

Pretending Not to Know Reveals a Capacity for Model-Based Self-Simulation

Matan Mazor¹ , Chaz Firestone² , and Ian Phillips² 

¹Department of Experimental Psychology and All Souls College, University of Oxford, and

²Department of Psychological & Brain Sciences and Department of Philosophy, Johns Hopkins University

Psychological Science

1–14

© The Author(s) 2026



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/09567976251409747

www.psychologicalscience.org/PS

 Sage

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. English-speaking adults ($N = 1,001$) 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 the effects of decision uncertainty on decision time. Although peers were unable to detect pretense, statistical analysis and computational modeling uncovered traces of overacting 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

Received 9/1/24; revision accepted 11/25/25

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 et al., 2012; Harley et al., 2004) without this knowledge. Such biases remain potent even when one instructs

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, people are able to provide relatively accurate descriptions of their own decision-making (Morris et al., 2023) and perception (Levin & Angelone, 2008; Mazor et al., 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

Corresponding Author:

Matan Mazor, Department of Experimental Psychology and All Souls College, University of Oxford
 Email: matan.mazor@all-souls.ox.ac.uk

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 nonpretense 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 critical information in a game setting. Using online versions 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 normal version of the game (i.e., without pretense) as well as a pretend version in which they were given complete information about the hidden ships or the target word but were instructed to behave as if they did not have this information. Participants' pretense behavior mirrored broad patterns and subtle features of real players' decisions and decision times. At the same time, epistemic pretense was characterized by overacting, 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.

Research Transparency Statement

General disclosures

Conflicts of interest: The authors declare no conflicts of interest. **Funding:** This study was supported by a National Science Foundation (Division of Behavioral and Cognitive Sciences) Grant No. 2021053 awarded to C. Firestone. M. Mazor 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 disclosures (Battleship)

Preregistration: The research aims, hypotheses, methods, and analysis plan were preregistered (<https://osf.io/v9zsb>) prior to data collection, and they were time-locked using cryptographic randomization-based time-locking (Mazor et al., 2019). For the confirmatory analyses, there were no deviations from the preregistration. Additional exploratory (not preregistered) analyses are identified as such in the manuscript. **Materials:** All study materials,

including demos of analysis experiments and experiment code, are publicly available (<https://osf.io/zma9b> and <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>). **Computational reproducibility:** The computational reproducibility of the results in the main article (but not the Supplemental Material) has been independently confirmed by the journal's STAR team.

Experiment 2 disclosures (Hangman)

Preregistration: The research aims, hypotheses, methods, and analysis plan were preregistered (<https://osf.io/3thry>) prior to data collection, and they were time-locked using cryptographic randomization-based time-locking (Mazor et al., 2019). For the confirmatory analyses, there were no deviations from the preregistration. Additional exploratory (not preregistered) analyses are identified as such in the manuscript. **Materials:** All study materials, including demos of the experiments and experiment code, are publicly available (<https://osf.io/zma9b> and <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>). **Computational reproducibility:** The computational reproducibility of the results in the main article (but not the Supplemental Material) has been independently confirmed by the journal's STAR team.

Open Science Framework (OSF)

To ensure long-term preservation, we registered all OSF files at <https://osf.io/95txn>.

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 (Experiment 1), 500 English-speaking players (recruited from Prolific.com) were presented with a 5×5 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, 501 English-speaking players attempted to reveal a

hidden word, name, or number (hereafter referred to broadly as a word) with as few letter-guesses as possible, on the basis of word length and category (a famous person, number, fruit, U.S. state, or body part). To ensure familiarity with U.S. states, Hangman participants were all U.S.-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, the players' task was to behave as if they were playing for real—that is, 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, participants were told this:

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 behavior rather than simulating how their behavior 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 behavior from this practice game. To minimize such learning effects, we distinguished practice games from the main experimental blocks, using a smaller 4×4 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, in which 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 nonpretend games. Our full preregistered results are available online, together with the report-generating code. Unless otherwise specified, all reported findings similarly hold when we analyzed only the first condition performed by each participant in a between-subject analysis, thereby ensuring that findings are not driven by learning effects.¹ Readers are invited to try demos of the experiments self-model.github.io/pretendingNotToKnow/experiments/demos/pretend.

Results

Measuring pretense quality

As a first measure of pretense quality, we compared the total number of guesses in pretend and nonpretend games. Among Battleship players, the number of cell selections was similar in pretend ($M = 15.83$, $SD = 2.91$)

Experiment 1: Battleship

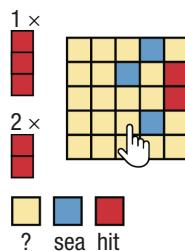
Randomized Order

a

Nonpretend Games

$\times 5$

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

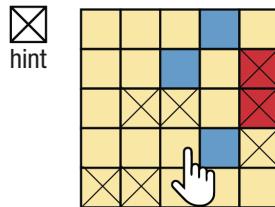


b

Pretend Games

$\times 5$

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.**

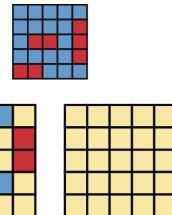


c

Judge Trials

$\times 5$

When you are ready to decide, click on the board of the player who had hints.



Experiment 2: Hangman

Randomized Order

d

Nonpretend Games

$\times 5$

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

$\times 5$

The next word is PIGEON, but **your task is to pretend you don't know that.**

Type PIGEON to confirm:

an animal

P	I	G	E	O	N
---	---	---	---	---	---

A B C D E F G H I
J K L M N O P Q R
S T U V W X Y Z

f

Judge Trials

$\times 5$

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

Fig. 1. Experimental design in Experiment 1 (upper panel) and 2 (lower panel). In nonpretend 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 we asked players to play as if they did not have this information. Last, players watched replays of the games of previous players and guessed which were pretend games (c and f).

and nonpretend games ($M = 16.05$, $SD = 2.18$, $t(499) = -1.43$, $p = .153$, Cohen's $d = 0.06$ (Fig. 2a)). Twenty pretenders who immediately discovered all ships without making errors were excluded from all further analyses, in accordance with our preregistered plan. With these subjects excluded, the number of cell selections remained very similar in pretend games ($M =$

16.11) and nonpretend games ($M = 15.94$, $t(479) = 1.39$, $p = .164$; Fig. 2a). In Hangman, pretenders tended to make about one additional letter guess on average than did nonpretenders, controlling for word length (pretend: 2.80 misses, $SD = 2.77$; nonpretend: 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

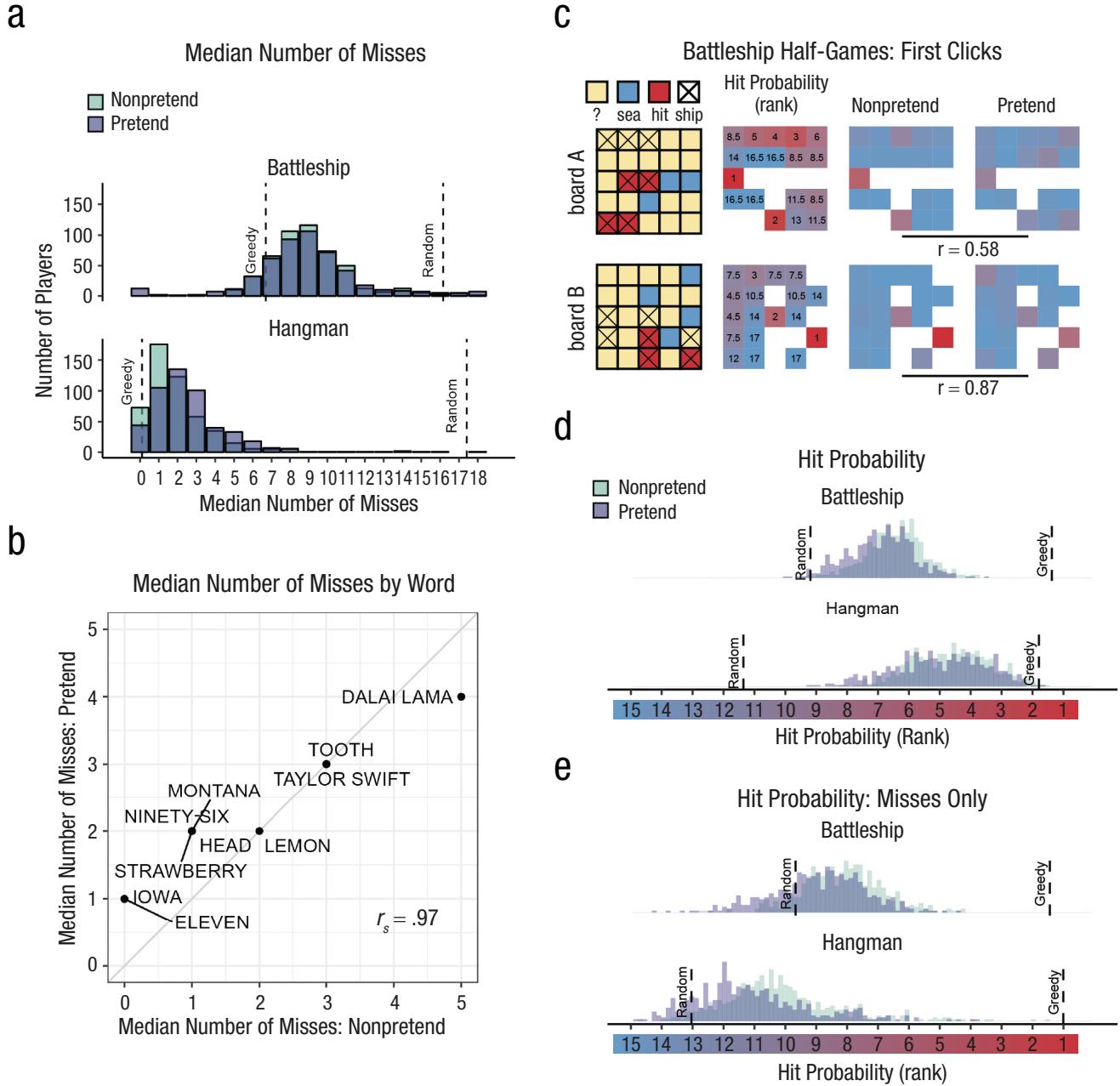


Fig. 2. Battleship and Hangman guesses in pretend and nonpretend games. In (a) we show the median number of misses in Battleship and Hangman games, in nonpretend (green) and pretend (purple) games. For reference, the expected number of misses is indicated by a reference line for a fully random agent and for a greedy agent that maximizes the probability of a hit in each step. In (b) we show the median number of misses in Hangman for pretend and nonpretend games, as a function of the target word. In (c) are illustrated spatial-guess distributions for pretend and nonpretend half-games (in which players continued the game from a half-finished state) alongside their corresponding hit-probability maps. In (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). The median rank per participant is plotted for pretend and nonpretend games, with reference lines for the expected rank probability for both a random agent and 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. We show results in (e) as in (d), except that all guesses that resulted in a hit were discarded.

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.

Having established an alignment in the total number of guesses, we next turned to the content of pretend and nonpretend guesses. In order to directly compare pretend and nonpretend 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 nonpretenders. 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 nonpretend guesses (board A: $r = .58$, $p = .012$; board B: $r = .87$, $p < .001$; see 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 nonpretend 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 et al., 2014). To this end, we ranked cells on the basis of 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 on the number of visits to Wikipedia entries for famous people. The full prior distributions for each category were included in the preregistration (for details, see the Supplemental Material available online). Similar to Battleship, in deriving hit probability we assumed 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 nonpretend 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 nonpretend guesses (Battleship: $t(479) = 11.04$, $p < .001$, Cohen’s $d = 0.50$; Hangman: $t(500) = -4.57$, $p < .001$, Cohen’s $d = 0.20$; see 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 $d = 0.86$; see 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’ behavior 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 nonpretenders, 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 whether 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 whether 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 nonpretend games. Like nonpretenders, pretenders were faster in their successful guesses (difference in decision time between hits and misses in Battleship: $\Delta_{\text{nonpretend}} = -109$ ms, $\Delta_{\text{pretend}} = -293$ ms; Hangman: $\Delta_{\text{nonpretend}} = -386$ ms, $\Delta_{\text{pretend}} = -297$ ms) and slowed down immediately after a hit (difference in decision time between guesses that followed hits vs. misses in Battleship: $\Delta_{\text{nonpretend}} = 182$ ms, $\Delta_{\text{pretend}} = 236$ ms; Hangman: $\Delta_{\text{nonpretend}} = 986$ ms, $\Delta_{\text{pretend}} = 667$ ms; Fig. 3a). All effects are significant at the 0.001 level with the preregistered within-subject t test, except for the posthit slowing down in Battleship, which, because of outliers with extreme effects in the opposite direction (> 10 s), was only significant in a nonparametric Wilcoxon

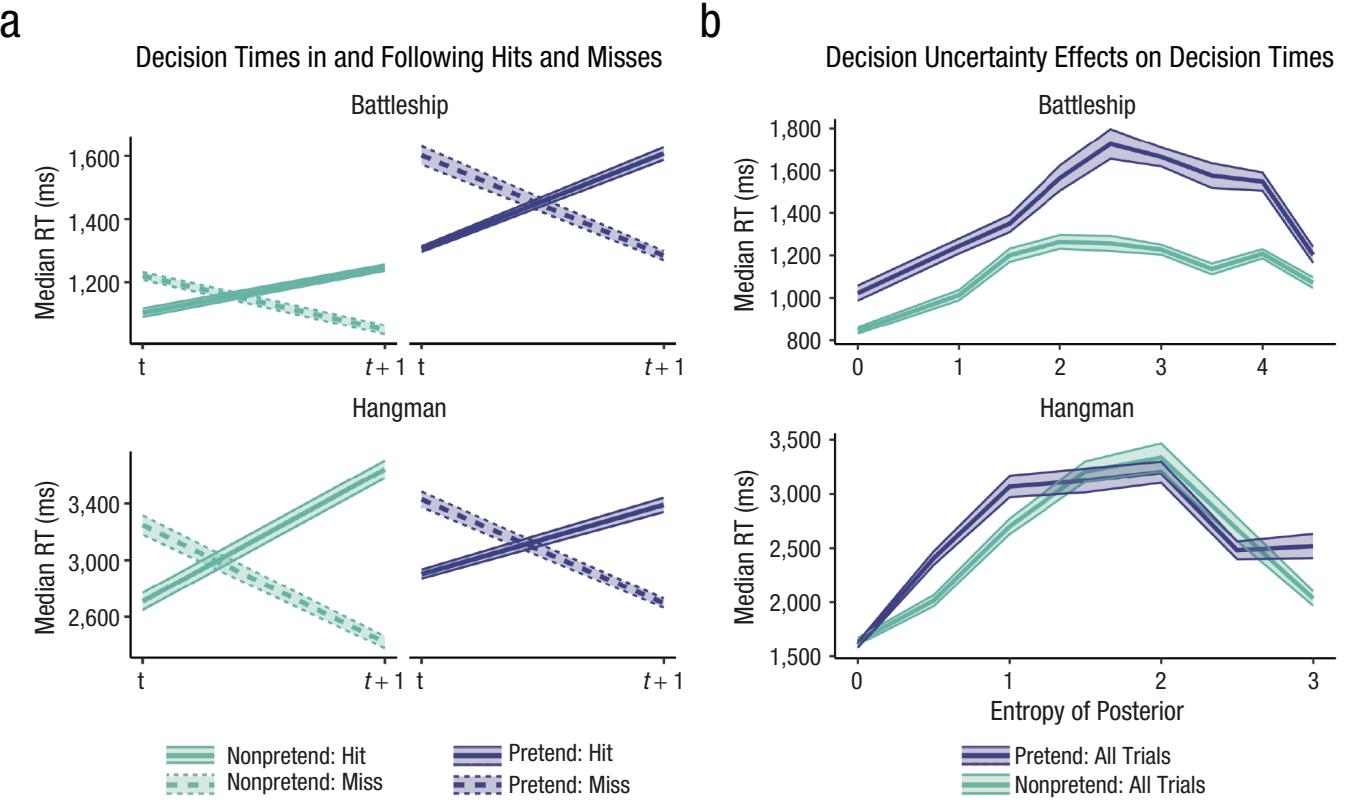


Fig. 3. Patterns of decision time in pretend and nonpretend games. In (a) we show median decision times for hits and misses, as well as the decisions following them. In both Battleship and Hangman, hits were faster on average than misses, but guesses following a hit were slower on average than those following a miss. This pattern was mimicked in pretend games. In (b) are 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 midrange levels of entropy, and this pattern was mimicked in pretend games. Shaded areas represent the bootstrapped standard errors of the median. RT = response time.

sign-rank test ($V = 87,876.50$, $p < .001$). Effects remained 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 response times from the linear and quadratic expansions of decision entropy, and contrasted the coefficients against zero in a group-level t test. In the nonpretend versions of the games, the quadratic coefficients were significantly negative, with the slowest responses associated with midrange levels of entropy (Battleship: $t(479) = -4.20$, $p < .001$, Cohen's $d = 0.19$; Hangman: $t(500) = -8.70$, $p < .001$, Cohen's $d = 0.39$; see Fig. 3b). When restricting the analysis to those Battleship players who pretended after playing normally, this effect was significant only in a Wilcoxon rank-sum test, because of outliers in the sample: $V = 3,892.00$, $p < .001$). Critically, the quadratic

coefficients were significantly negative also in pretend games (Battleship: $t(479) = -15.65$, $p < .001$, Cohen's $d = 0.71$; Hangman: $t(500) = -3.49$, $p = .001$, Cohen's $d = 0.16$; see Fig. 3b). In other words, despite knowing the game's solution with full certainty, pretenders successfully feigned subtle qualitative effects of counterfactual uncertainty on their decision times.

Stereotypical, imperfect self-simulation

Though impressive, the capacity for simulating a state of ignorance was not perfect. Importantly, the limitations and biases we observed 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 nonpretend games—a form of overacting. Specifically, the difference in response 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 nonpretend games. Despite a highly similar average number of misses in pretend and nonpretend games (Fig. 2a), the number of unsuccessful guesses was overwhelmingly less variable in pretend relative to nonpretend games (Battleship: $SD = 1.61$ in pretend vs. 2.60 in nonpretend games, $t(499) = -15.65$, $p < .001$, Cohen's $d = 0.70$; Hangman: $SD = 1.53$ in pretend vs. 2.65 in nonpretend games, $t(500) = -12.65$, $p < .001$, Cohen's $d = 0.56$; see Fig. 4a). Moreover, although 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 et al., 2021; Mazor & Fleming, 2021), with an attempt to avoid extreme outcomes to appear convincing to a hypothetical observer (Oey et al., 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 corresponds to a tendency to repeat the same sequence of guesses in different games. For example, if a player always starts games by clicking in the top left corner, their guess entropy for the first click will be $H([1,1,1,1]) = 0$. Unsurprisingly, the within-participant sequential guess entropy increased as a function of guess number, consistent with players adjusting their behavior 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 ($p < .001$ for a within-subject t test of guess entropy in guesses number 1 through 4 in both Battleship and Hangman; see the Supplemental Material for guess-specific statistics). Thus, when pretending, participants produced similar guess sequences across different games. In Hangman, for example, this meant that nonpretenders more flexibly adjusted their first-letter guess to the word category and number of letters than pretenders did; pretenders tended to open the game with the same letter guess regardless of the specific game state. This seems consistent with an attempt to enact what they saw as typical, representative, or average behavior. 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 nonpretend 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' wrongly calibrated 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 nonpretend games (and was even positive for Battleship: $r_s = .11$, $p = .019$; Hangman: $r_s = -.01$, $p = .832$). We interpret this as evidence that the reduction in variance reflects wrongly calibrated beliefs not only about randomness, but also about participants' behavior 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.) regardless of the game state, compared with genuine players (Fig. 4c). This suggests that in their attempt to behave as if they did not 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 the Supplemental

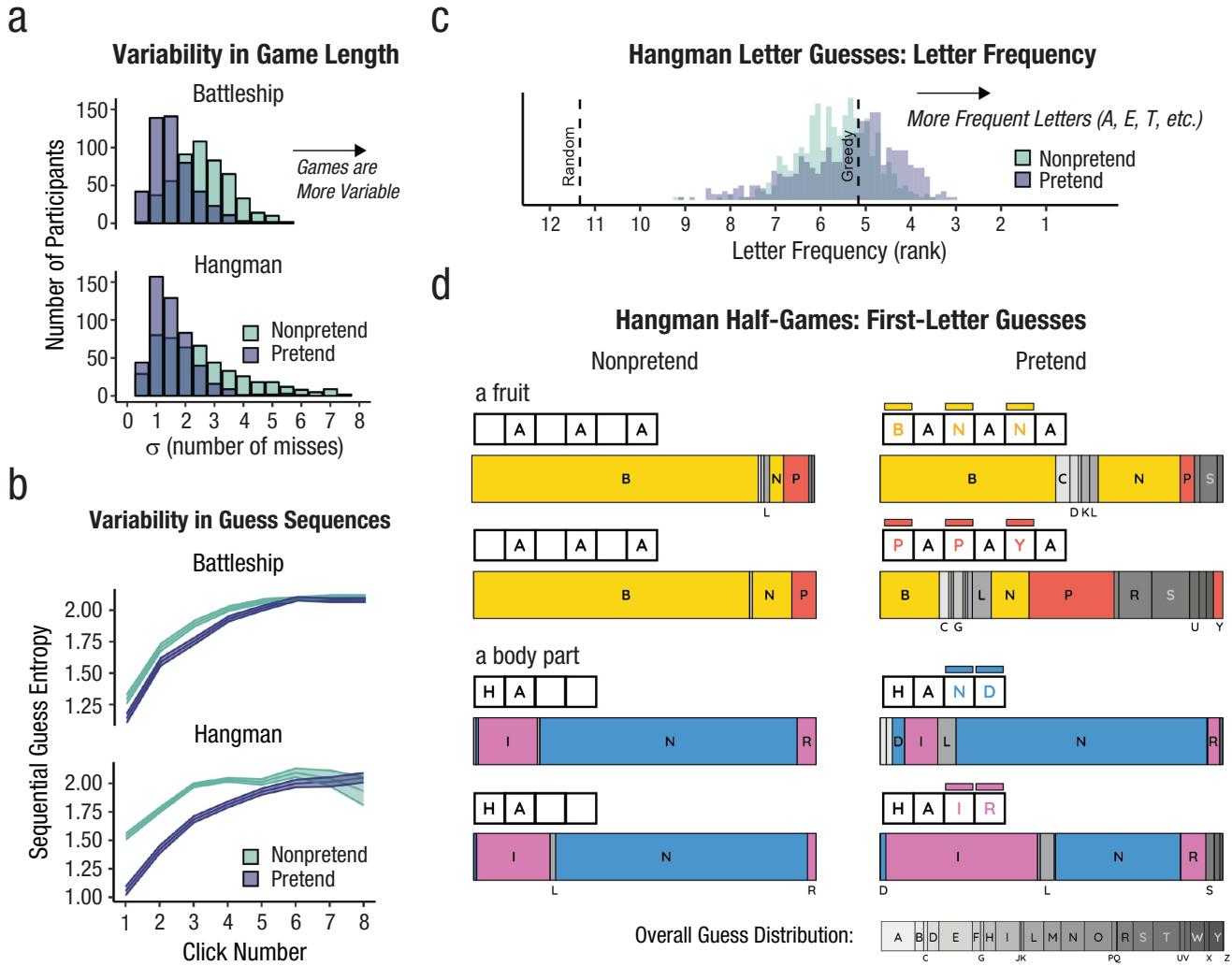


Fig. 4. Limitations on flexible decision-making when pretending. In (a) we show that variability in the number of misses (extracted individually for each player and then averaged) was lower in pretend games. Sequential guess entropy—a measure of the (inverse) predictability of individual players' guesses as a function of click number and guess number—is shown in (b). In both Battleship and Hangman, sequential guess entropy increased with click number, and was overall lower in pretend games. Shaded areas represent the mean ± 1 SE. Letter frequency of Hangman guesses is shown in (c)—the median rank frequency per participant in pretend and nonpretend 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. First-letter guesses in Hangman half-games are shown in (d) 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.

Material for evidence that heuristic use alone cannot fully explain pretenders' behavior).

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 nonpretenders 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: $p = .002$). Crucially, 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 were not consistent with either PAPAYA or BANANA (see the gray bars in Fig. 4d). This reveals that many pretenders had the knowledge that PAPAYA would be a hard fruit to guess and therefore avoided 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).

Failure to detect pretending

These systematic deviations of pretend games from behavior under true ignorance made pretend games discernible from nonpretend games. To illustrate, a support vector machine (SVM) algorithm reached an accuracy level of 72% (Battleship, two-alternative forced choice) and 60% (Hangman, yes/no) in linearly classifying condition (pretend/nonpretend) on the basis of 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 nonpretend 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 the 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 that game 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 nonpretend 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 nonpretend behavior, 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; Roes & Vohs, 2012; Wood, 1978). In some cases, attempts to suppress thoughts even give rise to the paradoxical enhancement of suppressed representations (Earp et al., 2013; Giuliano & Wicha, 2010; Wegner et al., 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 reproduced 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 et al., 2009, 2017; Richardson & Keil, 2022) or implicitly (Liuet et al., 2017; Onishi & Baillargeon, 2005). Here we have proposed a complementary approach: Asking participants to generate behavior on the basis of 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 asked participants to run

the model forward, taking counterfactual beliefs and desires as input and producing behavior as output.

Because of the unconstrained space of possible behaviors in our task (cell selections \times decision latencies), successfully pretending not to know demands a rich model of cognition and is much harder to achieve on the basis of a quasienteristic theory of mental states (Gopnik & Wellman, 1994). Consequently, our findings support a simulation model of epistemic pretense, and perhaps of mentalizing more generally. Critically, however, unlike classic self-simulation accounts of mind-reading and theory of mind (Gallese & Goldman, 1998; 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 (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 on-line fashion as well as their consistent biases and limitations relative to behavior when in a true state of ignorance (Saxe, 2005).

We interpret participants' success in emulating a state of ignorance as revealing a nontrivial 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 nonpretend 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, in which the distribution of generated samples is often narrower than the distribution of training data (an effect known as *mode collapse*; Kossale et al., 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, overacted 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 in such a way that their entire cognitive machinery was available to play the game. This would

not require self-simulation but rather a capacity to intentionally "unsee," or forget, relevant evidence. Although 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 behavior 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 in a debrief question how they had performed the task, a significant majority of participants gave responses aligned with self-simulation or rule-following, and our main findings hold when we excluded the 32 Battleship and 10 Hangman players who mentioned suppression in response to this question (see the exploratory analysis).

Findings from Battleship and Hangman mostly aligned: For both environments, the median number of guesses was similar in pretend and nonpretend games, guesses (correct and incorrect) made sense within the context of the game, and response times were similarly sensitive to guess accuracy and uncertainty. We also observed a similar tendency to produce representative and stereotypical behavior 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 with 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 behavior 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 et al., 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 behavior (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 et al., 2003). Others have focused on the interaction between liars and recipients, modeling 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 behavior ("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"). Consequently, although 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 nontrivial 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 have shown 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 do not 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.

Transparency

Action Editor: Clayton R. Critcher

Editor: Simine Vazire

Author Contributions

Matan Mazor: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Resources; Software; Validation; Visualization; Writing – original draft; Writing – review & editing.

Chaz Firestone: Conceptualization; Funding acquisition; Investigation; Methodology; Resources; Writing – review & editing.

Ian Phillips: Conceptualization; Investigation; Methodology; Writing – review & editing.

Declaration of Conflicting Interests

The authors declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This study was supported by a National Science Foundation (Division of Behavioural and Cognitive Sciences) Grant No. 2021053 awarded to C. Firestone. M. Mazor 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.

Open Practices

Open practices for this article are described in the Research Transparency Statement section, which appears at the end of the Introduction section in the main text.

Open Science Framework (OSF)

To ensure long-term preservation, we registered all OSF files at <https://osf.io/95txn>.

ORCID iDs

Matan Mazor  <https://orcid.org/0000-0002-3601-0644>
 Chaz Firestone  <https://orcid.org/0000-0002-1247-2422>
 Ian Phillips  <https://orcid.org/0000-0003-2932-8045>

Supplemental Material

Additional supporting information can be found at <http://journals.sagepub.com/doi/suppl/10.1177/09567976251409747>

Note

1. In both experiments, the order of pretend and nonpretend blocks was counterbalanced between participants. We observed no significant interaction between the strength of any of our effects (i.e., differences between pretend and nonpretend conditions) and the part of the experiment (i.e., first vs. second). For full details, see the Supplemental Material.

References

- Allen, K., Brändle, F., Botvinick, M., Fan, J., Gershman, S. J., Gopnik, A., Griffiths, T. L., Hartshorne, J. K., Hauser, T. U., Ho, M. K., de Leeuw, J. R., Ma, W. J., Murayama, K., Nelson, J. D., van Opheusden, B., Pouncy, T., Rafner, J., Rahwan, I., Rutledge, R. B., . . . Schulz, E. (2024). Using games to understand the mind. *Nature Human Behaviour*, 8, 1035–1043. <https://doi.org/10.1038/s41562-024-01878-9>
- Audinot, M., Bonnet, F., & Viennot, S. (2014). *Optimal strategies against a random opponent in Battleship*. The 19th Game Programming Workshop. <https://cir.nii.ac.jp/crid/1050292572119748352>
- 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, Article 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
- 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, 62–74.
- Gopnik, A., & Wellman, H. M. (1994). The theory theory. In Hirschfeld, L. A., & Gelman, S. A. (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 257–290). Cambridge University 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>
- Gordon, R. M. (2007). *Moorean pretense*. Clarendon Press.
- Graziano, M. S., & Webb, T. W. (2015). The attention schema theory: A mechanistic account of subjective awareness. *Frontiers in Psychology*, 6, Article 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 Dunning-Kruger 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). Mode collapse in generative adversarial networks: An overview. In *2022 8th International Conference on Optimization and Applications (ICOA)* (pp. 1–6). IEEE. <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*, 152(7), 1951–1966. <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. In P. Carruthers & P. K. Smith (Eds.), *Theories of theories of mind* (pp. 90–104). 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, Article 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.
- Verschueren, B., Lin, C.-C., Huismann, S., Kleinberg, B., Willemse, M., Mei, E. C. J., van Goor, T., Löwy, L. H. S., Appiah, O. K., & Meijer, E. (2023). The use-the-best heuristic facilitates deception detection. *Nature Human Behaviour*, 7(5), 718–728. 1–11. <https://doi.org/10.1038/s41562-023-01556-2>
- Walczuk, 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 possibility—a 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), 5–13.
- 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>