

Pretending not to know reveals a capacity for model-based self-simulation

Matan Mazor<sup>1</sup>, Chaz Firestone<sup>2</sup>, & Ian Phillips<sup>2</sup>

<sup>1</sup> University of Oxford

<sup>2</sup> 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](mailto:matan.mazor@all-souls.ox.ac.uk)

## Abstract

Pretending not to know requires appreciating how one would behave without a given piece of knowledge, and acting accordingly. Here, two game-based experiments reveal a capacity to simulate decision-making under such counterfactual ignorance. 1001 English-speaking adults saw the solution to a game (ship locations in Battleship, the hidden word in Hangman) but attempted to play as though they never had this information. Pretenders accurately mimicked broad aspects of genuine play, including the number of guesses required to reach a solution, as well as subtle patterns, such as effects of decision uncertainty on decision time. While peers were unable to detect pretense, statistical analysis and computational modeling uncovered traces of ‘over-acting’ in pretenders’ decisions, suggesting 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:* 7745

Pretending not to know reveals a capacity for model-based self-simulation

## Research Transparency Statement

### General Disclosures

The authors declare no conflicts of interest. Funding: This study was supported by an NSF BCS grant #2021053 awarded to C.F. M.M is supported by a post-doctoral research fellowship at All Souls College. Artificial intelligence: No artificial intelligence assisted technologies were used in this research or the creation of this article. Ethics: The research complied with all relevant ethical regulations and was approved by the Institutional Review Board of Johns Hopkins University.

### Experiment 1 (Battleship)

Preregistration: The hypotheses and methods were preregistered on 2021-12-21, prior to data collection, and following the analysis of an independent pilot sample. A detailed pre-registration can be accessed at <https://osf.io/v9zsb>. The pre-registration was time-locked using cryptographic randomisation-based time-locking (Mazor, Mazor, & Mukamel, 2019) (protocol sum: 60c270410375e8a192468fc1a0e9c93da60d5e203eb2760b621a8631a26f4c5c; [link to relevant lines in experimental code](#), making experimental randomisation causally dependent on the content of the pre-registration and thus ensuring that all data were collected after pre-registration. All pre-registered analyses are publicly available, including the report-generating R script (<https://self-model.github.io/pretendingNotToKnow/docs/exp.-1-battleship.html>). Exploratory analyses are flagged as such. Materials: All study materials, including demos of analysis experiments, are publicly available (<https://github.com/self-model/pretendingNotToKnow>). Data: All primary data are publicly available

(<https://osf.io/zma9b>). Analysis scripts: All analysis scripts are publicly available

(<https://osf.io/zma9b>).

## Experiment 2 (Hangman)

Preregistration: The hypotheses and methods were preregistered on 2022-06-22, prior to data collection, and following the analysis of an independent pilot sample. A detailed pre-registration can be accessed at [osf.io/3thry](https://osf.io/3thry). The pre-registration was time-locked using cryptographic randomisation-based time-locking, making experimental randomisation causally dependent on the content of the pre-registration and thus ensuring that all data were collected after pre-registration (Mazor, Mazor, & Mukamel, 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](#)). All pre-registered analyses are publicly available, including the report-generating R script (<https://self-model.github.io/pretendingNotToKnow/docs/exp.-2-hangman.html>). Exploratory analyses are flagged as such. Materials: All study materials, including demos of the experiments, are publicly available (<https://github.com/self-model/pretendingNotToKnow>). Data: All primary data are publicly available (<https://osf.io/zma9b>). Analysis scripts: All analysis scripts are publicly available (<https://osf.io/zma9b>).

## 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?"*). Similar examples abound in higher-stakes contexts such as diplomacy, warcraft and law. 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.

Our research question is whether people can mentally simulate their actions under a counterfactual knowledge state of ignorance. To that end, we had participants “pretend not to know” 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. Participants’ pretense behaviour mirrored broad patterns and subtle features of real players’ decisions and decision times. At the same time, epistemic pretense was characterized by over-acting, stereotypical behavior, and suboptimal incorporation of new information: all markers of model-based simulation. Together, we take these findings as evidence for a capacity to mentally simulate decisions and actions using a simplified and schematic self-model.

## **Method**

The research complied with all relevant ethical regulations and was approved by the Institutional Review Board of Johns Hopkins University. In two experiments, online participants played online versions of two information-seeking games. In Battleship (Exp. 1), 500 English-speaking players (recruited from Prolific.com) were presented with a 5x5 grid of yellow squares,

and attempted to reveal one size-3 submarine and two size-2 patrol boats with as few guesses as possible. In our version of the game, ships could only touch corner-to-corner, but not side-to-side (this was explained to participants before playing), and participants were not notified once they had sunk a ship (only whether their guess was a hit or a miss). In Hangman, N=501 English-speaking players attempted to reveal a hidden word, name or number (hereafter referred to broadly as “word”) with as few letter-guesses as possible, based on word length and category (a famous person, number, fruit, US state, or body part). To ensure familiarity with US states, Hangman participants were all US-based.

Both 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 and remained visible on the screen throughout the entire game (pretend-Battleship ship locations were marked with a cross, pretend-Hangman words were presented visually and had to be typed by players before the game, to ensure encoding; see Fig. 1). In these games, players’ task was to behave as if they were playing for real – i.e. to play as though they did not have this information.

Throughout the game, participants accrued points that were later converted to a monetary bonus. In non-pretend games, participants received points for revealing the ships, or the target word, with as few guesses as possible. In pretend games, participants were given different instructions:

“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."

#### And in Hangman:

"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."

We intentionally included no reference to an observer in these instructions, to have participants focusing on simulating their own behaviour rather than simulating how their behaviour would be perceived by another person. In reality, participants' games were presented to other participants, and they received bonus points if they tricked these other participants into believing they did not have hints.

Players played pretend and standard games in separate blocks that were presented in random order after a first 'practice' game. In principle, participants could learn about their own behaviour from this practice game. To minimize such learning effects, we distinguished practice games from the main experimental blocks, using a smaller 4x4 grid with only two size-2 ships in Battleship, and a word category (animals) that was not used in the main experiment in Hangman.

Each experimental block was followed by a half-game, where players were instructed to complete the game from a half-finished state. Finally, players were presented with replays of the games of previous players, and judged which were standard and which were pretend games. We measured players' 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. Unless otherwise specified, all reported findings similarly hold when analysing only the first condition performed by each subject in a between-subject analysis, thereby ensuring that findings are not driven by learning effects<sup>1</sup>. Readers are invited to try [demos of the experiments](#).

---

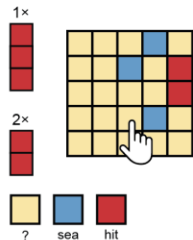
<sup>1</sup> In both experiments, the order of pretend and non-pretend blocks was counterbalanced between participants. We observed no significant interaction between the strength of any of our effects (i.e., differences between pretend and non-pretend conditions) and part of the experiment (i.e., first versus second). For full details see Supplementary Materials.

## Exp. 1: Battleship

randomized order

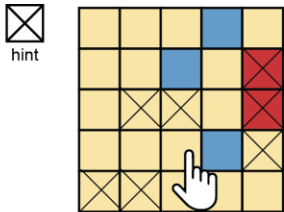
**A. non-pretend games** ×5

your task is to sink all ships located in a grid with as few clicks as possible.



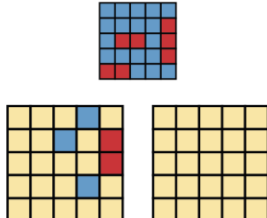
**B. pretend games** ×5

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.**



**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:

\_\_\_\_\_

hint A

an animal

P \_ I \_ 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

**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

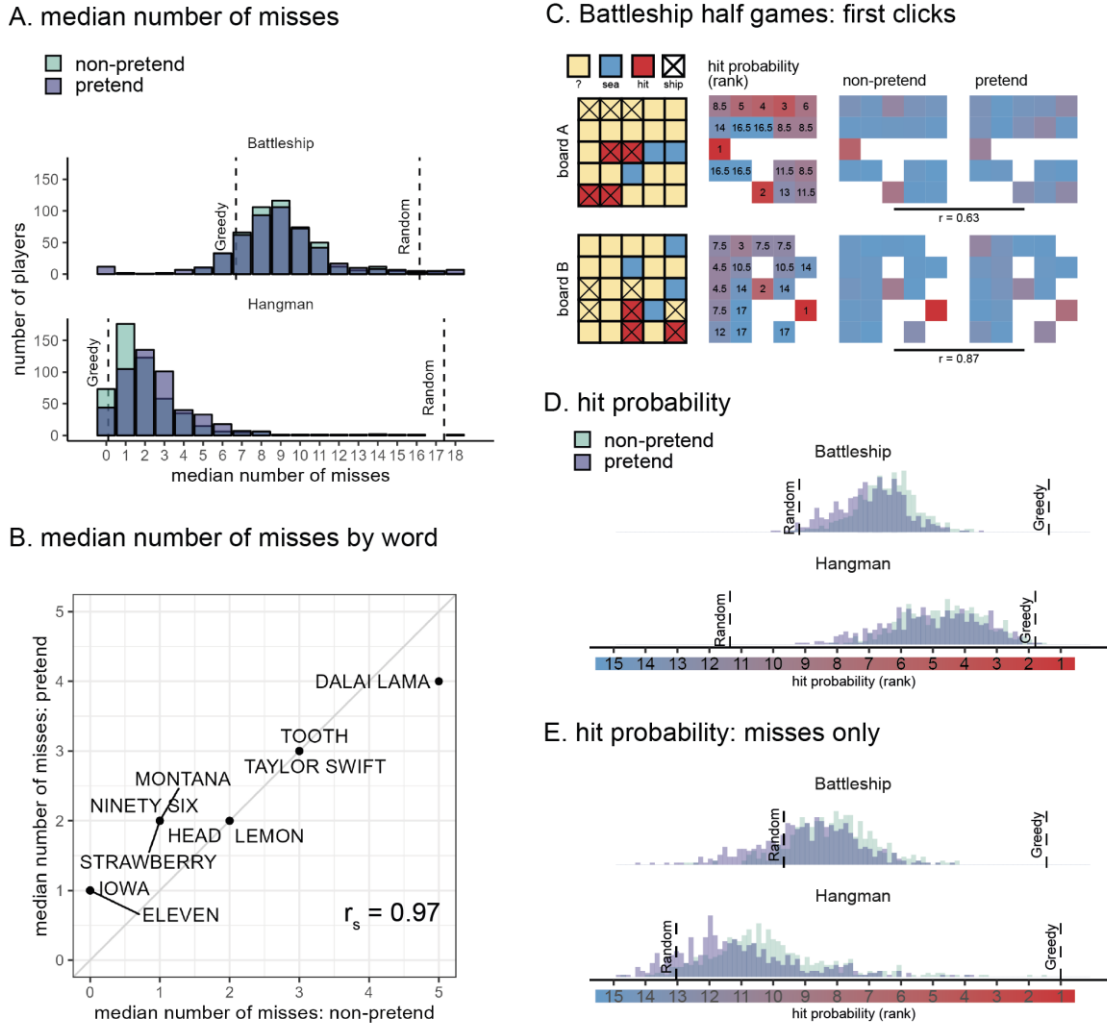
Replaying game: 00:00:43:87

*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).

## Results

### 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, SD = 2.91) and non-pretend games (mean = 16.05, SD = 2.18;  $t(499) = -1.43$ ,  $p = .153$ ; Cohen's  $d = 0.06$ ; 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;  $t(479) = 1.39$ ,  $p = .164$ ; 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, SD = 2.77; non-pretend: 1.94 misses, SD = 1.76;  $t(500) = 6.47$ ,  $p < .001$ ; Cohen's  $d = 0.29$ ; Fig. 2B). Despite an overall bias in the number of guesses, pretend Hangman games showed a near-perfect item-specific alignment: pretenders were successful in making more incorrect 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 given the players’ knowledge at the time of making the decision (dynamically updated after each guess). 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. Note that the expected rank for a greedy agent is greater than 1 because there was not always a single optimal choice. 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. In standard games, players start in the same (blank) board state but quickly diverge as they make different guess sequences. Including half games allowed us access to hundreds of cell selections for the same board state from pretenders and non-pretenders. This way, we had sufficient statistical power to compare the two guess distributions. We find 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, & Viennot, 2014). To this end, we ranked cells based on the Bayesian probability of a hit given players’ knowledge at the time of making the decision. Critically, hit probability maps were dynamically updated after each guess. In Battleship, this model assumed that all legal board configurations are equally likely a priori, but board configurations were ruled out as the game progressed and the content of individual cells was revealed. In Hangman, we used the category information (e.g., ‘a fruit’), to obtain a probability-weighted list of category-compatible words (or names, in the case of famous people). We relied on prototypicality norms (Uyeda & Mandler, 1980) for words, and number of visits to Wikipedia entries for famous people. The full prior distributions for each category were

included in the pre-registration (for details, see Supplementary Materials). Similar to Battleship, in deriving hit probability we assume access to the full list of options that is consistent with the game state (the number of hidden letters, the revealed letters and their positions, and the list of letters that do not appear in the game solution) at the time of making the decision.

In the non-pretend versions of both games, guesses were more rational according to this measure than expected by chance (Battleship:  $t(479) = 49.18, p < .001$ , Cohen's  $d = 2.24$ ; Hangman:  $t(500) = 86.88, p < .001$ , Cohen's  $d = 3.88$ ). Pretend guesses were also more rational than expected by chance (Battleship:  $t(479) = 38.51, p < .001$ , Cohen's  $d = 1.76$ ; Hangman:  $t(500) = 72.29, p < .001$ , Cohen's  $d = 3.23$ ), but significantly less rational than non-pretend guesses (Battleship:  $t(479) = 11.04, p < .001$ , Cohen's  $d = 0.50$ ; Hangman:  $t(500) = -4.57, p < .001$ , Cohen's  $d = 0.20$  Fig. 2D). Critically, pretend guesses were more rational than random guesses even when restricting the analysis to unsuccessful guesses (Battleship:  $t(479) = 10.25, p < .001$ , Cohen's  $d = 0.47$ ; Hangman:  $t(487) = 18.91, p < .001$ , Cohen's  $d = 0.86$ ; 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.

A specific example of this effect in the game of Battleship can be observed in players' behaviour immediately after hitting the last cell of a size-2 patrol boat (players attempted to reveal two size-2 patrol boats and one size-3 submarine). Among non-pretenders, the next cell selection was often directed at checking whether the two cells were part of the size-3 submarine, but this was only true if the size-3 submarine had not been sunk yet (52% of all cell selections), and not when it had been sunk (4% of all cell selections, and significantly lower than 52%:  $t(395) = 30.47, p < .001$ , Cohen's  $d = 1.53$ ). Despite knowing with full certainty that the size-

2 patrol boat was not a size-3 submarine, pretenders showed the same qualitative pattern: pretending to check if the revealed cells were part of a size-3 submarine only when they pretended not to know that it was fully sunk (22% of all cell selections), but not when the size-3 submarine had been sunk (4% of all cell selections,  $t(366) = 12.09$ ,  $p < .001$ , Cohen's  $d = 0.63$ ). The tendency to check if the two cells were part of a bigger ship was weaker among pretenders ( $t(467) = -18.07$ ,  $p < .001$ , Cohen's  $d = 0.84$ ).

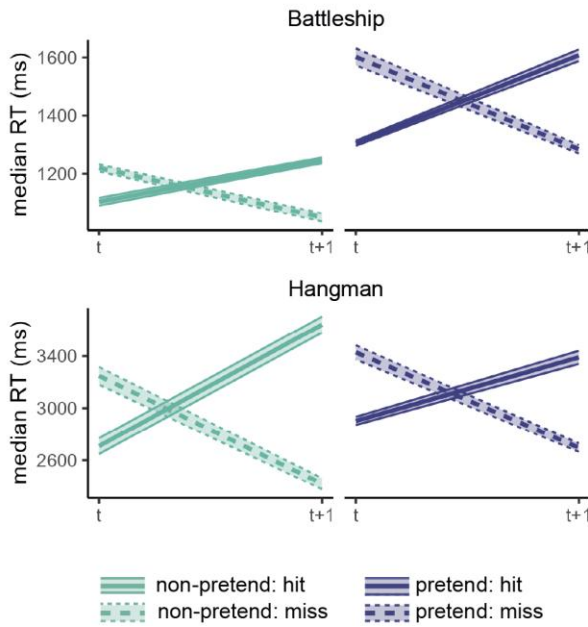
Good pretense is a function not only of the number and content of players' decisions, but also of their timing. Here too, pretend games showed the same qualitative patterns as non-pretend games. Like non-pretenders, pretenders were faster in their successful guesses (difference in decision time between 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). All effects are significant at the 0.001 level with the pre-registered within-subject t-test, except for the post-hit slowing down in Battleship, which, due to outliers with extreme effects in the opposite direction ( $>10$  s), was only significant in a non-parametric Wilcoxon sign-rank test ( $V = 87,876.50$ ,  $p < .001$ ). Effects remain significant at the 0.001 level when statistically controlling for the serial position of guesses within the game.

We also examined the effect of decision uncertainty, quantified as the Shannon entropy of the posterior distribution over cell or letter options, on decision times. To this end we fitted subject-level linear models, predicting reaction times from the linear and quadratic expansions of decision entropy, and contrasted the coefficients against zero in a group-level t-test. In the non-pretend versions of the games, the quadratic coefficients were significantly negative, with the

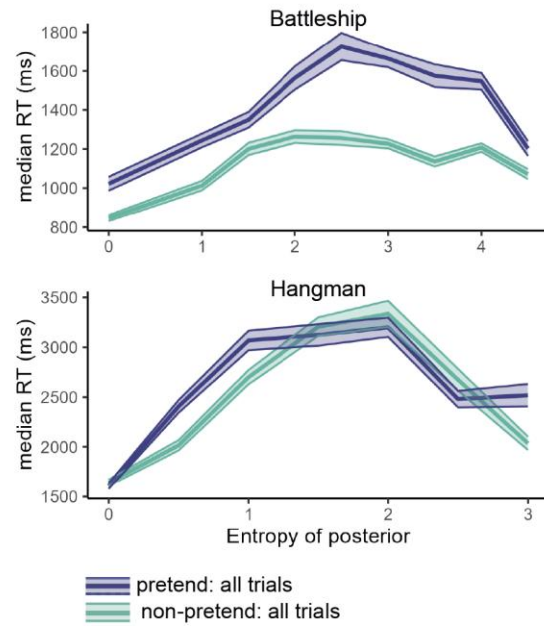
slowest responses associated with mid-range levels of entropy (see Fig. 3B; Battleship:  $t(479) = -4.20, p < .001$ , Cohen's  $d = 0.19$ ; Hangman:  $t(500) = -8.70, p < .001$ , Cohen's  $d = 0.39$ ).

When restricting the analysis to those Battleship players who pretended after playing normally, this effect was significant only in a Wilcoxon rank-sum test, due to outliers in the sample:  $V = 3,892.00, p < .001$ ). Critically, the quadratic coefficients were significantly negative also in pretend games (see Fig. 3B; Battleship:  $t(479) = -15.65, p < .001$ , Cohen's  $d = 0.71$ ; Hangman:  $t(500) = -3.49, p < .001$ , Cohen's  $d = 0.16$ ). 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.

A. decision times in and following hits and misses



B. decision uncertainty effects on 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. Shaded areas represent the bootstrapped standard error of the median.

### 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 as would be expected if pretenders’ success was due to efficient, but imperfect, suppression of their knowledge of the game solution. 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”. Specifically, the difference in reaction times as a function of guess outcome (Fig.3A) was larger in pretend games, both when measured with respect to the current guess ( $t(479) = 10.69, p < .001$ , Cohen’s  $d = 0.49$ ), and with respect to the following guess ( $t(479) = 2.69, p = .007$ , Cohen’s  $d = 0.12$ ). Similarly, the quadratic effect of decision entropy on decision times was stronger in pretend games ( $t(479) = 4.92, p < .001$ , Cohen’s  $d = 0.22$ ).

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,  $t(499) = -15.65, p < .001$ , Cohen’s  $d = 0.70$ ; Hangman:  $sd=1.53$  in pretend versus  $2.65$  in non-pretend games,  $t(500) = -12.65, p < .001$ , Cohen’s  $d = 0.56$ ; Fig. 4A). Moreover, while pretenders produced more letter misses for harder words (Fig. 2B), they underestimated the difficulty of the very hard “DALAI LAMA” and overestimated the difficulty of the easy number (“ELEVEN” and “NINETY SIX”) and state (“MONTANA” and “IOWA”) words. That is, pretenders consistently enacted what they saw as a ‘typical’ or a ‘representative’ game, one that is not unusual in the number of lucky or unlucky guesses. This is again consistent with shrinkage toward the mean of a generative self-model (Jansen, Rafferty, & Griffiths, 2021; Mazor & Fleming, 2021), with an attempt to avoid extreme outcomes to appear convincing to a hypothetical observer (Oey, Schachner, & Vul, 2023), and with representativeness skewing intuitions about randomness (Kahneman & Tversky, 1972).

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, and low entropy to a tendency to repeat the same sequence of guesses in different games. For example, if a player always starts their games by clicking in the top left corner, their guess entropy for the first click will be  $H([1,1,1,1,1])=0$ . 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). If pretend games were a similar but noisier version of standard games, their associated guess entropy would be higher, reflecting the additional noise in the decision-making process, or the game-specific biases that are associated with the suppression of specific words or game states. Critically, however, entropy was systematically reduced in pretend games ( $p<0.001$  for a within-subject t-test of guess entropy in guesses number 1-4 in both Battleship and Hangman, see Supplementary Materials for guess-specific statistics). Thus, when pretending, participants produced similar guess sequences across different games. In Hangman, for example, this meant that non-pretenders more flexibly adjusted their first letter guess to the word category and number of letters, compared to pretenders, who tended to open the game with the same letter guess irrespective of the specific game state. This seems consistent with an attempt to enact what they saw as typical, representative, or average behaviour. In contrast, a reduction in the guess sequence entropy is inconsistent with leakage of suppressed knowledge into the decision-making process, as would be expected if differences between pretend and non-pretend games reflected the imperfect suppression of the game's solution.

One possible account of the reduced decision entropy in pretend games is that it reflects pretenders' miscalibrated intuitions about randomness, conforming to a prototype of randomness that is itself too ordered. If the same prototype of randomness is used by pretenders to determine the number of unsuccessful guesses per game, the two measures should be correlated across

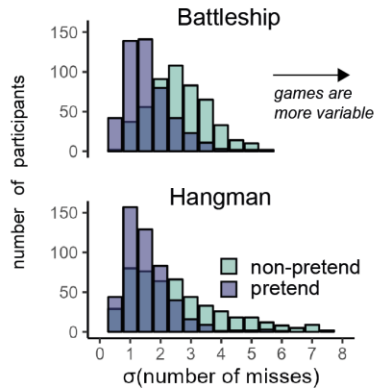
participants. Crucially, we find the exact opposite pattern: a negative correlation between variability in the number of unsuccessful guesses and game entropy (Battleship:  $r_s = -.20$ ,  $p < .001$ , Hangman:  $r_s = -.12$ ,  $p = .009$ ). This negative correlation was not observed in non-pretend games (and was even positive for Battleship:  $r_s = .11$ ,  $S = 16,454,287.34$ ,  $p = .019$ ; Hangman:  $r_s = -.01$ ,  $S = 21,158,163.09$ ,  $p = .832$ ). We interpret this as evidence that the reduction in variance reflects miscalibrated beliefs not only about randomness, but also about participants' behaviour under a counterfactual knowledge state. Those players who thought they would strictly follow a particular sequence of guesses (low entropy), ended up producing games of more variable lengths, as their success depended more on luck. Other players adjusted their decision strategy more flexibly, perhaps attempting to produce games that are not too long or short, in line with their intuitions about randomness.

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 (but see Supplementary Materials for evidence that heuristic use alone cannot fully explain pretenders' behaviour).

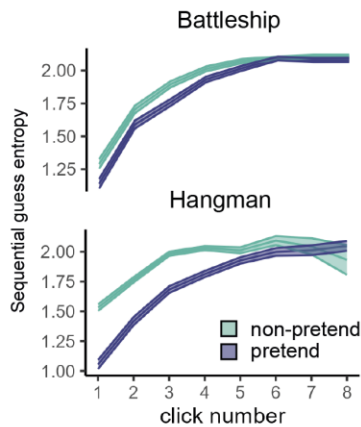
This limitation on incorporating evidence into the (simulated) decision-making process was especially evident in Hangman half-games, where players completed the game from a half-completed state. When asked to reveal the hidden fruit “\_A\_A\_A”, 90% of the non-pretenders guessed one of the letters ‘B’ 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% (this drop was significant in a binomial test:  $p = .002$ ). 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', with many guessing letters that are not consistent with either PAPAYA or BANANA (gray bars in Fig. 4D), revealing that many pretenders had the knowledge that PAPAYA would be a hard fruit to guess, therefore avoiding the letter 'P', but were still unable to predict that BANANA would have immediately come to their minds (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).

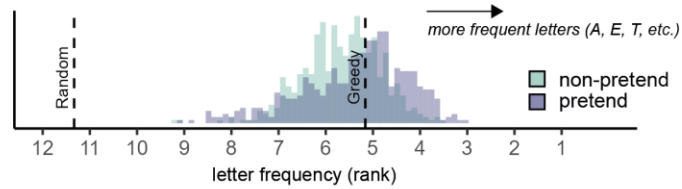
A. Variability in game length



B. Variability in guess sequences



C. Hangman letter guesses: letter frequency



D. Hangman half games: first letter guesses

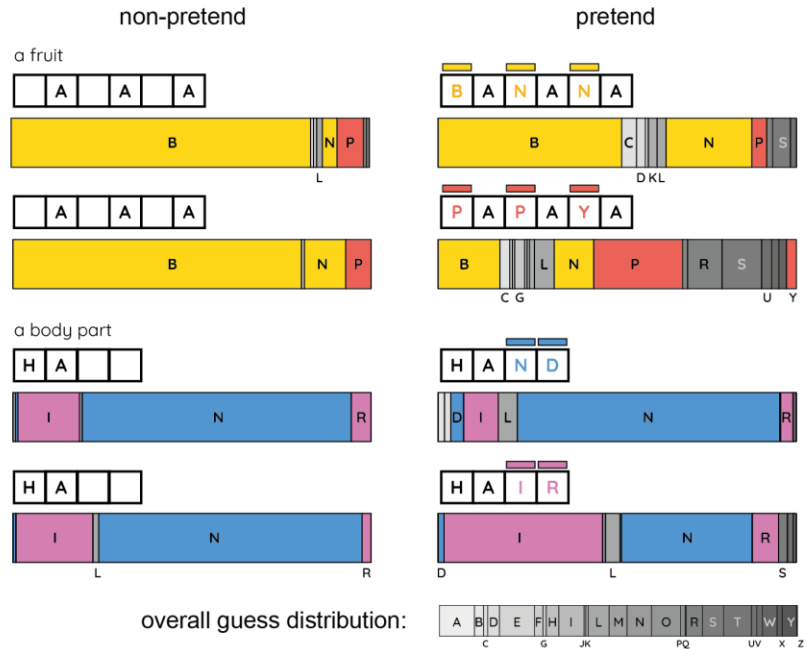


Figure 4. Limitations on flexible decision-making when pretending. A: variability in the number of misses (extracted individually for each player and then averaged) was lower in pretend games. B: sequential guess entropy, a measure of the (inverse) predictability of individual players' guesses as a function 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  $\pm$  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**

These systematic deviations of pretend games from behaviour under true ignorance made pretend games discernible from non-pretend games. To illustrate, a support vector machine (SVM) algorithm reached an accuracy level of 72% (Battleship, 2AFC) and 60% (Hangman, yes/no) in linearly classifying condition (pretend / non-pretend) based on just three basic summary features: median decision latency, game optimality score, and number of irrational guesses per game. Despite this, human observers were entirely fooled by players' pretending. 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 Battleship, judges were presented with five pairs of game replays from different players. Within each pair, one player was a pretender, and judges' task was to identify the pretender (or, for a random half of the judges, identify the real player). In Hangman, judges were presented with five game replays, each from a different player attempting to reveal a different word. For each individual game replay, the probability of it being a pretend game was 0.5, such that the number of pretend game replays was anywhere between 0 and 5 following a binomial distribution. Judges' task was to decide, for each game, whether it was played as a pretend or a non-pretend game. In both games, pretense detection was at chance (Battleship: 51%,  $t(499) = 1.45$ ,  $p = .147$ ; Hangman: 51%,  $t(452) = 0.57$ ,  $p = .568$ ). 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. By extracting the same statistical and model-derived measures from pretend and non-pretend behaviour, we were able to directly compare how people truly solve a puzzle with how they believe they would solve the puzzle had they not known the solution. This approach revealed that people are capable of 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 are most consistent with epistemic pretense involving model-based self-simulation, based on a simplified model of participants' own cognition

Previous research has identified limitations in our capacity to prevent knowledge from influencing our decisions and behavior (Fischhoff, 1975, 1977; Harley et al., 2004; Roesse & Vohs, 2012; Wood, 1978). In some cases, attempts to suppress thoughts even give rise to the paradoxical enhancement of suppressed representations (Earp, Dill, Harris, Ackerman, & Bargh, 2013; Giuliano & Wicha, 2010; Wegner, Schneider, Carter, & White, 1987). 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.

Internal simulations of decision-making processes are often studied (for example, in research on Bayesian Theory of Mind) 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). As such, our findings support a simulation model of epistemic pretense, and perhaps of mentalizing more generally. Critically, however, unlike classic “self-simulation” accounts of mindreading and theory of mind (Gallese & Goldman, 1998; R. M. Gordon, 1986; Perner, 1996), which, in their purest form, entail that simulating ignorance should require effectively deleting or hiding mental representations from one's self (R. Gordon, 2007), here the simulation is not of one's actual cognitive machinery, but of a simplified, “cartoon” model of it that depicts its most salient surface-level aspects while ignoring details (Graziano & Webb, 2015). A simulation of a schematic model explains both participants' ability to mimic subtle patterns of true ignorance in

an online fashion, as well as their consistent biases and limitations relative to behaviour when in a true state of ignorance (Saxe, 2005).

We interpret participants’ success in emulating a state of ignorance as revealing a non-trivial capacity for model-based counterfactual simulation, over and beyond any ability to suppress or ignore information (here, the game’s solution). This interpretation is supported by our finding, observed in both experiments, that pretend games were more similar to each other than were non-pretend games to each other, consistent with an attraction to the mean of a prior distribution (Mazor & Fleming, 2021), or with an attempt to simulate representative behavior (Kahneman & Tversky, 1972). Such a tendency to avoid extreme events has been observed in the way people lie to an opponent (Oey et al., 2023), and in the generation of pseudorandom sequences of coin flips (Bar-Hillel & Wagenaar, 1991; Falk & Konold, 1997; Nickerson, 2002). A similar effect is observed in Generative Adversarial Networks (GANs) where the distribution of generated samples is often narrower than the distribution of training data (an effect known as “mode collapse,” Kossale, Airaj, & Darouichi, 2022). This underestimation of variability in game length cannot be explained by suppression alone. Additional support for a model-based simulation interpretation comes from the exaggerated, over-acted response-time profiles in pretend Battleship.

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 players’ successful pretense, for at least three reasons over and above the tendency to produce representative behaviour described

above. First, we tried to make such suppression as hard as possible, by presenting the game solution 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 (Earp et al., 2013; Giuliano & Wicha, 2010; Wegner et al., 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, and our main findings hold when excluding the 32 Battleship and 10 Hangman players who mentioned suppression in response to this question (see exploratory analysis).

Findings from Battleship and Hangman mostly aligned: For both environments, the median number of guesses was similar in pretend and non-pretend games, guesses (correct and incorrect) made sense within the context of the game, and reaction times were similarly sensitive to guess accuracy and uncertainty. We also observed a similar tendency to produce representative and stereotypical behaviour in both experiments. At the same time, some differences are worth noting. First, fewer participants reported suppression as a strategy in pretend-Hangman games (2% of all pretenders) compared to pretend-Battleship games (6% of all pretenders;  $p < .001$  in a Chi-square test of independence). This may be related to the fact that only in Hangman were players required to type in the target word before pretending, making suppression much harder. A second notable difference is the failure of many participants to predict their behaviour in Hangman half-games — most notably, their inability to appreciate that a high frequency word (e.g., BANANA) would immediately come to mind — when knowing that the solution is a low-frequency word (e.g., PAPAYA). This failure may have to do with an important difference between the two games: in Battleship, success in the game depends on players' ability to weigh the relative likelihood of a relatively constrained set of hypotheses (grid configurations), which

are fully specified by the rules of the game. In Hangman, in contrast, even though the set of hypotheses may be tightly constrained, these hypotheses are not evident from the rules of the game themselves. As a result, success in Hangman depends also on specific hypotheses coming to mind: a process that is largely masked from awareness (Bear, Bensinger, Jara-Ettinger, Knobe, & Cushman, 2020). It is possible that, having conscious access to the process of deliberation between existing hypotheses but not to the process of generating new hypotheses, participants can successfully simulate the first but not the second. An additional, not mutually exclusive explanation is that successful pretense requires suppressing available representations as a precondition for the model-based simulation process, and that words are harder to suppress than grid configurations. Either way, identifying the limiting conditions on epistemic pretense would be an important next step for understanding the underlying cognitive mechanisms, and in identifying the scope and content of people's models of their own minds

Our findings speak not only to people's ability to simulate counterfactual mental states, but also to their ability to pretend, deceive and lie more broadly. Previous research has mostly focused on the simulation of counterfactual world states, with theoretical models that suggest a key role for model-based simulations in pretense behaviour (Nichols & Stich, 2000; Weisberg & Gopnik, 2013), a role for pretense in the development of reasoning about causation (Walker & Gopnik, 2013), and hard constraints on the capacity to deceive (DePaulo et al., 2003; Verschuere et al., 2023; Walczyk, Roper, Seemann, & Humphrey, 2003). Others have focused on the interaction between liars and recipients, modelling the effect of liars' models of recipients' mental states (Oey et al., 2023) and showing consistently poor ability of observers to detect lies or pretense in others (Bond & DePaulo, 2006). In contrast, our focus here is on a special kind of pretense, one involving simulations of a counterfactual internal belief state rather than a counterfactual state of the external world, and with no reference to a specific recipient. Such

simulations are required not only in adversarial settings such as pretense and deceit, but also in teaching and explaining (“*Would I have understood my explanation if I was not familiar with the subject matter?*”), fairness judgments (“*Would I have been so impressed with this candidate if I didn’t know they went to Harvard?*”), intelligence attribution based on observed behaviour (“*They solved the puzzle faster than it would have taken me to solve it had I not known the solution?*”) and legal settings (“*Please ignore this witness’s testimony in your decision, as they were found unreliable?*”). As such, while our findings should be considered within the broader context of people’s ability to behave in accordance with an imaginary world state, we focus not on the dependence of deceit on models of the world or of other agents, but on its reliance on a model of the self. We suggest that this novel perspective may open entirely new avenues for research about self-models and metacognitive knowledge.

Together, our findings reveal a non-trivial capacity for pretending not to know. Complementing previous work on cognitive and perceptual hindsight biases, which traditionally focus on people’s inability to emulate ignorance, we show that people are in fact capable of accurately simulating diverse aspects of their decision-making processes, although they exhibit systematic shortcomings. We speculate that these shortcomings are consistent with the simulation of a simplified model of cognition, over and above any suppression of knowledge or sensory input. In revealing this powerful capacity, our findings raise many new theoretical questions to which we don’t yet have answers. Are there specific aspects of our knowledge, beliefs, or inferences that are harder than others to simulate, and is this related to a lack of metacognitive understanding of these aspects? Does pretending not to know rely on explicit, reportable self-knowledge, or on an implicit self-model? Is the ability to overcome the curse of knowledge in the context of pretending predictive of the ability to overcome it in communicating information to a

naive audience? Further research into these and similar limitations may continue to reveal the simplifications, abstractions, and biases in people's models of their own minds.

### 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.
- Bar-Hillel, M., & Wagenaar, W. A. (1991). The perception of randomness. *Advances in Applied Mathematics*, 12(4), 428–454. [https://doi.org/10.1016/0196-8858\(91\)90029-I](https://doi.org/10.1016/0196-8858(91)90029-I)
- Bear, A., Bensinger, S., Jara-Ettinger, J., Knobe, J., & Cushman, F. (2020). What comes to mind? *Cognition*, 194, 104057.
- 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](https://doi.org/10.1207/s15327957pspr1003_2)
- DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, 129(1), 74–118. <https://doi.org/10.1037/0033-2909.129.1.74>
- Earp, B. D., Dill, B., Harris, J. L., Ackerman, J. M., & Bargh, J. A. (2013). No sign of quitting: incidental exposure to “no smoking” signs ironically boosts cigarette-approach tendencies in smokers. *Journal of Applied Social Psychology*, 43(10), 2158–2162. <https://doi.org/10.1111/jasp.12202>
- Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review*, 104(2), 301–318. <https://doi.org/10.1037/0033-295X.104.2.301>
- 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>

- Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, 2(12), 493–501. [https://doi.org/10.1016/S1364-6613\(98\)01262-5](https://doi.org/10.1016/S1364-6613(98)01262-5)
- Giuliano, R. J., & Wicha, N. Y. (2010). Why the white bear is still there: Electrophysiological evidence for ironic semantic activation during thought suppression. *Brain Research*, 1316, 6274.
- 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.
- Gordon, R. (2007). *Moorean pretense*. Clarendon Press.
- Gordon, R. M. (1986). Folk Psychology as Simulation. *Mind & Language*, 1(2), 158–171. <https://doi.org/10.1111/j.1468-0017.1986.tb00324.x>
- Graziano, M. S., & Webb, T. W. (2015). The attention schema theory: A mechanistic account of subjective awareness. *Frontiers in Psychology*, 6, 500.
- 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>
- Jansen, R. A., Rafferty, A. N., & Griffiths, T. L. (2021). A rational model of the DunningKruger effect supports insensitivity to evidence in low performers. *Nature Human Behaviour*, 5(6), 756–763. <https://doi.org/10.1038/s41562-021-01057-0>

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness.

*Cognitive Psychology*, 3(3), 430–454. [https://doi.org/10.1016/0010-0285\(72\)90016-3](https://doi.org/10.1016/0010-0285(72)90016-3)

Kossale, Y., Airaj, M., & Darouichi, A. (2022, October). 2022 8th international conference on optimization and applications (ICOA). 1–6.

<https://doi.org/10.1109/ICOA55659.2022.9934291>

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., & Fleming, S. M. (2021). The Dunning-Kruger effect revisited. *Nature Human Behaviour*, 5(6), 677–678.

<https://doi.org/10.1038/s41562-021-01101-z>

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>

Nichols, S., & Stich, S. (2000). A cognitive theory of pretense. *Cognition*, 74(2), 115–147.

[https://doi.org/10.1016/S0010-0277\(99\)00070-0](https://doi.org/10.1016/S0010-0277(99)00070-0)

Nickerson, R. S. (2002). The production and perception of randomness. *Psychological Review*, 109(2), 330–357. <https://doi.org/10.1037/0033-295X.109.2.330>

Oey, L. A., Schachner, A., & Vul, E. (2023). Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General*, 152(2), 346–362.

<https://doi.org/10.1037/xge0001277>

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

Perner, J. (1996). *Simulation as explication of predication-implicit knowledge about the mind: Arguments for a simulation-theory mix* (P. Carruthers & P. K. Smith, Eds.). Cambridge University Press.

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>

- Saxe, R. (2005). Against simulation: the argument from error. *Trends in Cognitive Sciences*, 9(4), 174–179. <https://doi.org/10.1016/j.tics.2005.01.012>
- Uyeda, K. M., & Mandler, G. (1980). Prototypicality norms for 28 semantic categories. *Behavior Research Methods & Instrumentation*, 12(6), 587–595.
- Verschuere, B., Lin, C.-C., Huismann, S., Kleinberg, B., Willemse, M., Mei, E. C. J., ... Meijer, E. (2023). The use-the-best heuristic facilitates deception detection. *Nature Human Behaviour*, 1–11. <https://doi.org/10.1038/s41562-023-01556-2>
- Walczyk, J. J., Roper, K. S., Seemann, E., & Humphrey, A. M. (2003). Cognitive mechanisms underlying lying to questions: response time as a cue to deception. *Applied Cognitive Psychology*, 17(7), 755–774. <https://doi.org/10.1002/acp.914>
- Walker, C. M., & Gopnik, A. (2013). Pretense and possibilitya theoretical proposal about the effects of pretend play on development: Comment on lillard et al. (2013). *Psychological Bulletin*, 139(1), 40–44. <https://doi.org/10.1037/a0030151>
- 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.
- Weisberg, D. S., & Gopnik, A. (2013). Pretense, Counterfactuals, and Bayesian Causal Models: Why What Is Not Real Really Matters. *Cognitive Science*, 37(7), 1368–1381. <https://doi.org/10.1111/cogs.12069>
- 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>

## Supplementary Materials for

**Pretending not to know reveals a capacity for model-based self-simulation.**

## Extended Materials and 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](https://osf.io/v9zsb). The pre-registration was time-locked using cryptographic randomization-based time-locking (40) (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](https://osf.io/3thry). The pre-registration was time-locked using cryptographic randomization-based time-locking (40). 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:

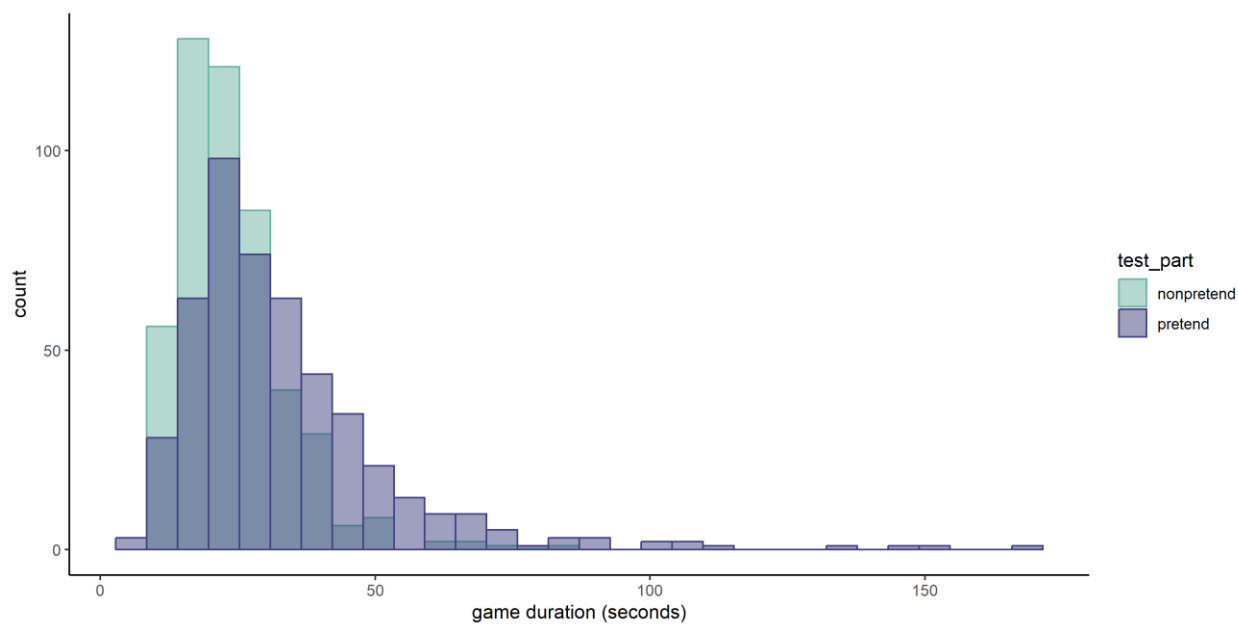
*“We would appreciate it if you could share any thoughts you had about the experiment, or anything we should take into account when analyzing your data.”*

Readers are invited to try a [demo of the experiment](#).

**Pre-registered analysis**Exp. 1: Battleship

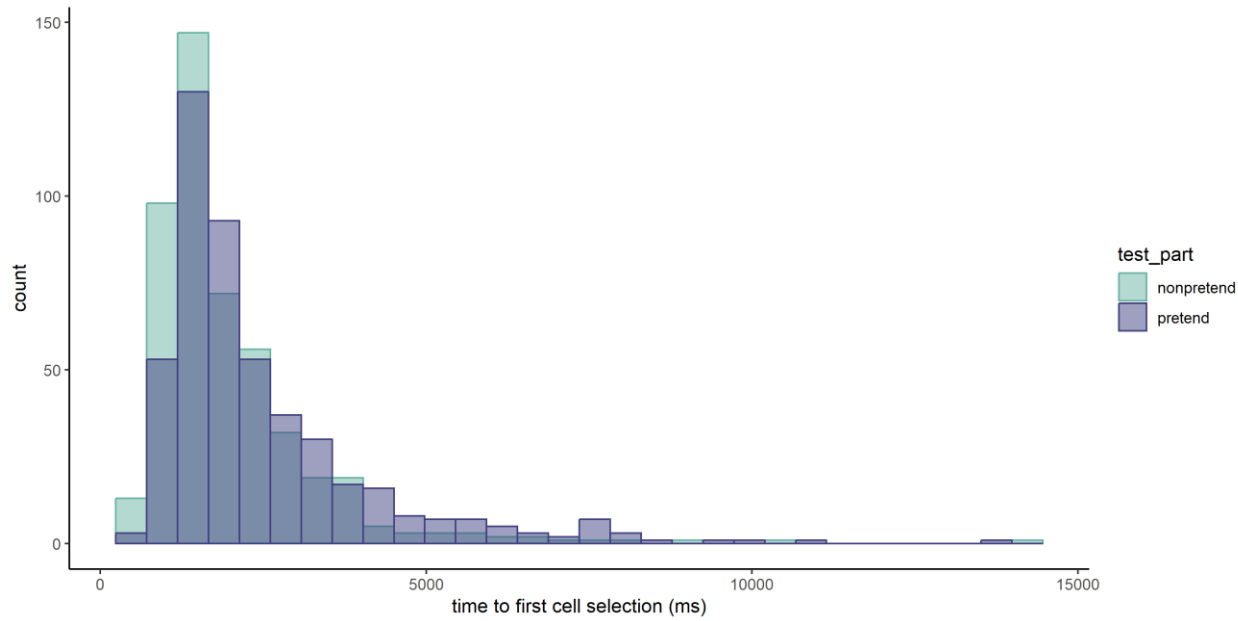
We designed our analyses to explore subjects' capacity for self-simulation under a counterfactual knowledge state, and the limits of this capacity. We focused on where subjects clicked and when, and asked whether this differed between pretend and non-pretend games. All analyses were pre-registered unless otherwise specified. In our pre-registration document, we committed to separately analyzing participants according to whether they pretended before or after completing a non-pretend block. Whenever there is a mismatch between the two groups, we focus on the first block of each participant (only pretend games from participants who pretended first, and only standard games from participants who played normally first). We do so to ensure that any successful pretending is not due to memory of one's own behavior in a previous block, and that non-pretend games are not biased by experience with the pretend block.

***Hypothesis 1 (game duration).*** Pretend games were longer than non-pretend games by 6 seconds on average. This was true in the pretend-first group ( $t(233) = 11.59, p < .001$ ), and in the pretend-second groups ( $t(245) = 6.41, p < .001$ ).



**Figure S1: game duration in Battleship.**

***Hypothesis 2 (first click latency).*** The first click in pretend games took longer to execute by 321 ms. This was true in the pretend-first group ( $t(233) = 8.24, p < .001$ ), but not in the pretend-second groups ( $t(245) = -1.03, p = .306$ ). As per our pre-registration document, we focused our analysis on the first blocks only, using a between-subject t-test. This test revealed a significant difference in the first click latency ( $\Delta M = 750.61, 95\% \text{ CI } [374.56, 1,126.67], t(391.22) = 3.92, p < .001$ ).



**Figure S2: first RT in Battleship**

***Hypothesis 3 (number of clicks).*** To sink all ships, players had to click on at least 7 and at most 25 squares. A simulated player that clicks randomly had a mean total click number of 22.75, and a near-optimal greedy player that consistently selected the square with the highest objective probability of containing a ship had a mean total click number of 13.62. Among our players, the mean number of clicks was 16.05 in non-pretend games and 15.83 in pretend games.

No significant difference in the total number of clicks was observed in the pretend-first group ( $t(243) = 1.43, p = .153$ ), but a significant difference was observed in the pretend-second group ( $t(255) = -3.71, p < .001$ ). Note that the two effects point in opposite directions, reflecting a tendency to make fewer clicks in the second block of the game. Crucially, when focusing on the first block only, we observed no significant difference in the number of clicks between the two conditions ( $t(440.10) = 0.41, p = .682$ ).

In 62 pretend games from 20 players, games were completed after 7 clicks only, without ever missing a ship. This never happened in non-pretend games. We assumed that these participants did not follow the task instructions, and excluded them from all analyses.

***Hypothesis 4 (click latency by outcome).*** In non-pretend games, cell selections that resulted in a hit were faster by 109 ms. than those that resulted in a miss (pretend-first group:  $t(233) = -8.73, p < .001$ ; pretend-second group:  $t(245) = -8.80, p < .001$ ).

In pretend games, cell selections that resulted in a hit were faster by 293 ms. than those that resulted in a miss (pretend-first group:  $t(233) = -11.83, p < .001$ ; pretend-second group:  $t(245) = -11.32, p < .001$ ). The effect of hitting versus missing a ship was significantly stronger in pretend games (pretend-first group:  $t(233) = -8.90, p < .001$ ; pretend-second group:  $t(245) = -6.20, p < .001$ ).

***Hypothesis 5 (click latency by previous outcome).*** In non-pretend games, cell selections that followed a hit were slower by 182 ms. than those that followed a miss (pretend-first group:  $t(233) = 10.83, p < .001$ ; pretend-second group:  $t(245) = 8.19, p < .001$ ).

In pretend games, cell selections that followed in a hit were slower by 236 ms. than those that followed a miss (pretend-first group:  $t(233) = 7.96, p < .001$ ; pretend-second group:  $t(245) = 9.52, p < .001$ ). The effect of following a hit versus a miss was significantly stronger in pretend games only in the pretend-second group ( $t(245) = 2.70, p = .007$ ; pretend-first group:  $t(233) = 1.40, p = .162$ ). Crucially, the effect was significantly stronger when focusing on the first blocks only ( $t(377.12) = 2.75, p = .006$ ).

***Hypothesis 6 (click latency by next outcome).*** In non-pretend games, cell selections that preceded a hit were slower than those that preceded a miss only in the group that played normally first (pretend-second group:  $M = -78.17, 95\% \text{ CI } [-103.34, -53.00], t(245) = -6.12, p < .001$ ), but not in the group that pretended first ( $M = 4.57, 95\% \text{ CI } [-22.11, 31.26], t(233) = 0.34, p = .736$ ).

In pretend games, the timing of cell selections did not covary with the outcomes of future cell selections (pretend-first group:  $t(233) = 1.13, p = .261$ ; pretend-second group:  $t(245) = 1.43, p = .155$ ). The effect of the next outcome was significantly weaker in pretend games only in the pretend-second group ( $t(245) = 3.95, p < .001$ ) but not in the pretend-first group ( $t(233) = 0.94, p = .346$ ). Crucially, the effect was significantly weaker when focusing on the first blocks only ( $t(301.81) = 3.26, p = .001$ ).

***Model based analysis.*** Pretend games were similar to non pretend games in the total number of clicks — but were they also similar in *where* participants clicked? More specifically, did cell selections in pretend games make sense given the limited information those participants pretended to have? To ask this, we approximated optimal behavior by calculating the probability that a ship is hidden in

each cell given available information,  $p(\text{ship}(x_i))$ , and the posterior probability that one should click on a square, assuming a uniform prior over cells  $P(x_i) = \frac{p(\text{ship}(x_i))}{\sum_{j=1}^k p(\text{ship}(x_j))}$ . Critically, in modeling pretend games we did not treat hints as part of this available information for extracting  $p(\text{ship}(x_i))$ , because an optimal player should ignore hints in choosing where to click next. Given this posterior map, a rational player should choose cells where  $P(x_i)$  is high [this behavior is not strictly optimal, but approximates optimal behavior in most cases; (23), Section 3.3]. To quantify optimality, before each cell selection we computed the posterior probability map for all ‘unknown’ cells. Then, we ranked cells from high to low according to their posterior probability and recorded the rank of the chosen cell: a lower rank indicating more optimal behavior.

***Hypothesis 7 (game optimality).*** The mean posterior rank of non-pretend cell selections was 6.44 and significantly lower (more optimal) than that of a simulated random agent (9.19,  $t(479) = -49.18$ ,  $p < .001$ ). Pretend games were significantly less optimal than non-pretend games (6.93; pretend-first group:  $t(233) = 8.49$ ,  $p < .001$ ; pretend-second group:  $t(245) = 7.11$ ,  $p < .001$ ), but still more optimal than those of a random agent ( $t(479) = -38.51$ ,  $p < .001$ ). Critically, the same pattern was observed when restricting analysis to cell selections that resulted in a miss (non-pretend - pretend:  $t(469.71) = -6.30$ ,  $p < .001$ ; pretend - random:  $t(479) = -10.25$ ,  $p < .001$ ). In other words, the optimality of pretend games relative to random cell selection was not merely due to the fact that pretenders clicked on ships more than expected by chance. Even when missing a ship, their cell selections made sense given the limited information they pretended to have.

***Hypothesis 8 (entropy-RT relation).*** When playing Battleships, it is sometimes clear what the next cell selection should be, and sometimes more difficult to decide where to click next. To capture this notion of decision uncertainty, we calculated the entropy of the posterior distribution over cell selections  $H(P) = -\sum_{i=1}^k P(x_i) \log P(x_i)$ , where  $P(x_i) = \frac{p(\text{ship}(x_i))}{\sum_{j=1}^k p(\text{ship}(x_j))}$ , and asked how this measure relates to decision latency, or the time taken to click on the next cell.  $H(P)$  is high when players need to

decide between multiple cells with a similar probability of hiding a ship, and low when there are only a few candidates with a high probability of hiding a ship. For every player and condition separately, we fitted a multiple linear regression to predict decision latency based on  $H(P)$  and  $H(P)^2$ . The resulting coefficients were then subjected to a group-level inference. The first cell selection of each game was excluded from this analysis, because entropy was constant for the first click.

In non-pretend games, we found no evidence for a linear relation between decision entropy and decision latency among participants who played normally in the first block ( $t(245) = -0.06, p = .950$ ). In contrast, a significantly negative linear modulation was observed in the group that pretended in the first and played normally in the second block ( $t(233) = -10.24, p < .001$ ).

In pretend games, this negative relation between  $H(P)$  and click latency was significant in both groups (pretend-first group:  $t(233) = -4.93, p < .001$ ; pretend-second group:  $t(245) = -5.54, p < .001$ ). We found no significant difference between the magnitude of the linear effect in pretend and non-pretend games (pretend-first group:  $t(233) = 0.63, p = .528$ ; pretend-second group:  $t(245) = -1.77, p = .079$ ).

Similarly, we found no evidence for a quadratic relation between decision entropy and decision latency in non-pretend games of participants who played normally in the first block ( $t(245) = -1.33, p = .184$ ). Again, a significantly negative quadratic modulation was observed in the non-pretend games of players who pretended in the first and played normally in the second block ( $t(233) = -13.26, p < .001$ ). Like the linear effect, this negative quadratic relation was significant in the pretend games of both groups (pretend-first group:  $t(233) = -10.21, p < .001$ ; pretend-second group:  $t(245) = -12.14, p < .001$ ). This negative quadratic effect was

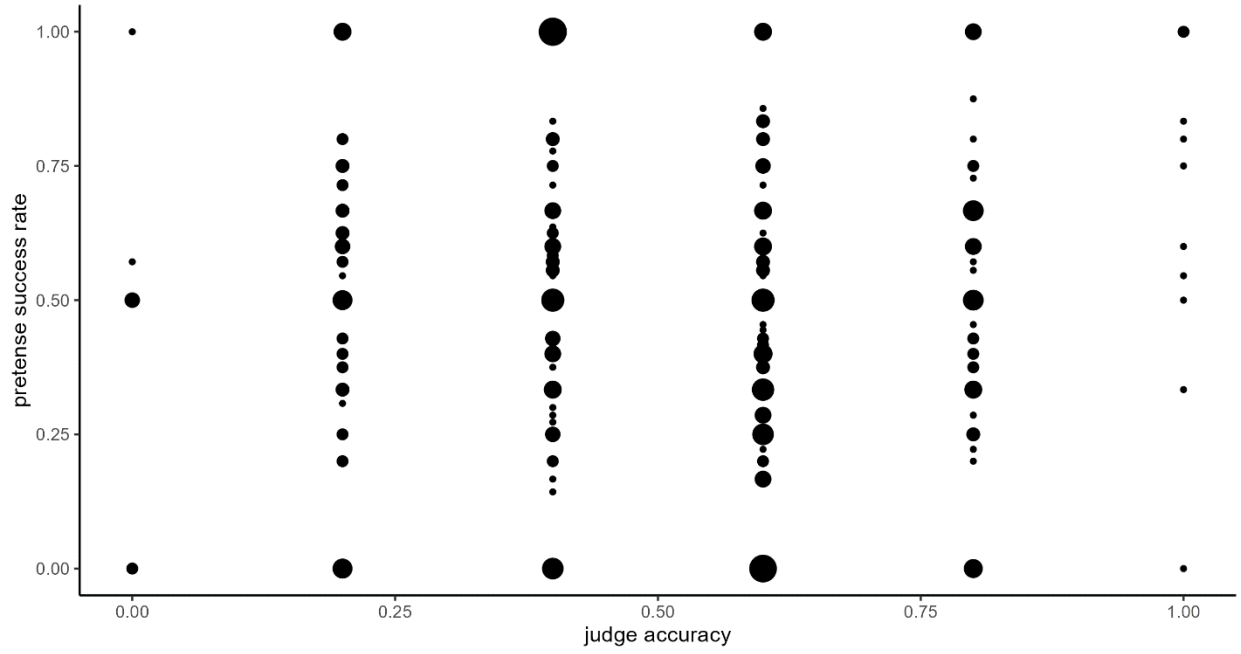
stronger in pretend compared to non-pretend games (pretend-first group:  $t(233) = -4.32, p < .001$ ; pretend-second group:  $t(245) = -3.49, p = .001$ ).

***Hypothesis 9 (judge accuracy).*** Participants were not significantly above chance in judging which of the two presented games came from a pretender (mean accuracy: 51%; t test against 50:  $t(499) = 1.45, p = .147$ ).

For a subset of 288 players, one of the five boards presented in the judging block was already presented both in the pretend and in the non-pretend blocks. Still, even this subset of participants who had just experienced a board both as pretenders and as non-pretenders were not significantly above chance in telling which of two other players had hints for this same board (mean accuracy: 55%,  $p = .141$ ).

***Hypothesis 10 (judge-pretend relation).*** Pretense quality was quantified as the number of pretend games that were misclassified by other participants as non-pretend games, divided by the number of pretend games that were presented to other participants. Due to randomization, this denominator is different for different participants. We tested for a correlation between pretense quality and judge accuracy at the group level. To maximize statistical power, this hypothesis was tested on the pooled data from both groups of participants. We find no evidence for such a correlation ( $r_s = -.05, S = 7,779,376.67, p = .328$ ). Furthermore, we find no significant correlation between participants' accuracy in detecting the pretender and their decision optimality in pretend games (measured as the mean rank posterior probability of their cell selections;  $r_s = .02, S = 17,999,242.64, p = .608$ ), nor with the cost to

optimality relative to non-pretend games ( $r_s = .02$ ,  $S = 18,089,259.81$ ,  $p = .685$ ).

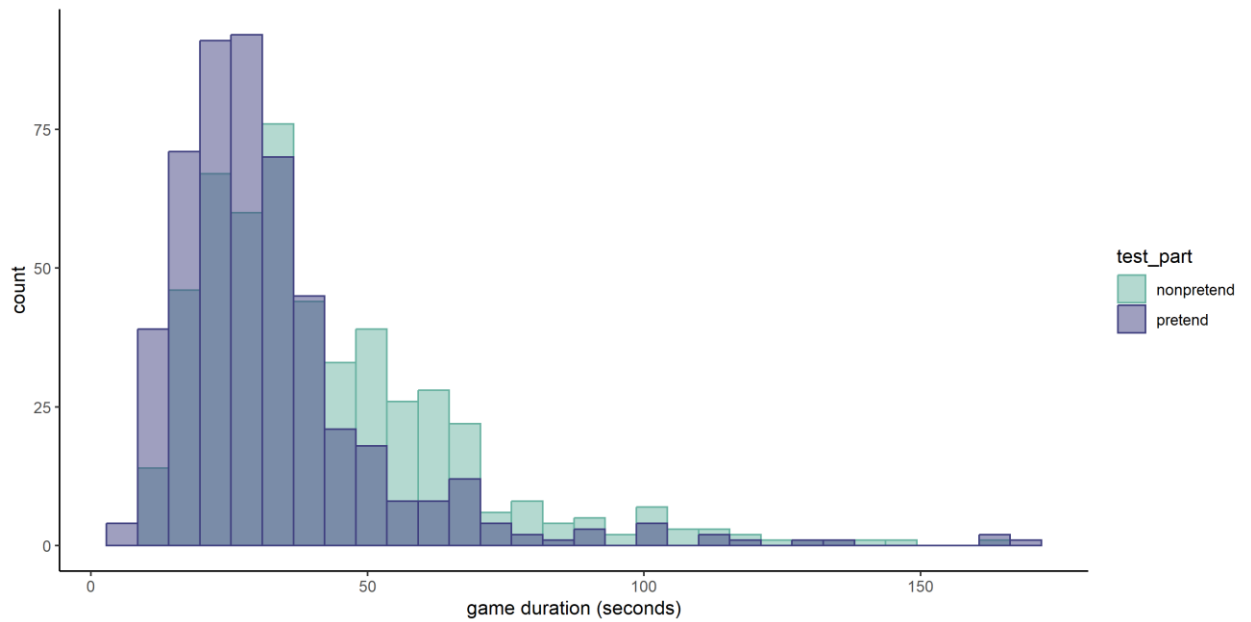


**Figure S3: Judge accuracy against pretense success rate in Battleship.** Marker sizes are proportional to the number of participants in each coordinate.

**Exp. 2: Hangman**

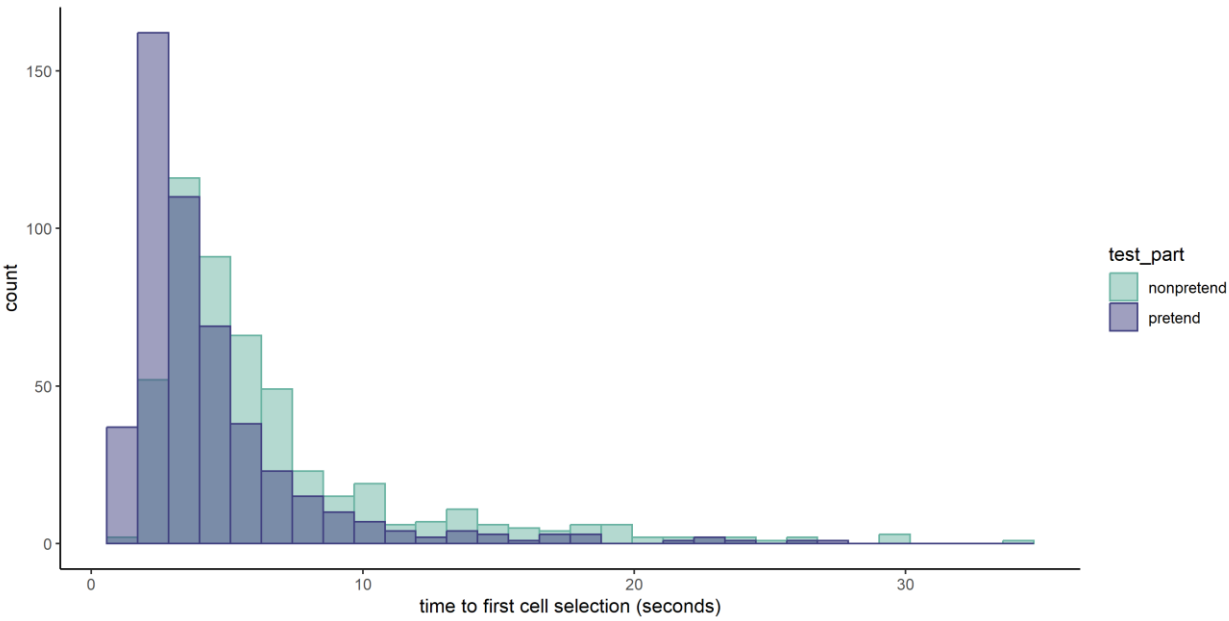
***Hypothesis 1 (Game duration).***

The median game duration in non-pretend games was 36 seconds in non-pretend games and 28 seconds in pretend games. Pretend games were significantly shorter among subjects that pretended in the first block ( $t(232) = -3.67, p < .001$ ) and in the second block ( $t(267) = -7.82, p < .001$ ).



**Figure S4: game duration in Hangman**

***Hypothesis 2 (first click latency).*** The median time taken to make the first letter selection was 3.32 seconds in pretend games and 4.98 in non-pretend games. First letter selections were slower in non-pretend games in both groups (pretend-first group:  $t(232) = 17.93, p < .001$ ; pretend-second group:  $t(267) = 20.02, p < .001$ ).



**Figure S5: first RT in Hangman**

***Hypothesis 3 (number of misses).*** The mean number of misses was 1.94 in non-pretend games and 2.80 in pretend games. The difference between the two was significant among players that pretended in the first part ( $t(232) = -5.96, p < .001$ ), and among those that pretended in the second part ( $t(267) = -3.26, p = .001$ ).

***Hypothesis 4 (click latency by outcome).*** In non-pretend games, cell selections that resulted in a hit were faster by 386 ms than those that resulted in a miss (pretend-first group:  $t(227) = -4.62, p < .001$ ; pretend-second group:  $t(265) = -4.28, p < .001$ ).

In pretend games, cell selections that resulted in a hit were faster by 297 ms than those that resulted in a miss (pretend-first group:  $t(228) = -3.36, p = .001$ ; pretend-second group:  $t(252) = -2.42, p = .016$ ). The effect of hitting versus missing a ship was weaker in pretend games in the group that pretended first ( $t(223) = 1.98, p = .049$ ) but not in the group that played normally first ( $t(250) = 0.24, p = .812$ ). Crucially, pretend and nonpretend games were not different according to this measure when focusing on the first block only ( $t(459.12) = 1.59, p = .112$ ).

***Hypothesis 5 (click latency by previous outcome).*** In non-pretend games, cell selections that followed a hit were not faster or slower than those that followed a miss (pretend-first group:  $t(231) = 0.25, p = .803$ ; pretend-second group:  $t(267) = 0.33, p = .741$ ).

In contrast, pretend cell selections that followed a hit were slower by 667 ms. than those that followed a miss (pretend-first group:  $t(229) = 2.84, p = .005$ ; pretend-second group:  $t(256) = 3.44, p = .001$ ). The effect of following a hit versus a miss was not significantly stronger in pretend games (pretend-first group:  $t(228) = 1.12, p = .264$ ; pretend-second group:  $t(256) = 0.15, p = .883$ ).

***Hypothesis 6 (click latency by next outcome).***

Letter selections preceding a hit were faster than those preceding a miss in non-pretend games only in the group that pretended first ( $t(219) = -3.40$ ,  $p = .001$ ) but not in the group that played normally first ( $t(257) = -1.86$ ,  $p = .064$ ). This effect was not observed in pretend games (pretend first:  $t(223) = -0.57$ ,  $p = .567$ ; non-pretend first:  $t(241) = -1.01$ ,  $p = .311$ ). We find a significant interaction between these effects in the group that pretended first ( $t(210) = 2.75$ ,  $p = .006$ ) but not in the other group ( $t(231) = 1.67$ ,  $p = .097$ ). The difference was not significant when analyzing first blocks only ( $t(260.60) = -1.80$ ,  $p = .072$ ).

***Model based analysis.***

The next analyses were designed to test for differences in game optimality between pretend and non-pretend games, and for a relationship between decision difficulty and click latency. To do so, we approximated optimal behavior by approximating the posterior probability that a letter appears in the word, given available information. Critically, in modelling pretend games we do not treat hints as part of this available information, because an optimal player should ignore this information in choosing where to click next. Given this posterior, a rational player should choose letters that have a high posterior probability of appearing in the word.

To approximate the posterior probability of letters given a game state, we followed the following procedure:

1. We used the category information (e.g., ‘a fruit’), to obtain a probability-weighted list of category-compatible words (or names, in the case of famous people). The lists were obtained in the following way: for US states and number names, we used an exhaustive list (in the case of numbers, of numbers of 1-2 words), fruit names were taken from Wikipedia ([simple.wikipedia.org/wiki/List\\_of\\_fruits](https://simple.wikipedia.org/wiki/List_of_fruits)), famous people names from a [crowdsourced document](#), and body-part words from existing prototypicality norms (Uyeda and Mandler, 1980)). We extended Uyeda and Mandler’s list of body part words by adding all body part words commonly appearing in lists on the internet. In the case of famous people, names were given prior probabilities in proportion to the number of visits their Wikipedia entries received in 2021. The top 100 most popular entries were given a prior probability twice that of the next 100, three times that of the next 100, and so forth. All entries from 600 and on were given the same prior probability, seven times smaller than that of entries at the top 100 positions. In the cases of fruits and body parts, we used prototypicality norms from Uyeda and Mandler (1980) to assign higher prior probability to more prototypical words (mapped to lower numbers on the 1-7 scale used by Uyeda

and Mandler). Words that were not included in the norm were given the maximum score (that is, the lowest perceived prototypicality), of 7. Body parts and fruits were included in both singular and plural forms. The plural forms were assigned a prototypicality score of 100, due to the fact that category names were presented in the singular form ('a fruit' and 'a body part'). We then took the reciprocal for each score, and normalized it by the total sum to get a probability distribution over words  $p(w)$ . The full prior distributions for each category are included in this pre-registration.

2. The likelihood of a game state given a target word  $p(s|w)$  equals 0 when the word is inconsistent with the information available to the player (this includes the word length, the identity of letters that do not appear in the word based on previous guesses, and the identity and position of letters that do appear in the word based on previous guesses). When consistent, the likelihood is a non-zero quantity that is equal for all consistent words. Bayes rule was used to extract the posterior over words given game state  $p(w|s) = \frac{p(s|w)p(w)}{\sum_{w'} p(s|w')p(w')}$ . The full non-zero  $p(w|s)$  distributions for the initial states of all games are included in this pre-registration.
3. The probability that an individual letter appears in the target word  $p(l|s)$  is the sum of posterior probabilities of words that contain this letter  $p(l|s) = \sum_w p(w|s) \times \text{includes}(w, l)$  where  $\text{includes}(w, l)$  returns 1 if  $l$  appears in  $w$  and 0 otherwise.
4. A posterior probability over letter selections was obtained by dividing the probabilities of individual letters by their total sum  $\frac{p(l|s)}{\sum_l p(l|s)}$ .

To quantify this notion of optimality, before each cell selection we computed the posterior probability that each of the unclicked letters appeared in the target word, given the game state. Then, we ranked letters from highest to lowest according to their posterior probability and recorded the rank of the chosen cell.

**Hypothesis 7 (optimality).** Letter selections in non-pretend games had a higher posterior probability to be included in the word than in random games (pretend first:  $t(232) = -59.66, p < .001$ ; non-pretend first:  $t(267) = -63.07, p < .001$ ). The same was true for letter selections in pretend games (pretend first:  $t(232) = -46.83, p < .001$ ; non-pretend first:  $t(267) = -55.96, p < .001$ ). Non-pretend letter selections had a higher posterior probability than pretend letter selections among players who pretended first ( $t(232) = 5.96, p < .001$ ) but not in the group that played normally first ( $t(267) = 0.72, p = .469$ ). Comparing data from the first condition only in a between-subject manner revealed a significant difference ( $\Delta M = -0.52, 95\% \text{ CI } [-0.77, -0.28], t(453.21) = -4.25, p < .001$ ).

When focusing on misses, non-pretend letter selections were more optimal than random (pretend first:  $t(231) = -20.23, p < .001$ ; non-pretend first:  $t(267) = -22.41, p < .001$ ), as well as pretend games (pretend first:  $t(229) = -13.46, p < .001$ ; non-pretend first:  $t(257) = -13.37, p < .001$ ). In both cases, non-pretend games were more optimal than pretend games (pretend first:  $t(228) = 6.60, p < .001$ ; non-pretend first:  $t(257) = 5.36, p < .001$ )

***Hypothesis 7b (Letter frequency).*** Non-pretenders selected high-frequency letters more than expected by chance (pretend first:  $t(232) = -70.56, p < .001$ ; non-pretend first:  $t(267) = -72.51, p < .001$ ). The same was true for letter selections in pretend games (pretend first:  $t(232) = -53.04, p < .001$ ; non-pretend first:  $t(267) = -65.57, p < .001$ ). Letters selected by pretenders had a higher overall frequency compared to letters selected by non-pretenders (pretend first:  $t(232) = -2.41, p = .017$ ; non-pretend first:  $t(267) = -5.43, p < .001$ ). Comparing data from the first condition only in a between-subject manner revealed a significant difference ( $\Delta M = -0.52, 95\% \text{ CI } [-0.77, -0.28], t(453.21) = -4.25, p < .001$ ).

When focusing on misses only, pretend letter selections were again more aligned with the frequency of letters in English words than were non-pretend letter selections (pretend first:  $t(228) = -0.79, p = .433$ ; non-pretend first:  $t(257) = 5.36, p < .001$ ; first block only:  $\Delta M = -0.98, 95\% \text{ CI } [-1.31, -0.64], t(489.98) = -5.80, p < .001$ ).

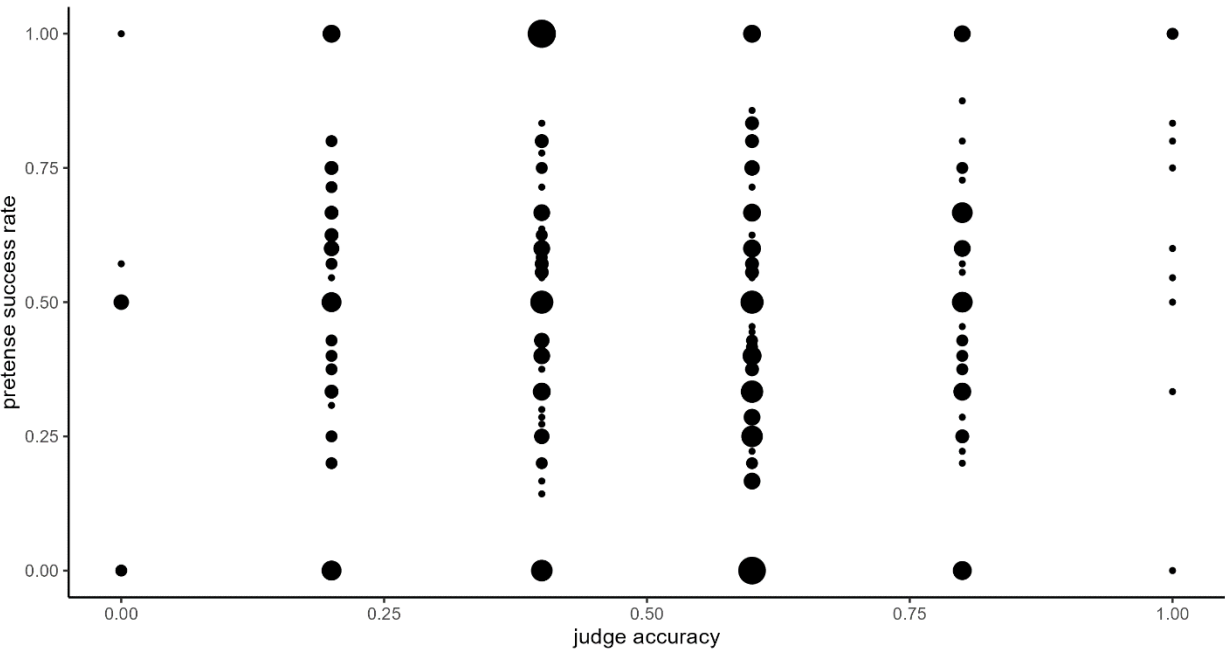
***Hypothesis 8 (Entropy/RT relationship).*** In non-pretend games, high uncertainty was associated with longer decision times (pretend-first group:  $t(232) = 9.50, p < .001$ ; pretend-second group:  $t(267) = 10.68, p < .001$ ). The same was true for pretend games (pretend-first group:  $t(232) = 4.66, p < .001$ ; pretend-second group:  $t(267) = 7.65, p < .001$ ). The effects were significantly weaker in pretend games (pretend-first group:  $t(232) = -2.51, p = .013$ ; pretend-second group:  $t(232) = -2.51, p = .013$ ).

In non-pretend games, the relationship between uncertainty and RT was negatively quadratic, such that the longest responses were the one associated with intermediate uncertainty levels (pretend-first group:  $t(232) = -7.75, p < .001$ ; pretend-second group:  $t(232) = -7.75, p < .001$ ). The same was true for pretend games (pretend-first group:  $t(232) = -1.79, p = .075$ ; pretend-second group:  $t(267) = -3.43, p = .001$ ). The effects were significantly weaker in pretend games (pretend-first group:  $t(232) = 3.72, p < .001$ ; pretend-second group:  $t(232) = 3.72, p < .001$ ).

***Hypothesis 9 (judge accuracy):*** Participants were not significantly above chance in judging whether the presented game came from a pretender or from a genuine player ( $M = 0.51$ , 95% CI [0.49,0.53],  $t(452) = 0.57, p = .568$ ). This was the case also when focusing on words that judges encountered when pretending ( $M = 0.52$ , 95% CI [0.49,0.55],  $t(452) = 1.30, p = .194$ ), and when focusing on words that judges encountered when playing normally ( $M = 0.49$ , 95% CI [0.46,0.52],  $t(452) = -0.57, p = .572$ ).

***Hypothesis 9B (judge bias):*** Participants were slightly biased to classify games as pretend games ( $M = 0.53$ , 95% CI [0.51,0.55],  $t(452) = 3.46, p = .001$ ).

***Hypothesis 10 (judge-pretend relation):*** No significant correlation between judge accuracy and ability to trick others ( $r_s = .00, S = 5,089,032.77, p = .941$ ), nor with the optimality cost to pretending ( $r_s = -.10, S = 5,613,970.32, p = .082$ ).



**Figure S6: Judge accuracy against pretense success rate in Hangman.** Marker sizes are proportional to the number of participants in each coordinate.

***Hypothesis 12 (irrational letter selections in half-games).*** In non-pretend games, subjects almost always chose letters that made sense, that is, that were consistent with at least one word that matched the presented category and game state (pretend-first group: 0.97; pretend-second group: 0.98). In pretend games, these figures were significantly lower (pretend-first group: 0.73,  $p < .001$ ; pretend-second group: 0.82,  $p < .001$ ).

***Hypothesis 13 (target word effect in half-games).*** Excluding irrational clicks, non-pretenders had an overall preference to select letters that are consistent with the words ‘BANANA’ and ‘HAND’ (pretend-first group: 0.82; pretend-second group: 0.86). Pretenders’ preference for the same letters was weaker (pretend-first group: 0.66,  $p < .001$ ; pretend-second group: 0.70,  $p < .001$ ).

Within pretend games, preference for letters that are consistent with HAND or BANANA was significantly stronger when the word itself was BANANA or HAND, compared to when it was PAPAYA or HAIR (pretend-first group: 0.80 vs. 0.52,  $p < .001$ ; pretend-second group: 0.97 vs. 0.41,  $p < .001$ ).



## Exploratory analysis

### Exp. 1: Battleship

#### Half games

Our optimality analysis showed that pretenders' click selections closely resemble those of non-pretenders, at least in that they are not random, but rather guided by where a ship might be. However, due to the high number of possible board configurations, data from full games provide limited opportunity to compare cell selections for specific game states. In addition to asking, "What guides cell selections in pretend and non-pretend games?", we also wanted to ask, "Where exactly would pretenders and non-pretenders click, given a specific board configuration?".

To achieve this, the sixth game in each block started not with an empty grid, but with the contents of some cells already revealed by a previous player. As before, pretenders also knew where the remaining ships were hidden, but tried to play as if they only knew what was known to this previous player. Having cell selections from 250 players for each board configuration and condition allowed us to plot and compare the distribution of clicks under a genuine, or pretend, knowledge state.

In the third column we plot the distribution of clicks for non-pretend players. This distribution is in agreement with the hit probability map (board A:  $r = .81$ , 95% CI [.55, .93],  $t(16) = 5.48$ ,  $p < .001$ ; board B:  $r = .87$ , 95% CI [.68, .95],  $t(16) = 7.01$ ,  $p < .001$ ). Finally, in the fourth column we plot the distribution of cell selections for pretend players. Although noisier, this distribution is also in agreement with the hit probability map (board A:  $r = .49$ , 95% CI [.03, .78],  $t(16) = 2.27$ ,  $p = .037$ ; board B:  $r = .73$ , 95% CI [.40, .89],  $t(16) = 4.28$ ,  $p =$

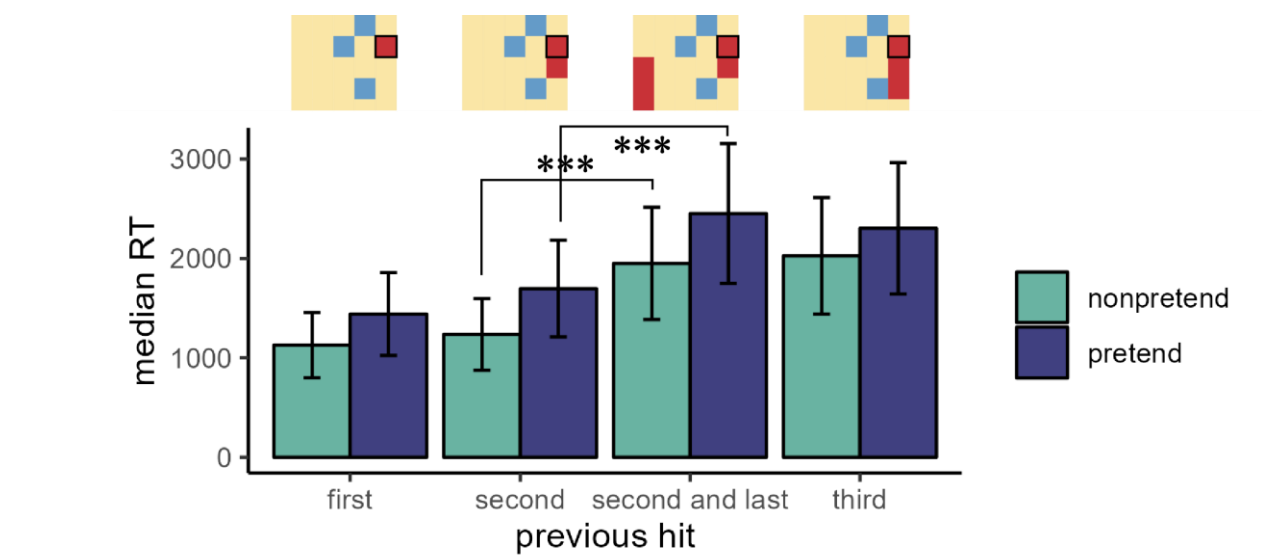
.001), and more importantly, with the hit distribution of non-pretenders for the same board configuration (board A:  $r = .58$ , 95% CI [.15, .82],  $t(16) = 2.83$ ,  $p = .012$ , board B:  $r = .87$ , 95% CI [.69, .95],  $t(16) = 7.23$ ,  $p < .001$ ).

**Ship completion** 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. S7). 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 ( $p < .001$ ).

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

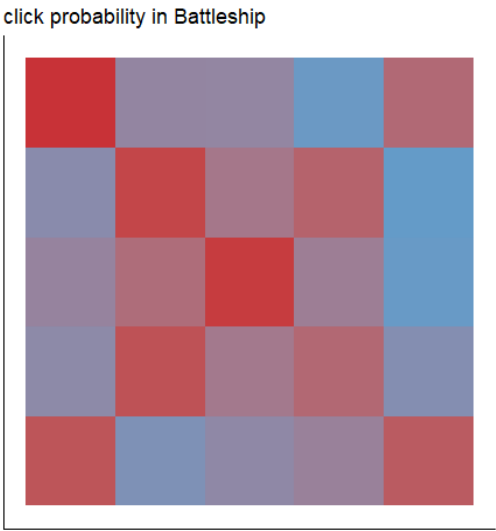
ship that is in fact of size-two ( $p < .001$ ), and when controlling for the physical distance traversed by the mouse from the previous click ( $p < .001$ ).



**Figure S7: The effect of ship completion on decision times in pretend and non-pretend games.** Error bars and shaded areas represent the bootstrapped standard error of the median.

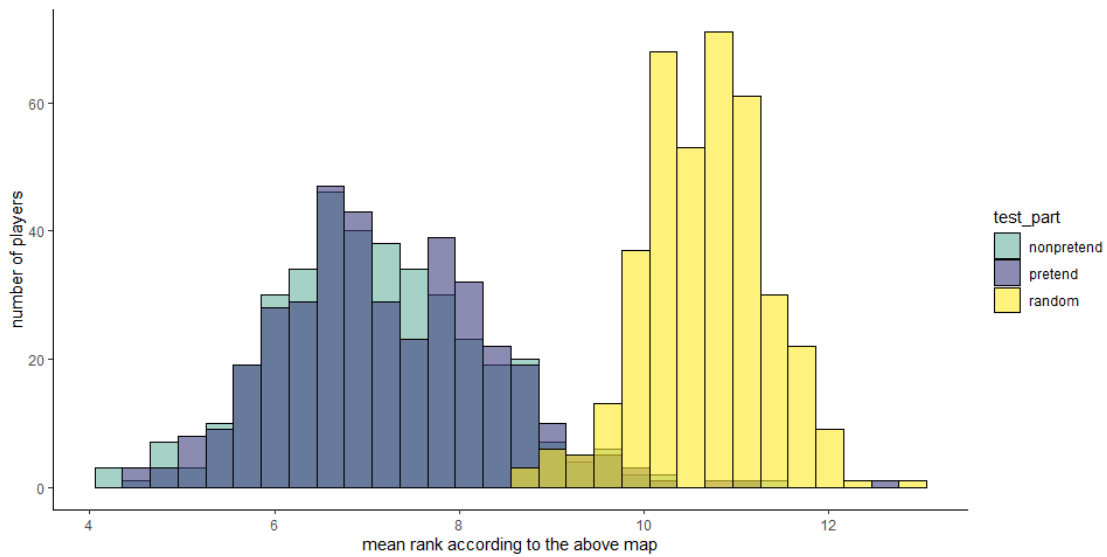
#### Heuristic use

In Battleship, players sometimes have a fixed sequence of clicks that they follow, at least for the first moves. To quantify this, we plot the distribution of the first 5 clicks from all games of a random subset of 100 players:



**Figure S8: Guess distribution in Battleship, first 5 guesses.** The map was extracted from all games of a random subset of 100 players.

As can be seen, subjects prefer to click on the diagonals, with a slight preference for the left-hand side. We then asked how well aligned are players' clicks with this map in pretend and non-pretend games. For this analysis, we only used data from the remaining 380 players whose games were not used for generating the above map.



**Figure S9: Alignment of pretend and non-pretend guesses with mean guess probability.** Higher values indicate lower alignment.

We find no difference in this measure between pretend and non-pretend games ( $t(379) = -1.51, p = .131$ ). In the next analysis we use a more sensitive, player-specific measure of players' tendency to follow a fixed heuristic, versus playing flexibly.

#### Variability in number of misses

We quantified the within-subject variability in the number of misses in pretend and non-pretend games. This is a function of how similar games were to each other, within condition and within individual. The standard deviation in non-pretend games was on average 2.60, and significantly higher than in pretend games (mean SD = 1.58,  $t(479) = 18.12, p < .001$ ). In other words, pretend games were significantly more similar to each other, in number of guesses, than non-pretend games were to each other.

#### Within-participant sequential guess entropy

In this analysis, we asked how flexible players were in their click sequences, across games. We quantified flexibility as the Shannon entropy of cell selections across games, for a given click number. For example, if a player always starts their games by clicking in the top left corner, their flexibility score for the first click will be  $H([1,1,1,1,1]) = 0$ . Flexibility was quantified separately for pretend and non-pretend games.

Flexibility scores increased as a function of click number for both pretend and non-pretend games, as expected if players adjust their behavior based on the outcomes of previous clicks. This process reached a plateau of around  $H = 2.1$  at click number 6.

Importantly, before that point, flexibility was persistently lower in pretend games (click #1:  $t(479) = -3.93, p < .001$ ; click #2:  $t(479) = -3.40, p = .001$ ; click #3:  $t(479) = -4.88, p < .001$ ; click #4:  $t(479) = -3.26, p = .001$ ; click #5:  $t(479) = -2.64, p = .009$ ).

We reasoned that this tendency to rigidly follow a pre-defined plan may underlie at least some of the difference in game optimality between pretend and non-pretend games. Interestingly, however, we find no significant correlation between the pretend/non-pretend differences in flexibility and optimality in the first 5 clicks ( $r = -.07$ , 95% CI  $[-.15, .02]$ ,  $t(478) = -1.44$ ,  $p = .150$ ).

#### Suppression or simulation?

At the end of the experiment, players were asked whether they had a strategy for pretending or for detecting pretense. I (MM) manually labeled their responses according to three criteria: mentions of suppression (e.g., “Trying to ignore the hints”), simulation (e.g., “Trying to imagine how I would react”), and rules (e.g., “Checking the corners and center before random selection of tiles”). Some answers got more than one label and some got none. Overall, 32 mentioned using suppression, 127 mentioned using simulation, and 220 mentioned following rules.

In the following analysis, we focus on the first two labels and sort players into two groups: 23 players who reported using suppression but not simulation, and 118 who reported using simulation but not suppression.

**Click optimality** The mean negative optimality score of non-pretend games was 6.45 among suppressors and 6.33 among simulators ( $t(25.42) = 0.45$ ,  $p = .655$ ). The mean negative optimality score of pretend games was 7.14 among suppressors and 6.64 among simulators ( $t(28.77) = 1.94$ ,  $p = .062$ ). Note that a higher score here means lower optimality. To the very least, simulators did as well as suppressors in pretending.

**RT by outcome** Visual inspection reveals similar temporal profiles in both groups (the higher standard error among suppressors is due to the lower number of subjects in this group). Specifically, in pretend games, cell selections that resulted in a hit were faster than those

that resulted in a miss in both groups (suppressors:  $M = -735.59$ , 95% CI  $[-1,094.51, -376.68]$ ,  $t(21) = -4.26$ ,  $p < .001$ ; simulators:  $M = -441.08$ , 95% CI  $[-543.74, -338.42]$ ,  $t(117) = -8.51$ ,  $p < .001$ ), and cell selections that followed a hit were slower than those that followed a miss in both groups (suppressors:  $M = 322.87$ , 95% CI  $[27.88, 617.87]$ ,  $t(21) = 2.28$ ,  $p = .033$ ; simulators:  $M = 478.59$ , 95% CI  $[332.19, 624.99]$ ,  $t(117) = 6.47$ ,  $p < .001$ ).

### Exp. 2: Hangman

#### Correlation in number of misses

Some words were more difficult to reveal. This required in players making more unsuccessful letter guesses in attempting to reveal some words compared to others. When pretending, players made more unsuccessful letter guesses for the exact same words  $r_s = .97$ ,  $S = 5.48$ ,  $p < .001$ .

#### Variability in number of misses

Similar to Battleship, the within-subject standard deviation in number of misses in non-pretend games (2.65) was on average significantly higher than in pretend games (mean SD = 1.53,  $t(500) = 12.65$ ,  $p < .001$ ). Here too, pretend games were significantly more similar to each other, in terms of number of misses, than non-pretend games.

#### Within-participant sequential guess entropy

Like for Battleship, flexibility scores increased as a function of click number for both pretend and non-pretend games, as expected if players adjust their behavior based on the outcomes of previous clicks. This process reached a plateau of around  $H = 2$  and click number 6.

Importantly, before that point, flexibility was persistently lower in pretend games (click #1:  $t(500) = -12.67, p < .001$ ; click #2:  $t(500) = -10.17, p < .001$ ; click #3:  $t(500) = -10.39, p < .001$ ; click #4:  $t(469) = -8.30, p < .001$ ; click #5:  $t(99) = -1.61, p = .110$ ).

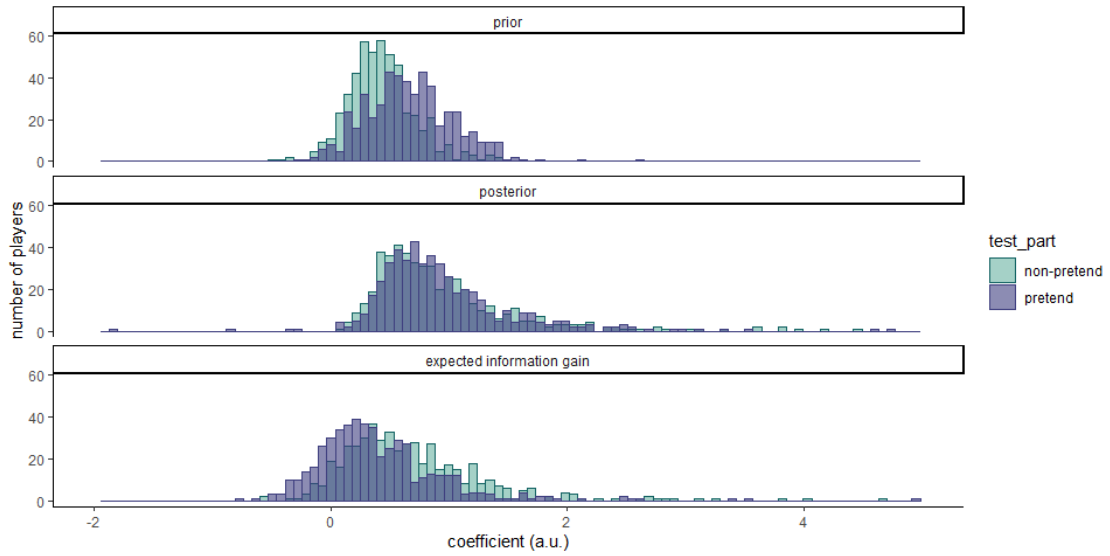
We reasoned that this tendency to rigidly follow a pre-defined plan may underlie at least some of the difference in game optimality between pretend and non-pretend games. We find evidence for such a correlation ( $r = -.13$ , 95% CI  $[-.21, -.04]$ ,  $t(499) = -2.91, p = .004$ ). Note that the direction is negative because high optimality is mapped to low values, but high flexibility is mapped to high values.

#### Model-based analysis

We considered three factors that contribute to letter selections in Hangman:

1. The prior probability of a letter in the English language  $p(X)$ .
2. The posterior probability that a letter is present in the target word, given the game state  $S$  and the category  $p(X|S)$ .
3. The expected information gain this letter provides regarding the identity of the target word  $H(S) - \sum p(x_v)H(S, x_v)$  (where  $x_v$  is one potential outcome of guessing the letter  $X$ , for example, finding it in the second and fifth positions).

To quantify the contribution of each of these factors to letter selection, we fitted a multinomial logistic regression model to the data of each player. The model was specified and fit using the *Turing* Julia library for Bayesian inference with probabilistic programming (25). We compared the ensuing coefficients in ‘pretend’ and ‘non-pretend’ conditions.

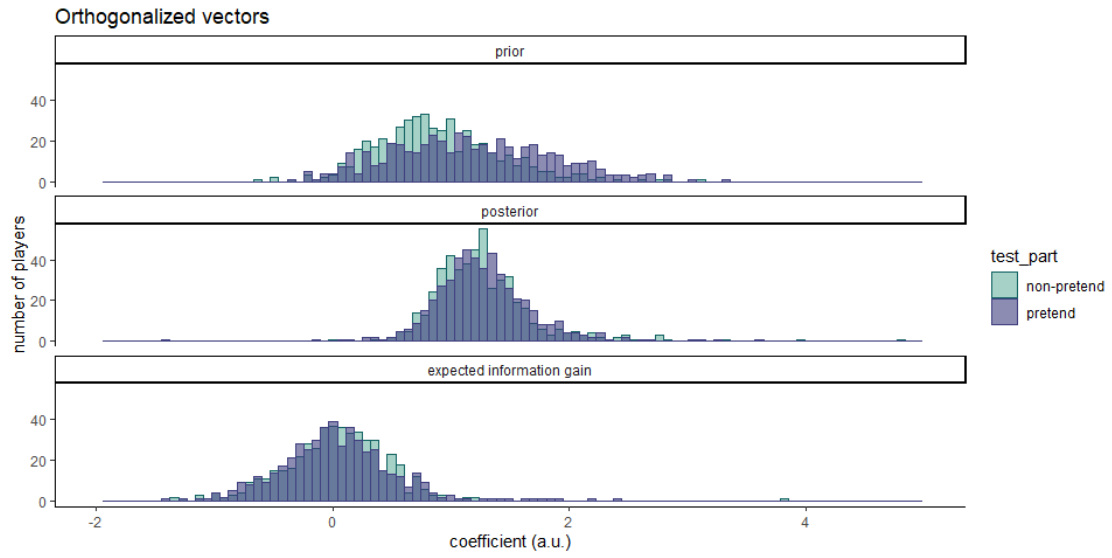


**Figure S10: Model coefficients in pretend and non-pretend Hangman games.**

Coefficients are derived from a multinomial regression model, trained to predict the next letter selection based on its prior probability, posterior probability given the game state, and expected information gain.

All three terms had a positive effect in both conditions. Prior probability (base rate) had a stronger effect in pretend games ( $M = 0.24$ , 95% CI  $[0.20, 0.27]$ ,  $t(500) = 12.34$ ,  $p < .001$ ). Hit posterior probability showed no such difference, and had a similar effect on players' decisions in both conditions ( $M = 0.01$ , 95% CI  $[-0.09, 0.10]$ ,  $t(500) = 0.13$ ,  $p = .900$ ). Finally, expected information gain had a weaker effect in pretend games ( $M = -0.29$ , 95% CI  $[-0.36, -0.21]$ ,  $t(500) = -7.62$ ,  $p < .001$ ).

Given collinearity between model predictors, we suspected that the effects of expected information gain may be driven by posterior probability effects. To isolate unique contributions of prior probability and expected information gain, we repeated the same analysis, but this time serially orthogonalizing the features (in the order [posterior, prior, expected information gain], and using the Gram Schmidt algorithm) prior to fitting the multinomial regression model. This time, expected information gain had no significant effect on decisions in non-pretend ( $t(500) = -0.32$ ,  $p = .750$ ) and pretend games ( $t(500) = 1.29$ ,  $p = .199$ ), rendering it possible that the previously observed effects were confounded with prior and posterior probability. In contrast, both prior and posterior probability had a significant positive effect, with prior probability having a stronger effect in pretend games ( $M = 0.26$ , 95% CI  $[0.18, 0.34]$ ,  $t(500) = 6.18$ ,  $p < .001$ ).



**Figure S11: Model coefficients in pretend and non-pretend Hangman games, after orthogonalization.**

### Suppression or simulation?

At the end of the experiment, players were asked whether they had a strategy for pretending or for detecting pretense. I (MM) manually labeled their responses according to three criteria: mentions of suppression (e.g., “I just tried to forget what the word was”), simulation (e.g., “I tried to focus on how I would have deduced the word if I hadn't known it”), and rules (e.g., “I started with vowels in the order we receive them in school”). Some answers got more than one label and some got none. Overall, 10 mentioned using suppression, 181 mentioned using simulation, and 328 mentioned following a heuristic.

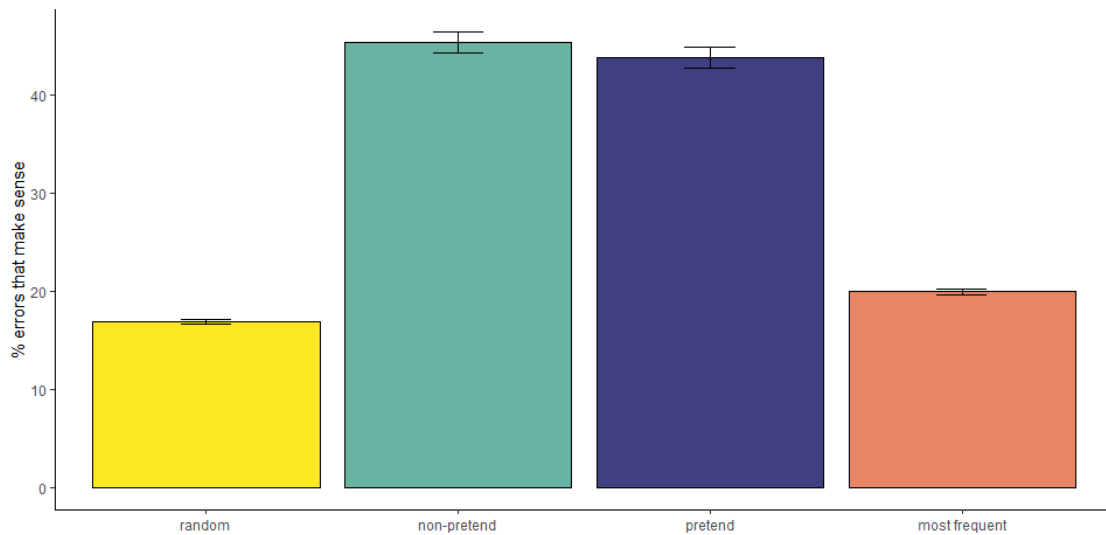
Since there were only 5 players who reported using suppression but not simulation, we focus our analysis only on the 176 who reported using simulation but no suppression.

**Click optimality** Among simulators, pretend games were significantly more optimal than random ( $M = -6.73$ , 95% CI  $[-7.01, -6.45]$ ,  $t(171) = -47.72$ ,  $p < .001$ ) even when restricting the analysis to letter selections that resulted in a miss ( $M = -2.41$ , 95% CI  $[-2.76, -2.06]$ ,  $t(168) = -13.66$ ,  $p < .001$ ). Like in the full sample, simulators were more optimal when playing normally than when pretending ( $M = 0.35$ , 95% CI  $[0.11, 0.59]$ ,  $t(171) = 2.85$ ,  $p = .005$ ), also when focusing on misses ( $M = 0.91$ , 95% CI  $[0.47, 1.36]$ ,  $t(168) = 4.07$ ,  $p < .001$ ).

**RT by outcome** The simulators group showed a similar temporal profile to that of the entire sample. Specifically, in pretend games, cell selections that resulted in a hit were faster than those that resulted in a miss ( $M = -1,268.01$ , 95% CI  $[-1,720.81, -815.20]$ ,  $t(167) = -5.53$ ,  $p < .001$ ), and cell selections that followed a hit were slower than those that followed a miss ( $M = 1,037.25$ , 95% CI  $[673.65, 1,400.84]$ ,  $t(168) = 5.63$ ,  $p < .001$ ).

***The letter frequency heuristic***

In this analysis, we asked whether pretenders are doing more than guessing frequent letters such as vowels. We therefore compared behaviour in pretend and non-pretend games to the behaviour of an agent that selects the most frequent letter among the ones that were not selected so far. By focusing our analysis on misses only (that is, selections of letters that are not in the target word), we ensured that any observed advantage in real players' behaviour is not simply due to making more correct guesses.



**Figure S12: Proportion of errors that make sense out of all errors.**

Our dependent measure was the proportion of guesses that “make sense”, that is, that are consistent with the board state and the word category. Pretend guesses made more sense according to this measure than what we would expect if people were only guessing the most frequent letters (difference in percentage of sensible errors between pretend and heuristic behaviour:  $M = 23.88$ , 95% CI [21.87,25.89],  $t(487) = 23.36$ ,  $p < .001$ ).

## Interaction between test part and condition

In both experiments, the order of pretend and non-pretend blocks was counterbalanced between participants. Here we ask whether differences between pretend and non-pretend effects are stronger in the first part of the experiment compared to the second part of the experiment. Overall, we find no significant interaction effects in any of our measures. Marginally significant effects appear for Hangman (a reduction of the effect of pretense on the number of misses, and on click optimality between the first and the second parts), and in Battleship (a reduction of the effect of pretense on the effect of guess accuracy on decision time between the first and second parts).

### Number of clicks

#### *Battleship:*

No interaction between test part (first/second) and condition (pretend/non-pretend) on number of clicks:  $\Delta M = -0.08$ , 95% CI  $[-0.41, 0.25]$ ,  $t(475.09) = -0.50$ ,  $p = .615$

#### *Hangman:*

No interaction between test part (first/second) and condition (pretend/non-pretend) on number of misses:  $\Delta M = -0.30$ , 95% CI  $[-0.61, 0.02]$ ,  $t(487.50) = -1.86$ ,  $p = .063$

A marginally significant effect was driven by a descriptively stronger tendency to underestimate the number of misses in the first part of the experiment (a contrast between pretend and non-pretend from the first part only:  $\Delta M = -1.18$ , 95% CI  $[-1.61, -0.74]$ ) compared to the second part of the experiment (a contrast between pretend and non-pretend from the second part only:  $\Delta M = -0.58$ , 95% CI  $[-0.96, -0.20]$ ).

### Correlation in number of guesses: Hangman

The Spearman correlation between the number of misses in pretend games and non-pretend games was 0.83 in the first part and 0.96 in the second part. The difference between the two correlations was not statistically significant in a Fisher z test ( $p > 0.19$ ).

**Click optimality**

***Battleship:***

No interaction between test part (first/second) and condition (pretend/non-pretend) on click

optimality:  $\Delta M = -0.05$ , 95% CI  $[-0.20, 0.11]$ ,  $t(475.52) = -0.57$ ,  $p = .568$

***Hangman:*** No interaction between test part (first/second) and condition (pretend/non-pretend) on click optimality:  $\Delta M = -0.16$ , 95% CI  $[-0.36, 0.03]$ ,  $t(488.02) = -1.66$ ,  $p = .097$

A marginally significant effect was driven by a descriptively bigger optimality cost for pretending in the first part of the experiment (a contrast between pretend and non-pretend from the first part only:  $\Delta M = -0.52$ , 95% CI  $[-0.77, -0.28]$ ) compared to the second part of the experiment (a contrast between pretend and non-pretend from the second part only:  $\Delta M = -0.20$ , 95% CI  $[-0.44, 0.05]$ ).

**Click optimality: misses only**

***Battleship:***

No interaction between test part (first/second) and condition (pretend/non-pretend) on click

optimality when restricting the analysis to misses only:  $\Delta M = -0.02$ , 95% CI  $[-0.23, 0.19]$ ,  $t(474.92) = -0.15$ ,  $p = .879$

***Hangman:*** No interaction between test part (first/second) and condition (pretend/non-pretend) on click optimality when restricting the analysis to misses only:  $\Delta M = 0.01$ , 95% CI  $[-0.26, 0.27]$ ,  $t(477.89) = 0.05$ ,  $p = .957$

**Effect of ship completion: Battleship**

No interaction between test part (first/second) and condition (pretend/non-pretend) on the effect of sinking the submarine on the probability of checking whether a size-2 patrol boat is a submarine:  $\Delta M = 0.02$ , 95% CI  $[-0.04, 0.07]$ ,  $t(319.44) = 0.58$ ,  $p = .563$

**Effect of guess accuracy on decision time**

***Battleship, previous guess:***

No interaction between test part (first/second) and condition (pretend/non-pretend) on the effect of the accuracy of the previous guess on the time taken to make the next one:  $\Delta M = -75.65$ , 95% CI  $[-155.04, 3.73]$ ,  $t(457.09) = -1.87$ ,  $p = .062$

A marginally significant effect was driven by a descriptively stronger effect in the first part of the experiment (a contrast between pretend and non-pretend from the first part only:  $\Delta M = -149.45$ , 95% CI  $[-256.31, -42.58]$ ) compared to the second part of the experiment ( $\Delta M = 1.86$ , 95% CI  $[-82.73, 86.44]$ ). Here, the effect in the second part of the experiment was not significant ( $t(473.52) = 0.04$ ,  $p = .966$ ).

***Battleship, current guess:*** No interaction between test part (first/second) and condition (pretend/non-pretend) on the difference in RT between accurate and inaccurate guesses:  $\Delta M = 55.96$ , 95% CI  $[-6.26, 118.19]$ ,  $t(452.94) = 1.77$ ,  $p = .078$

A marginally significant effect was driven by a descriptively stronger effect in the first part of the experiment (a contrast between pretend and non-pretend from the first part only:  $\Delta M = 339.40$ , 95% CI  $[249.78, 429.03]$ ) compared to the second part of the experiment ( $\Delta M = 227.48$ , 95% CI  $[156.33, 298.62]$ ).

***Hangman, previous guess:*** No interaction between test part (first/second) and condition (pretend/non-pretend) on the effect of the accuracy of the previous guess on the time taken to make the next one:  $\Delta M = 177.64$ , 95% CI  $[-401.56, 756.84]$ ,  $t(425.10) = 0.60$ ,  $p = .547$

***Hangman, current guess:*** No interaction between test part (first/second) and condition (pretend/non-pretend) on the difference in RT between accurate and inaccurate guesses:  $\Delta M = -299.17$ , 95% CI  $[-1,532.64, 934.31]$ ,  $t(420.51) = -0.48$ ,  $p = .634$

**Effect of uncertainty on decision time**  
***Battleship:***

No interaction between test part (first/second) and condition (pretend/non-pretend) on the quadratic effect of uncertainty on decision time:  $\Delta M = 89.37$ , 95% CI  $[-21.68, 200.43]$ ,  $t(386.83) = 1.58$ ,  $p = .114$

**Hangman:** No interaction between test part (first/second) and condition (pretend/non-pretend) on the quadratic effect of uncertainty on decision time:  $\Delta M = -90.17$ , 95% CI  $[-584.08, 403.74]$ ,  $t(496.63) = -0.36$ ,  $p = .720$

### Standard deviation of number of misses

#### **Battleship:**

No interaction between test part (first/second) and condition (pretend/non-pretend) on the inter-game variability in the number of misses:  $\Delta M = 0.05$ , 95% CI  $[-0.06, 0.17]$ ,  $t(497.40) = 0.90$ ,  $p = .369$

**Hangman:** No interaction between test part (first/second) and condition (pretend/non-pretend) on the inter-game variability in the number of misses:  $\Delta M = 0.10$ , 95% CI  $[-0.08, 0.27]$ ,  $t(493.93) = 1.09$ ,  $p = .276$

### Shannon entropy of decisions

#### **Battleship:**

No interaction between test part (first/second) and condition (pretend/non-pretend) on the Shannon entropy of participants' guesses (averaged across guesses 1-6):  $\Delta M = 0.03$ , 95% CI  $[-0.04, 0.10]$ ,  $t(458.56) = 0.95$ ,  $p = .344$

**Hangman:** No interaction between test part (first/second) and condition (pretend/non-pretend) on the Shannon entropy of participants' guesses (averaged across guesses 1-6):  $\Delta M = 0.04$ , 95% CI  $[-0.03, 0.11]$ ,  $t(472.67) = 1.19$ ,  $p = .235$