending not to know, with pretenders mirroring broad patterns and subtle features of real players' decisions and decision times. However, we also find that pretense was characterized by over-acting, stereotypical behavior, and suboptimal incorporation of new information, although these were undetected by human observers.

Results

In two experiments, players played online versions of two information-seeking games: Battleship (N=500 players) and Hangman (N=501 players). These games traditionally start in a state of ignorance, with a player's goal being to reveal an unknown world state (ship locations in Battleship, a hidden word in Hangman) in as few steps (cell or letter selections) as possible. Critically, in addition to playing five standard games, players in our experiments also completed five 'pretend' games in which the solution to the game was known to them from the start, but where their task was to behave as if it was not – i.e. to play as though they did not have this information (see Fig. 1). We measured participants' capacity to simulate a counterfactual state of ignorance by comparing patterns of decisions and decision times in pretend and non-pretend games. Our full pre-registered results are available online together with the report-generating code. Readers are invited to try demos of the experiments.

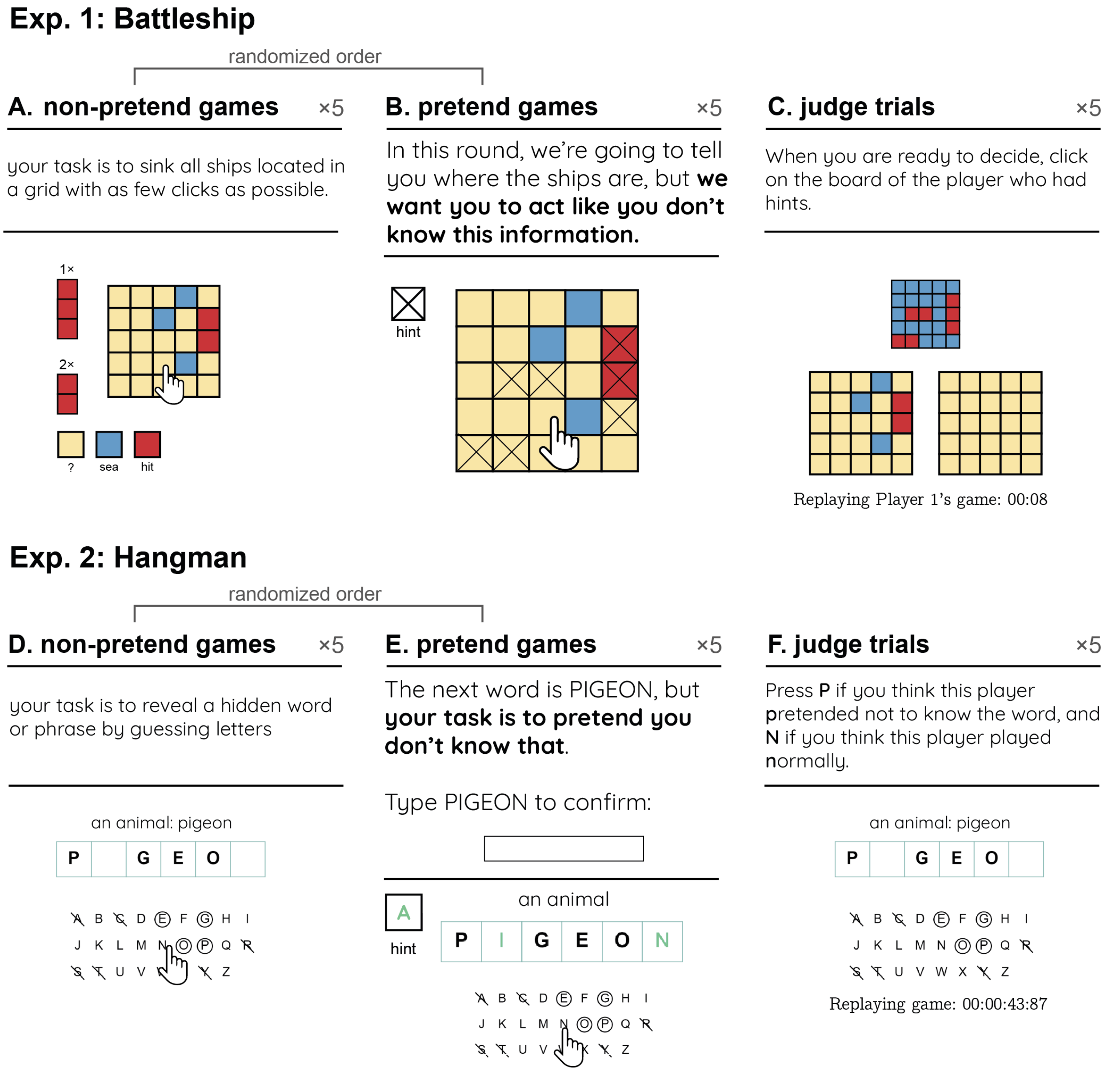
Exp. 1: Battleship
randomized order
A. non-pretend games ×5
your task is to sink all ships located in a grid with as few clicks as possible.
1×
2×
?
sea
hit
B. pretend games ×5
In this round, we're going to tell you where the ships are, but we want you to act like you don't know this information.
hint
C. judge trials ×5
When you are ready to decide, click on the board of the player who had hints.
Replaying Player 1's game: 00:08
Exp. 2: Hangman
randomized order
D. non-pretend games ×5
your task is to reveal a hidden word or phrase by guessing letters
an animal: pigeon
P G E O
A B C D E F G H I
J K L M N O P Q R
S T U V W X Y Z
E. pretend games ×5
The next word is PIGEON, but your task is to pretend you don't know that.
Type PIGEON to confirm:
A
hint
an animal
P I G E O N
A B C D E F G H I
J K L M N O P Q R
S T U V W X Y Z
F. judge trials ×5
Press P if you think this player pretended not to know the word, and N if you think this player played normally.
an animal: pigeon
P G E O
A B C D E F G H I
J K L M N O P Q R
S T U V W X Y Z
Replaying game: 00:00:43:87

***Figure 1**: Experimental Design in Exp. 1 (upper panel) and 2 (lower panel).* In non-pretend games, players revealed ships by guessing cells in a grid (A) or revealed a word by guessing letters (D). In pretend games, we marked ship locations with a cross (B) and revealed the target word from the start (E), but asked players to play as if they didn't have this information. Lastly, players watched replays of the games of previous players and guessed which were pretend games (C and F).

Measuring pretense quality

As a first measure of pretense quality, we compared the total number of guesses in pretend and non-pretend games. Among Battleship players, the number of cell selections was similar in pretend (mean

= 15.83) and non-pretend games (mean = 16.05; $p = .153$; Fig. 2A). 20 pretenders who immediately discovered all ships without making errors were excluded from all further analyses, in accordance with our pre-registered plan. With these subjects excluded, the number of cell selections remained very similar in pretend (mean = 16.11) and non-pretend games (mean = 15.94; $p =.164$; Fig. 2A)). In Hangman, pretenders tended to make about one additional letter guess on average than did non-pretenders, controlling for word length (pretend: 2.80 misses; non-pretend: 1.94 misses; $p < .001$; Fig. 2B). Despite an overall bias in the number of guesses, pretend Hangman games showed an impressive, item-specific alignment: pretenders were successful in making more letter guesses when attempting to reveal words that would have been harder to guess had they been playing for real ($r_s = .97$; Fig. 2B). This strong correlation provides evidence for a human capacity to act in accordance with a counterfactual knowledge state.

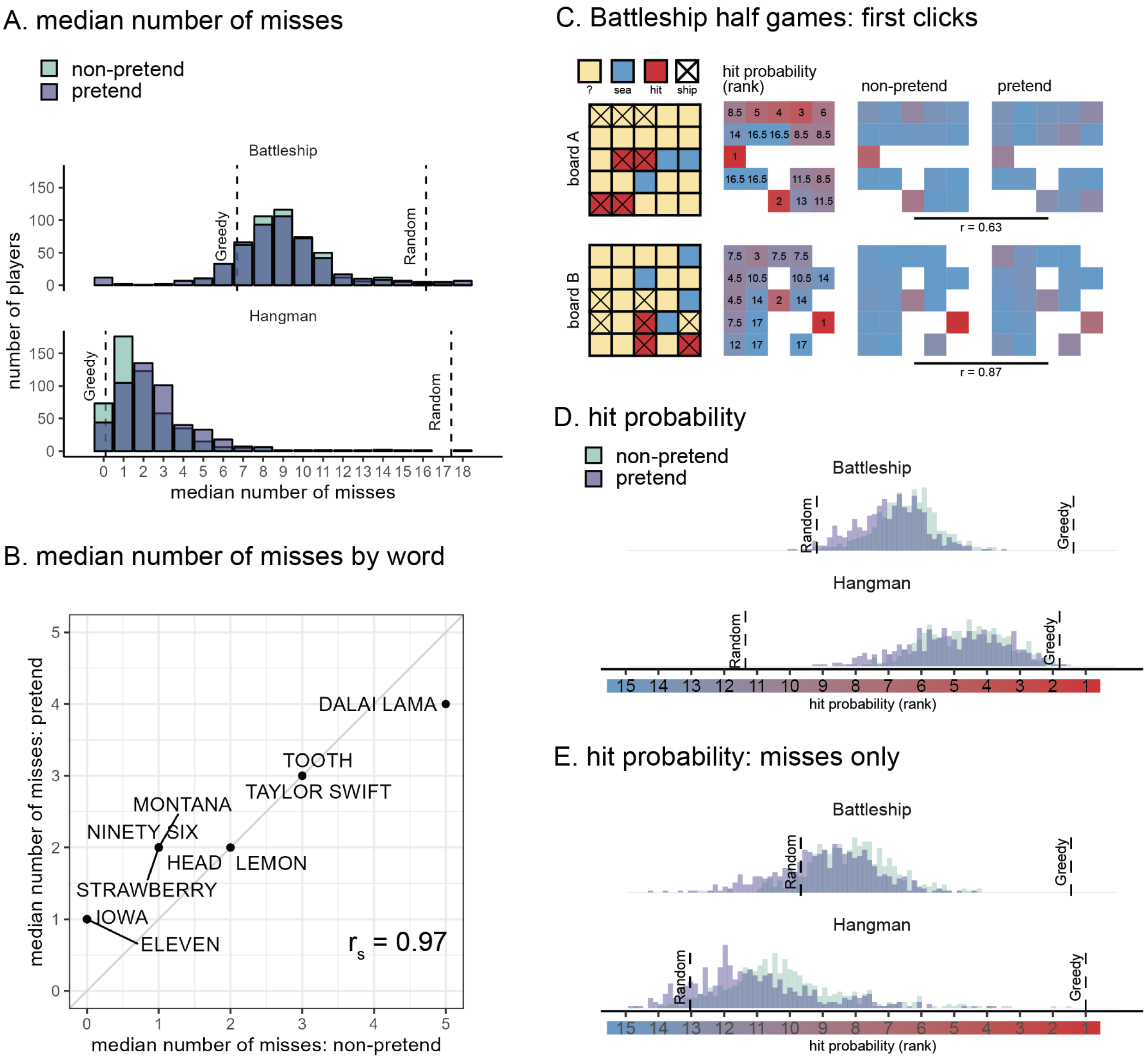

***Figure 2**: Battleship and Hangman guesses in pretend and non-pretend games.* A: median number of misses in Battleship and Hangman games, in non-pretend (green) and pretend (purple) games. For reference, the expected number of misses is indicated by a reference line for a fully random agent, and for a “greedy” agent that maximizes the probability of a hit in each step. B: The median number of misses in Hangman for pretend and non-pretend games, as a function of the target word. C: Spatial guess distributions for pretend and non-pretend half-games (where players continued the game from a half-finished state) alongside their corresponding hit probability maps. D: Cell and letter selections were ranked according to their relative hit probability. Plotting the median rank per subject in pretend and non-pretend games, with reference lines for the expected rank probability for a random agent, and for a “greedy” agent that maximizes the probability of a hit in each step. E: same as panel D but discarding all guesses that resulted in a hit.

Having established an alignment in the total number of guesses, we next turned to the content of pretend and non-pretend guesses. In order to directly compare pretend and non-pretend guesses for the same board state, Battleship players completed two half-games in which they were instructed to continue the game from a half-completed state. Hundreds of cell selections for the same board state revealed a strong correlation between the spatial distributions of pretend and non-pretend guesses (board A: $r = .63$, $p = < .001$; board B: $r = .87$, $p < .001$; Fig. 2C), confirming that pretenders were sensitive not only to the number of guesses they would have made had they been playing for real, but also to their content.

To further examine the decisional processes behind this strong alignment, we compared the degree to which pretend and non-pretend guesses made sense within the context of the game. When playing Battleship and Hangman, it makes sense to guess cells or letters for which the probability of hitting a ship or revealing a letter is high (this “greedy” behavior is not strictly optimal, but approximates optimal behavior in most cases Audinot, Bonnet, and Viennot (2014)). In the non-pretend versions of both games, guesses were more rational according to this measure than expected by chance (Battleship: $t(479) = 49.18, p < .001$, Hangman: $t(500) = 86.88, p < .001$). Despite being less rational than non-pretend guesses, pretend guesses were also more rational than expected by chance (Battleship: $t(479) = 38.51$, $p < .001$, Hangman: $t(500) = 72.29, p < .001$; Fig. 2D). Critically, the same was true when restricting the analysis to unsuccessful guesses (Battleship: $t(479) = 10.25, p < .001$, Hangman: $t(487) = 18.91$, $p < .001$; Fig. 2E): that is, even when incorrectly guessing a ship’s location or a letter’s identity, pretend guesses made sense given the limited information players pretended to have.

Good pretense is a function not only of the number and content of players’ decisions, but also of their timing. Here too, pretend games showed the same qualitative patterns as non-pretend games. Like non-pretenders, pretenders were faster in their successful guesses (difference in decision time between hits and misses in Battleship: $\Delta_{non-pretend}$ = -109 ms, $\Delta_{pretend}$ = -293 ms; Hangman: $\Delta_{non-pretend}$ = -386 ms, $\Delta_{pretend}$ = -297 ms) and slowed down immediately after a hit (difference in decision time between guesses that followed hits versus misses in Battleship: $\Delta_{non-pretend}$ = 182 ms, $\Delta_{pretend}$ = 236 ms; Hangman: $\Delta_{non-pretend}$ = 986 ms, $\Delta_{pretend}$ = 667 ms; Fig. 3A). Furthermore, decision uncertainty, quantified as the Shannon entropy of the posterior distribution over cell or letter options, had a similar quadratic effect on decision time in both non-pretend and pretend versions of the games, with the slowest responses associated with mid-range levels of entropy (see Fig. 3B). In other words, despite knowing the game's solution with full certainty, pretenders successfully feigned subtle qualitative effects of counterfactual uncertainty on their decision times.

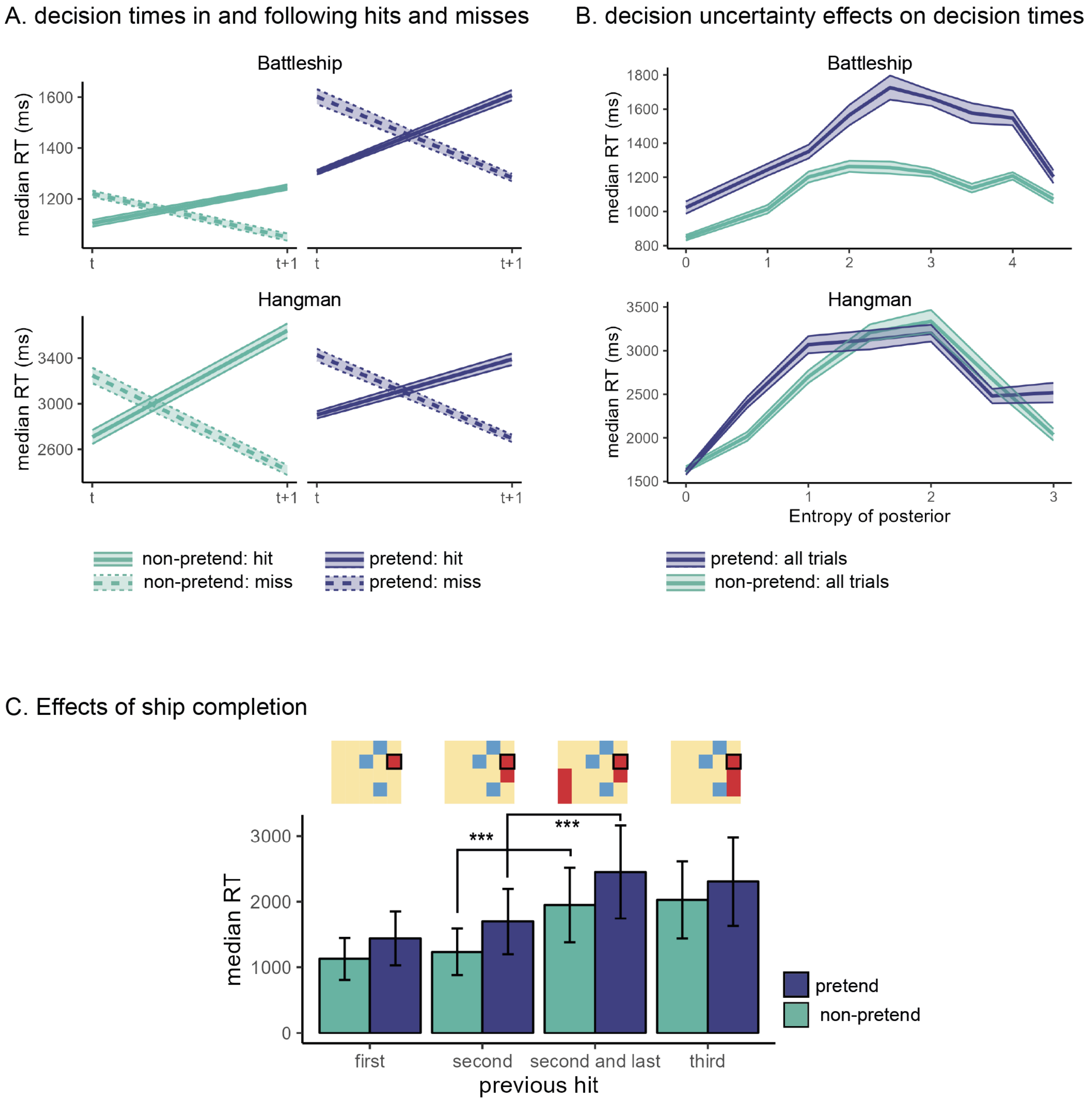

***Figure 3**: Patterns of decision time in pretend and non-pretend games*. A: median decision times for hits and misses, as well as the decisions following them. In both Battleship and Hangman, hits were faster on average than misses, but guesses following a hit were slower on average than those following a miss. This pattern was mimicked in pretend games. B: median decision times as a function of decision uncertainty, quantified as the entropy of the posterior over guess options. In both Hangman and Battleship, guesses were slowest for mid-range levels of entropy, and this pattern was mimicked in pretend games. C: Decision latency after a Battleship hit, as a function of belief state. Error bars and shaded areas represent the bootstrapped standard error of the median.

The previous analyses reveal robust associations between game state and decision latency, that are highly similar in pretend and non-pretend games. To get at subtler dynamics in a more direct way, in a follow-up exploratory analysis we focused on decision latencies following a Battleship hit. Battleship players attempted to reveal two size-2 patrol boats and one size-3 submarine. We categorized hits into four categories, based on players' knowledge at the time of guessing: first hit on a ship, second hit on a ship when the size-three submarine hasn't been sunk yet, second hit on a ship when the size-three submarine has already been sunk, and third hit on a submarine (see Fig. 3C). In the first category, players know the ship must continue in one of the neighboring cells. In the second category, there is a chance the ship continues (if this ship turns out to be a submarine). In the third and fourth categories, it is clear that the ship is fully sunk.

In non-pretend games, players were significantly slower to select the next cell when they knew they had just completed a ship (categories 3 and 4) compared to when they just hit a ship, but were not sure (second category) or knew they had not completely sunk it (first category). Specifically, we find that players were slower by 728 ms to make the next cell selection after hitting the second cell of a ship if the size-three submarine had already been sunk, compared to still hidden ($p < .001$).

Strikingly, we found the exact same pattern in pretend games. Players were faster to make the next cell selection when they pretended to think that the current ship might not be fully sunk. This was not merely a difference between the first, second and third hits: players were slower by 754 ms to make the next cell selection after hitting the second cell of a ship if the size-three submarine had already been sunk ($p < .001$). Further analysis confirmed that this effect remained significant when controlling for click number ($p < .001$), when restricting the analysis to the second hit of a ship that is in fact of size-two ($p < .001$), and when controlling for the physical distance traversed by the mouse from the previous click ($p < .001$).

This last finding bears emphasizing: In both categories two and three, pretenders *knew* that they had just sunk a size-two ship, but in the second case they *pretended not to know* this fact, and this affected their decision latency in the same way it would have been affected had they been in a non-pretend game.

Stereotypical, imperfect self-simulation

Though impressive, the capacity for simulating a state of ignorance was not perfect. Importantly, the limitations and biases we observe are consistent with the simulation of a stereotypical, “cartoon” model of decision-making, rather than leakage of concealed information into the decision-making process. First, despite showing the same qualitative effects, decision time patterns in Battleship pretend games (but not Hangman pretend games) were systematically more pronounced relative to non-pretend games: a form of “over-acting”. Furthermore, pretend games followed stereotypical patterns, and as a result were more homogeneous than non-pretend games. Despite a highly similar average number of misses in pretend and non-pretend games (Fig. 2A), the number of unsuccessful guesses was overwhelmingly less variable in pretend relative to non-pretend games (Battleship: sd=1.61 in pretend versus 2.60 in non-pretend games, $p < .001$; Hangman: sd=1.53 in pretend versus 2.65 in non-pretend games, $p < .001$; Fig. 4A).

Next, we examined variability not in the number of guesses, but in their contents. We separately computed the Shannon entropy of the guess distribution across different games for each player, condition (pretend or non-pretend), and serial guess number. High entropy then corresponds to pronounced variability in the guess sequences of different games. Unsurprisingly, the within-participant sequential guess entropy increased as a function of guess number, consistent with players adjusting their behaviour in light of the outcomes of previous guesses, making individual games increasingly more varied (Fig. 4B). Critically, however , entropy was systematically reduced in pretend games, in line with an attempt to enact typical, or average, behavior in a state of ignorance.

Finally, Hangman pretenders were more likely to guess letters that appear frequently in English words (E, T, A, etc.) irrespective of the game state, compared to genuine players (Fig. 4C). This suggests that in their attempt to behave as if they didn’t know the true state of the game, pretenders had an increased tendency to follow rigid heuristics and rules, ignoring useful information as a result.

This limitation on incorporating evidence into the (simulated) decision-making process was especially evident in Hangman half-games, where players completed the game from a half-completed state. When asked to reveal the hidden fruit “_A_A_A”, 90% of the non-pretenders guessed one of the letters ‘B’ or ‘N’ (Fig. 4D, yellow bars in left column). Among pretenders who knew that the hidden word was BANANA, this preference was reduced to 78%. Importantly, half of the pretenders were given different information: they were told that the hidden word was the less prototypical fruit PAPAYA. Although good pretenders should simulate their behavior had they not known this information, only 29% selected the letters ‘B’ or ‘N’, revealing that many pretenders were unable to predict that the first fruit that would have come to their minds was BANANA, not PAPAYA (Fig. 4D, yellow bars in right column). A similar pattern was observed for the prototypical body part word HA(ND) and its surprising counterpart HA(IR): when playing normally, 75% of the players selected letters that are consistent with the prototypical option HAND. This figure was 79% among pretenders for whom the target word was HAND, in contrast to only 39% among pretenders for whom the target word was HAIR (Fig. 4D, blue bars).

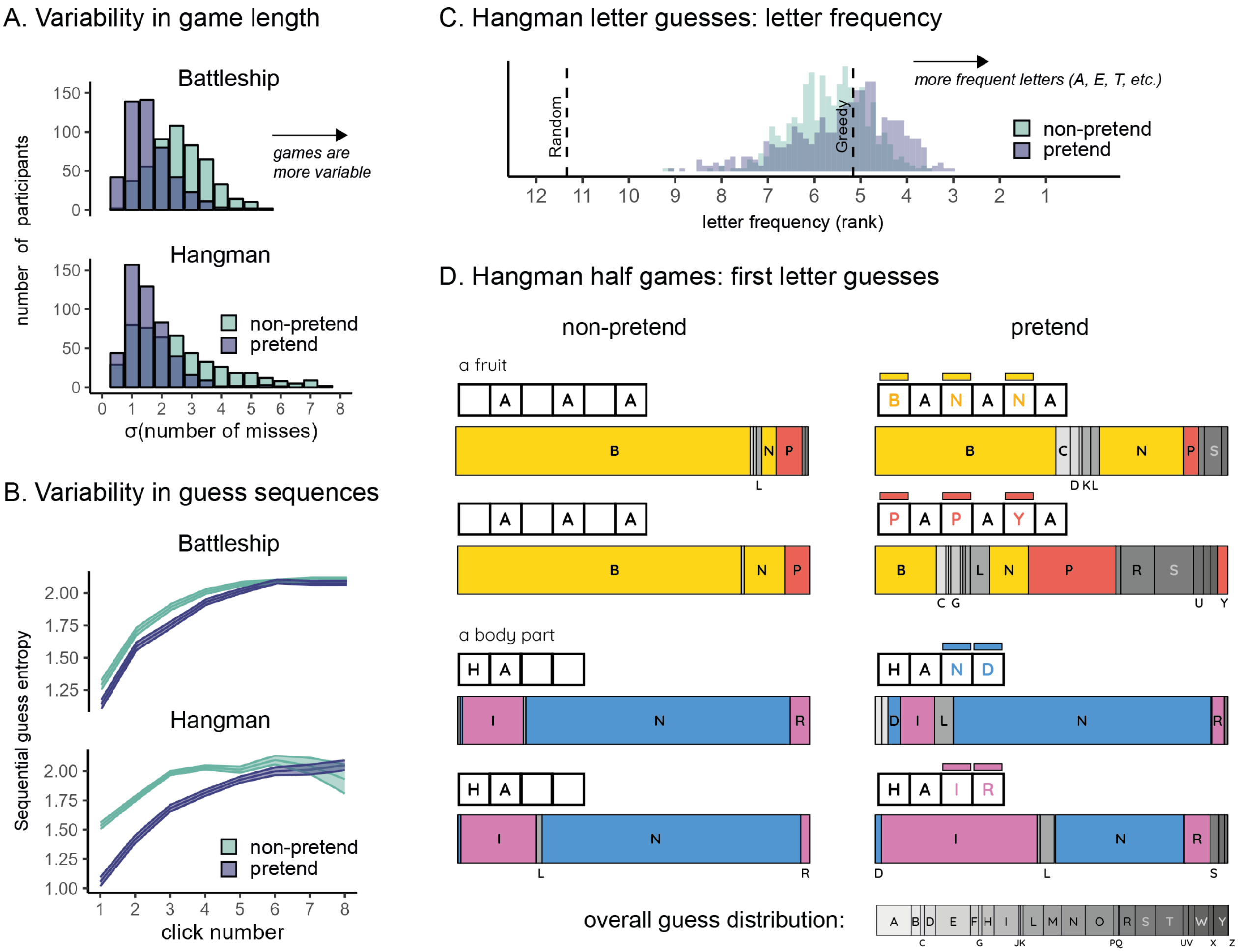

Figure 4: Limitations on flexible decision-making when pretending. A: variability in the number of misses (extracted individually for each player and then averaged) was lower in pretend games. B: sequential guess entropy, a measure of the (inverse) predictability of individual players' guesses as a function of click number and guess number. In both Battleship and Hangman, sequential guess entropy increased with click number, and was overall lower in pretend games. Shaded areas represent the mean +- one standard error. C: letter frequency of Hangman guesses: the median rank frequency per participant in pretend and non-pretend games, with reference lines for the expected rank frequency for a random agent, and for a "greedy" agent that maximizes the probability of a hit in each step. D: first letter guesses in Hangman half-games, as a function of pretense condition and target word. Letters appear in alphabetical order. letters that appear in the target word are marked in yellow (BANANA), red (PAPAYA), blue (HAND) and magenta (HAIR). For reference, the overall distribution of letter guesses across all games and conditions is given below.

Failure to detect pretending

Despite these limitations, pretenders' behavior resembled real players' behavior closely enough that they were able to fool other participants into believing they were playing for real. After performing pretend and non-pretend games, participants were presented with game replays of previous players and took on the new role of being a 'judge' who had to determine who was pretending and who was playing for real. In both games, pretense detection was at chance (Battleship: 51%, Hangman: 51%). This is in line with previous findings of near-chance accuracy in lie detection (Bond & DePaulo, 2006). Moreover, we find no sign of a correlation between pretense quality (measured as players' ability to trick judges into thinking they were not pretending) and pretense detection ability (measured as proportion correct; Battleship: $r_s = -.05$, Hangman: $r_s = .00$), indicating that pretense and pretense detection rely on at least partly different cognitive processes.

Discussion

In two experiments, we examined participants' ability to mimic a state of ignorance in a game setting, building on the recent recognition of games as a powerful tool for studying decision making (Allen et al., 2024). We find that pretenders were able to successfully emulate decisions taken under a true state of ignorance, including reproducing both broad patterns and subtle effects of guess accuracy and decision uncertainty on decision time. We also identify reliable signatures of pretend-ignorance on players' decisions, including a cost to decision rationality and an increased tendency to follow heuristics and rules, even though these signatures went undetected by 'judges' asked to discriminate real from pretend games. Collectively, our findings reveal a rich and impressive, but ultimately limited, capacity for online simulation of counterfactual belief states.

Previous research has identified limitations in our capacity to prevent knowledge from influencing our decisions and behavior (Fischhoff, 1975, 1977; Harley et al., 2004; Roese & Vohs, 2012; Wood, 1978). Our findings reveal that notwithstanding these limitations, humans are capable of approximating their hypothetical behavior had they not known what they in fact do know. This capacity goes beyond making similar decisions to the ones they would have made had they not known; pretenders were also able to generate decision times that reproduce subtle qualitative patterns observed under a true state of ignorance.

This alignment between pretend and non-pretend decision times may indicate that pretenders were capable of controlling their decision times at will, enacting their intuitive beliefs of how decision time should be affected by knowledge and uncertainty. In a recent developmental study, children aged 5-10 were capable of making inferences about knowledge from the observed response time of other agents (e.g., inferring that decision-makers who take longer to decide make higher-quality decisions Richardson and Keil (2022)). Our findings here may show that people are also capable of using this sort of intuitive knowledge to guide their own pretense behavior. Alternatively, this mimicry of decision time patterns may reflect a similarity in the computations leading to pretend and non-pretend decision making. According to this interpretation, pretenders may have paused for longer when making some decisions because the simulated computation that led to these decisions took longer.

Research on Bayesian Theory of Mind provides some support for such internal simulations of decision-making processes. These are often studied by measuring participants' ability to infer beliefs and desires from observed behavior, either explicitly (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Baker, Saxe, & Tenenbaum, 2009; Richardson & Keil, 2022), or implicitly (Liu, Ullman, Tenenbaum, & Spelke, 2017; Onishi & Baillargeon, 2005). Here we have proposed a complementary approach: Asking participants to generate behavior based on a counterfactual mental state—In this case, a counterfactual knowledge state in which a known piece of information is unknown. Instead of relying on model inversion (e.g., "Which belief states would give rise to this behavior?"), we ask participants to run the model forward, taking counterfactual beliefs and desires as input and producing behavior as output. Due to the unconstrained space of possible behaviors in our task (cell selections x decision latencies), successfully pretending not to know demands a rich model of cognition, and is much harder to achieve based on a quasi-scientific theory of mental states (Gopnik & Wellman, 1994).

An alternative interpretation of our results is that instead of simulating a counterfactual knowledge state, participants actively suppressed or ignored the revealed game state such that their entire cognitive machinery was available to play the game. This would not require self-simulation, only a capacity to intentionally ‘unsee’, or forget, relevant evidence. While we cannot fully rule out this interpretation, we think it is unlikely to explain our players’ successful pretense, for at least five reasons. First, we tried to make such suppression as hard as possible, by presenting the game solution on top of the game board for the entire duration of pretend games, and by having participants type the target word before pretend Hangman games. Second, suppressing thoughts on demand is notoriously difficult, and often has an opposite, positive effect on the suppressed content (Wegner, Schneider, Carter, & White, 1987). Third, when asked how they had performed the task in a debrief question, the responses of a significant majority of participants were aligned with self-simulation or rule-following (see exploratory analysis). Fourth, response time patterns were exaggerated in pretend Battleship relative to non-pretend Battleship games (see Fig. 3A): a finding that is consistent with the simulation of a simplified self-model, but is hard to explain if participants were merely suppressing their current knowledge. Finally, pretend games were more similar to each other than were non-pretend games to themselves. This is again consistent with an attempt to simulate typical behavior.

Together, our findings reveal a non-trivial capacity for pretending not to know. Complementing previous work on cognitive and perceptual hindsight biases, we show that people are capable of accurately simulating diverse aspects of their decision-making processes, although they exhibit systematic shortcomings. Further research into these and similar limitations may continue to reveal the simplifications, abstractions, and biases in people’s models of their own minds.

Methods

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

Exp. 1: Battleship

A detailed pre-registration can be accessed at osf.io/v9zsb. The pre-registration was time-locked using cryptographic randomization-based time-locking (Mazor, Mazor, & Mukamel, 2019) (protocol sum:

60c270410375e8a192468fc1a0e9c93da60d5e203eb2760b621a8631a26f4c5c; link to relevant lines in experimental code). All pre-registered analyses are available in this link.

Participants. The research complied with all relevant ethical regulations and was approved by the Institutional Review Board of Johns Hopkins University. 500 participants were recruited via Prolific (prolific.co) and gave their informed consent prior to their participation. They were selected based on their acceptance rate (>95%) and for being native English speakers. The entire experiment took approximately 20 minutes to complete. Participants' pay was equivalent to an hourly wage of 9.50 USD, in addition to a bonus payment (0.20 - 2 USD, mean = 0.90).

Procedure. Participants were first instructed that the experiment, based on the game Battleship, had three parts, and that they could accumulate 'points' that would later translate to a monetary bonus payment. They were then presented with a leaderboard of previous players, and given the rules of the game:

> *"In the game Battleship, your task is to sink all ships located in a grid with as few clicks as possible. What makes the game difficult is that you can't see the ships; all you can see is a grid of squares, and you have to guess where the ships are. To sink a ship, you need to click on all of the squares it is located in. If you hit part of a ship, the square will turn red. If there is no ship in the square, it will turn blue."*

We further explained that in this version of the game, ships can touch corners, but their sides can't touch. This explanation was accompanied by a visual presentation of legal and illegal ship configurations.

After completing a comprehension question and a practice round, participants completed one 'pretend' and one 'non-pretend' block, each comprising five full games and one half-game (see below for details). The order of pretend and non-pretend blocks was counterbalanced between participants. The allocation of boards to conditions was randomized between participants such that exactly one board was played in both pretend and non-pretend conditions, and this common board was different for different participants. The order of boards within a block was fully randomized, with the exception that half-games were always played last.

Non-pretend (normal) games. In non-pretend games (Fig. 1A), participants aimed to sink two 2-square patrol boats and one 3-square submarine with as few clicks as possible. An online counter of the number of clicks was displayed on the screen. After each game, feedback was given about the number of clicks and resulting number of points obtained.

Pretend games. Participants in pretend games were given the same explanation of Battleship, and played a practice round. However, they were then given an additional instruction:

> *"This time your goal is different. In this round, we're going to tell you where the ships are, but* ***we want you to act like you don't know this information****. We've marked the ships' locations with a cross, so you'll know where they are the whole time; but your job is to play the game as if these hints aren't there. To see how good you are at this, we're going to compare your games to the games of people who actually had no hints, and see how similar they are. We will measure where and when you clicked; if your clicks look similar to people who played like normal (trying to reveal all ships with as few clicks as possible, but without any hints), you'll get bonus points. But if your games look different, you won't get these bonus points. Your number of clicks in this part will not affect your bonus. Only your ability to play like you had no hints."*

We informed participants that both the location and timing of their cell clicks would be measured. After one practice round and one comprehension question, participants played five pretend games (Fig. 1B), followed by one pretend half-game. Each game was followed by a short message, reminding them that a game that looks similar to the games of participants who had no hints would be awarded 10 bonus points.

Half-games. In order to directly compare participants' pretend and non-pretend games for identical belief states (genuine or pretended ignorance about where the ships are hidden), participants completed one pretend and one non-pretend game given a partly finished board with the content of 7 cells already revealed (Fig. 1C). We designed our half-games to produce a strong expectation to find a ship in specific cells, but not in others. The assignment of half-completed boards to pretend and non-pretend conditions was randomized between participants.

Judge trials.

In the final part of the experiment, participants observed the games of previous players and tried to determine who were pretenders and who were non-pretenders. On each trial, two empty grids were presented side by side, with a smaller grid on top, displaying the hidden positions of ships on the grid (Fig. 1D). The two grids corresponded to the true games of two previous players who played a version of the top grid either as pretenders or as non-pretenders. For non-pretend games, only games from the group of participants that pretended in the second block (and played normally in the first block) were chosen for presentation in this part. For both pretend and non-pretend games, only games shorter than one minute (97% of included non-pretend games and 91% of pretend games) were presented. Judge participants observed a real time replay of the two grids, showing not only where participants clicked, but also when. After making a decision, participants were informed whether they would receive the 10 bonus points, or alternatively, whether the pretender would receive them in the event the pretender managed to trick them.

Readers are invited to try a demo of the experiment.

Hangman

A detailed pre-registration can be accessed at osf.io/3thry. The pre-registration was time-locked using cryptographic randomization-based time-locking (Mazor et al., 2019). Due to an error in the experiment code, time-locking took effect only from player number 221 (batch 3) and on (protocol sum: c4929c7fe33df1b7b52f15c789d98eab30a9cee09a8121807a3c59e28e7430a4; relevant lines in experimental code).

Participants. The research complied with all relevant ethical regulations and was approved by the Institutional Review Board of Johns Hopkins University. 501 Participants were recruited via Prolific (prolific.co) and gave their informed consent prior to their participation. They were selected based on their acceptance rate (>95%) and for being native English speakers. The entire experiment took approximately 20 minutes to complete. Participants' pay was equivalent to an hourly wage of 9.50 USD, in addition to a bonus payment of 1 USD that was awarded to 236 players who earned 100 points or more.

Procedure. The first instructions screen informed participants that the experiment, based on the game Hangman, had three parts, and that the points they accumulate translate to a monetary bonus payment. They were then presented with a leaderboard of previous players. Then, the rules of the game were presented:

"In the following game, your task is to reveal a hidden word or phrase by guessing letters. What makes the game difficult is that you can't see the word; all you can see is a row of squares - a square for each letter. Use your mouse to make letter guesses. We will have five types of words: body parts, numbers, US states, fruit, and famous people. You will start each game with 15 points and lose one point for every guess of a letter that is not in the word."

We then explained that "the words in this game are the kind of words that will be familiar to most English-speaking fifth-graders. We didn't pick any strange or particularly difficult words".

Once they responded correctly to a multiple-choice comprehension question ("the goal of the game is to…": "reveal the word with as few letter guesses as possible"), participants played a practice round, revealing the word PIGEON (see Fig. 1E).

After the main instructions, comprehension question and practice round, participants completed one pretend and one non-pretend block, each followed by one half-game (see below for details). The order of pretend and non-pretend blocks was counterbalanced between participants. Each block comprised five games played with five out of ten different words, and one half-game. The allocation of words to conditions was randomized between participants, with the constraint that both pretend and non-pretend blocks included exactly one word from each category. The order of words within a block was randomized, except for the half-game, which was always delivered at the end.

The ten words included two number words (ELEVEN, NINETY SIX), two famous people (DALAI LAMA, TAYLOR SWIFT), two fruits (STRAWBERRY, LEMON), two body parts (TOOTH, HEAD), and two US states (MONTANA, IOWA).

Non-pretend games. In non-pretend games, participants revealed a hidden word with as few letter guesses as possible. An online counter of the number of points was displayed on the screen, deducting one point for every guess of a letter that is not in the target word. After each game, feedback was given about the number of points obtained.

After completing the five games, participants performed one half-game (see below for details).

Pretend games. Participants were given the following instructions:

"In the next part of the experiment, you'll play 6 games where you reveal a hidden word by guessing letters.

However, this time your goal is different.

In this round, we're going to tell you the word in advance, but **we want you to act like you don't know this information**.

To see how good you are at this, we're going to compare your games to the games of people who played normally, without knowing what the word was, and see how similar they are. We will measure which letters you click and the timing of your guesses; if your clicks look similar to people who played like normal (trying to reveal the word with as few guesses as possible, but without any hints), you'll get bonus points. But if your games look different, you won't get these bonus points. Your number of clicks in this part will not affect your bonus. Only your ability to play like you didn't see the word in advance."

After one practice round, pretending not to know that the hidden word is PIGEON, and one comprehension question ("In this part of the experiment my goal is to…": "play the game as if I don't know what the word is so that I look like someone who had no hints"), participants played five pretend games (Fig. 1F). Each game was preceded by a short message informing subjects about the identity of the target word. To start pretending, players were asked to type in the target word on their keyboard. The target word remained on the screen, in green letters, until the end of the game. After pretending, we reminded players that a game that looks similar to the games of participants who had no hints will be awarded 10 bonus points.

After completing the five games, participants performed one half-game (see below for details).

Half-games. In order to directly compare participants' pretend and non-pretend games for identical belief states (true or pretended knowledge about the identity of the word), we asked participants to also complete one pretend and one non-pretend game, given a partly finished game with some letters already guessed (they were told that the computer made these guesses; Fig. 1G). The two half-game words were one fruit: PAPAYA or BANANA, with guessed letters [A, E, I, O, M, T], and one body part: HAND, or HAIR with guessed letters (A, E, O, M, T, H, P). The assignment of category (fruit or body part) to condition (pretend and non-pretend), as well as the identity of the target word within each category (e.g., PAPAYA or BANANA), was randomized between participants.

Instructions for the non-pretend half-game were:

"For the next game, the computer chose the first letters for you; you can take over from where it left off. Your challenge is to complete the game. Just like in the previous games, here also you will lose one point for each letter that you guess and is not in the word."

Instructions for the pretend half-game were:

"For the next game, the computer chose the first letters for you; you can take over from where it left off. Just like in the previous games, here also you will know what the word is, but your bonus points will depend on your ability to play as if you didn't know the word."

Judge trials. In the final part of the experiment, participants observed five games of previous players and determined who had hints and who didn't. Instructions for this part were:

"In this third and last part of the experiment, we ask you to be a judge for previous players, and see if you can tell which of the players were shown the word (but acted like they weren't). We will show you 5 replays of the games of previous players. Your task is to decide whether they played normally or pretended. For each game that you get right, you will receive 10 points. Good luck!"

Then, on each judge trial, one game of a previous player was replayed in real time, with the target word presented above. For non-pretend games, only games from the group of participants that pretended in the second block (and played normally in the first block) were chosen for presentation in this part. For both pretend and non-pretend games, only games shorter than 1.5 minutes (87% of included non-pretend games and 96% of pretend games) were presented. Judge participants indicated their decision by pressing the P and N keys on their keyboard. After making a decision, participants were informed whether they received the 10 points. Whenever a pretend game was classified as a non-pretend game, they were informed that the pretender received these 10 points instead of them.

Lastly, participants were asked the following debrief questions:

"Did you have a strategy that you used for pretending you did not see the word? What was most difficult about pretending? How about telling between players who pretenders and who played for real - did you have a strategy for that?"

And:

"We would appreciate it if you could share any thoughts you had about the experiment, or anything we should take into account when analyzing your data."

Readers are invited to try a demo of the experiment.

References

Allen, K. R., Brändle, F., Botvinick, M. M., Fan, J., Gershman, S. J., Gopnik, A., … Schulz, E. (2024). *Using games to understand the mind*. https://doi.org/10.31234/osf.io/hbsvj

Audinot, M., Bonnet, F., & Viennot, S. (2014). Optimal strategies against a random opponent in battleship. *The 19th Game Programming Workshop*.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 1–10.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Bernstein, D. M., & Harley, E. M. (2007). Fluency misattribution and visual hindsight bias. *Memory*, *15*(5), 548–560. https://doi.org/10.1080/09658210701390701

Bernstein, D. M., Wilson, A. M., Pernat, N. L. M., & Meilleur, L. R. (2012). Auditory hindsight bias. *Psychonomic Bulletin & Review*, *19*(4), 588–593. https://doi.org/10.3758/s13423-012-0268-0

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of Deception Judgments. *Personality and Social Psychology Review*, *10*(3), 214–234. https://doi.org/10.1207/s15327957pspr1003_2

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(3), 288–299. https://doi.org/10.1037/0096-1523.1.3.288

Fischhoff, B. (1977). Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(2), 349–358. https://doi.org/10.1037/0096-1523.3.2.349

Gopnik, A., & Wellman, H. M. (1994). The theory theory. *An Earlier Version of This Chapter Was Presented at the Society for Research in Child Development Meeting, 1991*. Cambridge University Press.

Harley, E. M., Carlsen, K. A., & Loftus, G. R. (2004). The "Saw-It-All-Along" Effect: Demonstrations of Visual Hindsight Bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(5), 960–968. https://doi.org/10.1037/0278-7393.30.5.960

Levin, D. T., & Angelone, B. L. (2008). The visual metacognition questionnaire: A measure of intuitions about vision. *The American Journal of Psychology*, *121*(3), 451–472. https://doi.org/10.2307/20445476

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

Mazor, M., Mazor, N., & Mukamel, R. (2019). A novel tool for time-locking study plans to results. *European Journal of Neuroscience*, *49*(9), 1149–1156. https://doi.org/10.1111/ejn.14278

Mazor, M., Siegel, M. H., & Tenenbaum, J. B. (2023). Prospective search time estimates reveal the strengths and limits of internal models of visual search. *Journal of Experimental Psychology. General*. https://doi.org/10.1037/xge0001360

Morris, A., Carlson, R. W., Kober, H., & Crockett, M. (2023). *Introspective access to value-based choice processes*. https://doi.org/10.31234/osf.io/2zrfa

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–258.

Pohl, R. F., & Hell, W. (1996). No reduction in hindsight bias after complete information and repeated testing. *Organizational Behavior and Human Decision Processes*, *67*(1), 49–58. https://doi.org/10.1006/obhd.1996.0064

Richardson, E., & Keil, F. C. (2022). Thinking takes time: Children use agents' response times to infer the source, quality, and complexity of their knowledge. *Cognition*, *224*, 105073. https://doi.org/10.1016/j.cognition.2022.105073

Roese, N. J., & Vohs, K. D. (2012). Hindsight Bias. *Perspectives on Psychological Science*, *7*(5), 411–426. https://doi.org/10.1177/1745691612454303

Wegner, D. M., Schneider, D. J., Carter, S. R., & White, T. L. (1987). Paradoxical effects of thought suppression. *Journal of Personality and Social Psychology*, *53*(1), 5.

Wood, G. (1978). The knew-it-all-along effect. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(2), 345–353. https://doi.org/10.1037/0096-1523.4.2.345