The Limits of Social Cognition: Production Functions and Reasoning in Strategic Interactions

Vered Kurtz-David[1], Adam Brandenburger[2,3,4], and Paul W. Glimcher[1,5]

[1] Grossman School of Medicine, New York University

[2] Stern School of Business, New York University

[3] Tandon School of Engineering, New York University

[4] NYU Shanghai

[5] Center for Neural Science, New York University

Abstract

Classical game theory assumes that players reason their way to Nash Equilibrium. This assumption has been challenged by behavioral approaches, which recognize that individuals face cognitive constraints, limiting their ability to achieve equilibria. Here, we introduce a new measure of a game-complexity, which decomposes each interaction into social and non-social arithmetic cognitive demands. Inspired by the economic concept of production functions, we develop a psychophysical approach that models sophistication as the product of subjects' capacities on each of these dimensions. In two independent studies, we show that social and arithmetic demands are contextual factors for sophistication that behave lawfully with psychophysical regularity, and that subjects trade-off these capacities as game-complexity varies. Our results are a hybrid, applying concepts from psychophysics and individual decision-making into strategic social reasoning. These findings present a new approach to behavioral game theory, and provide a framework for future neuroimaging and computational psychiatric studies.

Main text

In Conan Doyle's *Adventure of the Final Problem*, Sherlock Holmes must escape his nemesis, Professor Moriarty, after boarding a train to Dover[1]. Holmes can continue to Dover as planned, in hopes of escaping to the Continent, or, anticipating that Moriarty will entrap him at Dover, exit the train at Canterbury. Moriarty, also a strategic thinker, might double-cross Holmes and wait at Canterbury to kill him. Or, Holmes might choose the triple-cross and continue to Dover after all. Classical game theory cuts through such chains of reasoning by supposing that the players arrive at a *Nash equilibrium*[2], where each player makes an optimal choice that takes into account full knowledge of the other players' equally optimal strategies.

In recent decades, alternatives to this equilibrium theory have arisen — principally, level-k theory[3–5] and epistemic game theory[6,7]— that embrace the cognitive limitations of real human players in strategic interactions. These new theories allow for more realistic assumptions about how many steps (*levels of reasoning*) players can engage in when thinking about other players' internal reasoning. They assume that humans are often non-rational players who choose at random. One-step-rational players choose strategies that are arithmetically optimal, but without any understanding of the reasoning of others. Two-step-rational players, in turn, choose optimally in their arithmetic and social reasoning with respect to these one-step-rational players, but remain unaware of players with even deeper levels of understanding. And so on. Early experiments concluded that players were often limited to two or fewer steps in their social reasoning[5,8–12]. Recently, Kneeland (2015)[13] has presented a new methodology involving what are called *Ring Games*, which has allowed identification of reasoning under more plausible assumptions, and found that many players reason up to three or four steps, a finding similar to qualitative conclusions from social psychology[14,15]. It is tempting to conclude that classical game theory could be augmented by a newer theory that ascribes to each player a maximum level of reasoning that is characteristic of that player — what we call that player's *strategic sophistication*[16,17]. However, recent work by Alaoui and Penta (2016)[18] suggests that sophistication is significantly context-dependent, or endogenous in the language of economics. For a given cognitive ability, these experiments suggest that behaviorally-expressed sophistication varies with incentives and with the perceived sophistication of the other players. Other studies have found that strategic sophistication depends on incentives[19,20], identities[10,11,19,21], training[19], inattention[22], and the type of game being played[23,23,24,24,25].

In this paper, we introduce a new measure of the complexity of a strategic social interaction. Within the same class of games, we find that strategic sophistication changes in a lawful fashion as the complexity of the social interaction and the arithmetic problem increase. Traditional measures of complexity have included game-tree complexity[26] and state-space complexity[27], or have focused on complexity of solution concepts[28]. Our measure of complexity is inspired by neuroscience and cognitive psychology and consists of two components. The first component, in the spirit of Theory of Mind[29,30], captures the social demands of the task. We refer to this as the *social-complexity* of the task. The second component captures the non-social generalized cognitive demand imposed by the task. We refer to this as the *arithmetic-complexity* of the task. Following standard cognitive theory, we posit that for each level of reasoning in which a player engages, a human agent incurs a cognitive cost that is increasing in both the social-complexity and the arithmetic-complexity of the environment. We allow different agents to have different innate capacities in each of these domains (different *types* in the jargon of game-theory) and we also allow for the possibility that these capacities trade-off against each other in an individual. From these assumptions, we build a function which describes the chance that a player will express (or choose) a given strategy that corresponds to the maximum available levels of reasoning, where the chance of selecting that strategy decreases as the complexity of the interaction increases in both dimensions.

We leverage the merits of the *Ring Game* methodology and test our framework in two experiments. In the first, we present subjects with an array of strategic interactions, which vary stepwise in their social- and arithmetic-complexity. In the second, we manipulate the cognitive resources available to subjects by varying processing time. We find that social- and arithmetic-complexity are fundamental contextual variables for describing social interactions and we provide a simple mathematical formulation that describes the observed psychophysical trade-offs using an economic production function. We note that in spirit our work aligns closely with the production-function approach to cognitive capacity developed by Gabaix and Graeber (2023)[31].

Our work opens the door to a systematic study of the effect of complexity on strategic reasoning. While increased social-complexity prompts players to reason more deeply about the reasoning of others, we find that there is also an increased cognitive burden for subjects as social-complexity rises. Turning to arithmetic-complexity, we also find, perhaps less surprisingly, an inverse relationship, where the probability that subjects express strategic sophistication goes down as arithmetic-complexity rises.

We further find that most subjects can trade-off their allocation of cognitive resources across these two complexity dimensions to some degree, in a way that varies from individual to individual. We also show that the average sophistication level of a subject varies logarithmically with processing time. Lastly, we find that some aspects of the performance of an individual subject can be related to their performance on a battery of well-validated psychological tests.

Our experimental design and data analysis thus represent an academic hybrid. We apply concepts from individual decision making[32,33] and psychophysics[34] to strategic choice in games. We use an economic formalism to develop a well-calibrated psychophysical function that relates graduated changes in the game structure to a dynamic range in subjects' sophistication.

The conventional game-theoretic description of a social interaction involves a payoff matrix or game tree. Modern level-k theory and epistemic game theory add to this description with models of the levels of social reasoning engaged in by individual players. In this paper, we add a further layer of description, namely, a two-dimensional measure of complexity of the game. With this new feature we are able to relate levels of reasoning to complexity in two cognitively distinct domains. Our approach of including additional endogeneity in the description of a decision maker renders the behavior of the players both analyzable and, to some extent, predictable, in a way that has escaped previous approaches. We see our work as opening a door to the development of a robust neuroeconomic theory of reasoning in games, which is corroborated by more traditional psychological measures. Finally, our novel design is tailored for future neuroimaging work aimed at a much more fine-grained understanding of the neural representation of cognitive capacities and levels of reasoning in games[35–39].

Results

Task

We exploited the *Ring Game*, developed by Kneeland (2015)[13], to investigate the utilization of cognitive resources in strategic interactions (games). In the original game, each player's payoff depends on their own choice and on the choice of the player sitting to their right, whose payoff in turn depends on the choice of the player to their right, and so on around in a circle (Fig. 1a, upper row). In the 3-person ring depicted in Fig. 1a, Ann's payoffs depend on Bob's choices, whose payoffs depend on Charlie's choices, whose payoffs depend in turn on

Ann's choices. If, for example, Bob chooses action *c*, and Charlie chooses action *e*, then Bob will end up receiving $15. If Bob chooses action *d*, and Charlie chooses action *f*, then Bob will receive $5.

A player's degree of overall *strategic sophistication* (or, the number of *levels of reasoning*) refers to how far around the ring, from their position, a subject reasons about the behavior of other players. In Fig. 1a (middle), if a subject is *L*3 (level 3), then, when playing the role of Ann, they will reason about Bob's reasoning about Charlie's reasoning about themselves (Ann). Starting with Charlie, Ann can infer that Charlie will choose e, since this ensures Charlie a larger payoff regardless of Ann's choice (the choice e is "dominant" in the language of game theory). Ann then imputes this same reasoning to Bob, and infers that Bob will choose c, since this maximizes Bob's payoff when Charlie chooses e. Given that Bob chooses c, Ann will optimally choose a. Such a player is labeled *L*3, since the player infers all social and arithmetic dependencies around the full ring of three players. A subject with a lower degree of strategic sophistication, but who otherwise evaluates the ring correctly on the arithmetic dimension, neglects one or more level of reasoning. For example, an *L1* subject in the role of Ann does not even reason about Bob's reasoning. Likewise, we demonstrate – via our measure for cognitive complexity of the arithmetic dimension – that a subject unable to accurately reason arithmetically, but who accurately evaluates the social dependencies, also effectively neglects one or more levels of reasoning.

To measure how subjects handle interactions with increasing social-complexity, we systematically manipulated the number of players in the ring (Fig. 1b). At the same time, in our variant of the *Ring Game*, we also varied the number of choices faced by each player, to yield increasing arithmetic-complexity (Fig. 1c). By independently varying ring size (social-complexity) and matrix size (arithmetic-complexity), we created nine different *game types* (Fig. 1d). In the rings depicted in Fig 1b, there are four players, with two choices available to each player. In Fig. 1c, there are two players in a ring with four choices available to each player.

For each game type, we then generated three different versions of a given ring structure, by varying the expected value of the "dominant" strategy ($9, $12, or $18), for a total of 27 specific games (see Fig. 1d, Fig. 7a, and Methods). These experimental manipulations yielded 27 different ring types with increasing social- and arithmetic-complexities. We then modeled the effect of complexity on subjects' levels of reasoning (see *Model* section below).

On each trial, we revealed each matrix in the ring, for five seconds each, in order from right to left. After all matrices had been revealed, subjects were told which role they were playing (i.e., which matrix encoded their personal payoffs) in that round (Fig. 1e). After learning their role in a specific round, subjects had only five seconds to submit their choice – encouraging them to fully process the social information carried in each game before their role was revealed, as in the Kneeland (2015)[13] design. Subjects played all the roles in all ring types, randomized across subjects. Again following Kneeland (2015)[13], we built the *exclusion restriction* (*ER*) criterion into our design, to avoid misidentification of types (Fig. 1a bottom and Methods). Specifically, each ring was played twice, with the rows of the rightmost matrix flipped, so that the dominant strategy also flipped. The ER assumes that only players with sufficient strategic sophistication will adjust their behavior across these two scenarios, effectively controlling for lucky guesses.

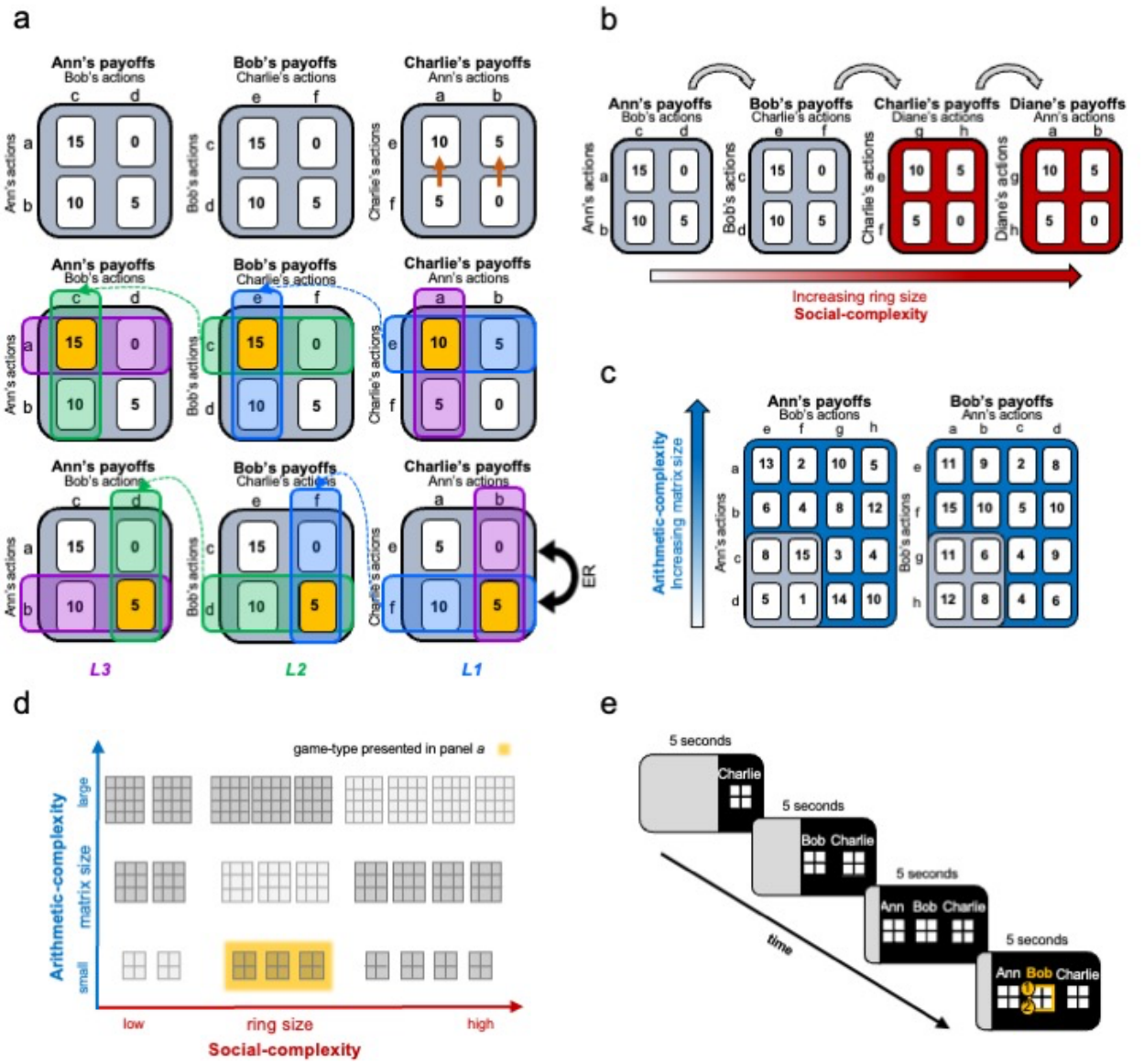


Fig. 1 | Design, Exp. 1. (a) *Top*: The *Ring Game*. Subjects played the *Ring Game*, a series of *strategic-form* (matrix) games, where the payoff to each player depends on the choice of the player to the right, and the payoff to the last (rightmost) player depends on the choice of the first payer, creating a ring. *Middle*: Player types. Rings are structured such that only the last (rightmost) player has a clearly best (dominant) strategy. An *L*1 player in the last position (Charlie) would choose the dominant action *e*. An *L*2 player in Bob's position would choose action *c*, since this is optimal when Charlie chooses e. An *L*3 player in Ann's position would choose action a, since this is optimal when Bob chooses c. *Bottom*: To identify a subject's type, we let subjects play each game twice, while flipping the rows for the last player (Charlie). The exclusion restriction[13] states that the flip will change Bob's choice only if he is a true *L*2 player. Similarly, the flip will change Ann's choice only if she is a true *L*3 player. (b) By adding more players to the ring, we increased the social-complexity of the task. (c) By adding more rows and columns to each matrix, we increased the arithmetic-complexity of the task. (d) We designed 27 rings from nine different game types that varied by the number of players in each ring (social-complexity), as well as by the number of alternatives in each matrix (arithmetic-complexity). The yellow shaded ring is the game type presented in panel *a*. (e) In each trial, game matrices were presented in order from the right to the left, with each matrix shown for five seconds. Subjects were then told which role they were playing on that round. Subjects played all the roles in all the different rings in random order for a total of 162 trials.

Model

To model the cognitive demand created by a specific game, we introduce the concept, adapted from economics, of a *cognitive production function* into psychophysics. (See Gabaix and Graeber (2023)[31] for a similar concept.) In our model, each strategic scenario faced by the subjects was described by its *social-complexity* (denoted by $m = 1, 2, ...$) and *arithmetic-complexity* (denoted by $n = 1, 2, ...$). The social-complexity m of a given ring is the total number of players in the ring, which is equal to the maximal number of iterations (levels of reasoning) needed for all players to identify their optimal strategies. The arithmetic-complexity n of a given ring is equal to the number of choice alternatives that each player faces. With each increment in m, players face the task of reasoning one more level about the reasoning of others required to understand the game. Likewise, each increment in n increases the cognitive demand on players required to identify the optimal alternative for each player in the ring. We assume that for each additional level of reasoning in which a player engages, there is a cognitive cost (effort) that is increasing in both m and n. On this basis, and by analogy with the economic concept of a Cobb-Douglas production function[31,40], we extend classical psychophysics by positing a *cognitive production function*[31] for each player that takes m and n as inputs (independent variables). The function outputs, for a given player, the probability that they will reason accurately to the maximum number of levels available given their position in the ring. Formally:

$$(1)\ \Pr(l_k) = C\left(\frac{1}{m}\right)^{\alpha}\left(\frac{1}{n}\right)^{1-\alpha},$$

where $\Pr(.)$ denotes probability, l_k is the maximum number of levels of reasoning available for a player in position k in the ring (measured from the rightmost position), and $C \geq 0$ and $0 \leq \alpha \leq 1$ are subject-specific constants. The constant C captures the idea of an overall cognitive capacity for each player, where a higher C increases the probability that the subject will reason the maximum number of levels. The exponent α and its complement $1 - \alpha$ capture the relative social and arithmetic-reasoning capabilities of the player. Formally, α is the elasticity (the ease of trade-off) of the inverse of social-complexity. It describes the percentage change in the probability of maximal reasoning divided by the percentage change in the number of levels about the reasoning of others required to understand the game. Similarly, the exponent $1 - \alpha$ is the elasticity of the inverse of arithmetic-complexity, with a parallel interpretation. (The general form of the Cobb-Douglas production function used in economics involves a different exponent α and β for each of two inputs – often taken to be "capital" and "labor," respectively. We impose the restriction $\beta = 1 - \alpha$ to limit degrees of freedom. Future psychophysical or economic models could relax this simplification).

For fixed values of C and α, we can calculate, for different probabilities of maximal reasoning $\Pr(l_k)$, the set of (m, n) pairs that satisfy Equation (1). This yields *iso-probability* curves (Fig. 1d, Fig. 2a), where the probability is decreasing as we move from one curve to another in a northeast direction. We can also depict the effect of increasing overall cognitive capacity C, which moves all iso-probability curves to the northeast (Fig. 2b). A value $\alpha > 1/2$ for the exponent indicates that the cognitive load of social-complexity weighs more heavily on the player than does arithmetic-complexity (indicating an *arithmetic-complexity orientation* of the player), while a value $\alpha < 1/2$ indicates the opposite case (*social-complexity orientation*); see Figs. 2c and 2d.

Beyond the previously mentioned exclusion restriction (ER) used to identify player types, the second identification challenge faced in inferring levels of reasoning in a game comes from a confounding effect when the ring size is varied. An increase in ring size m (social-complexity) necessarily also increases the overall cognitive load on a subject, as modeled by the right-hand side of Equation (1). But, at the same time, it motivates a subject to reason additional levels to arrive at the optimal choice. We wish to isolate the first effect (i.e., the overall cognitive load) and we achieve this by choosing as our left-hand variable in Equation (1) the probability of reasoning the maximum number of levels, at a given position, rather than some other measure, such as the number of levels reasoned, that would reflect both load and motivation effects simultaneously. By contrast, there is no confounding effect when we increase the number of options, m (arithmetic-complexity). However, if a subject chooses at random, there is a diminishing probability that the option corresponding to the maximum number of levels will be selected. We control for this effect empirically (see Performance Index and Methods).

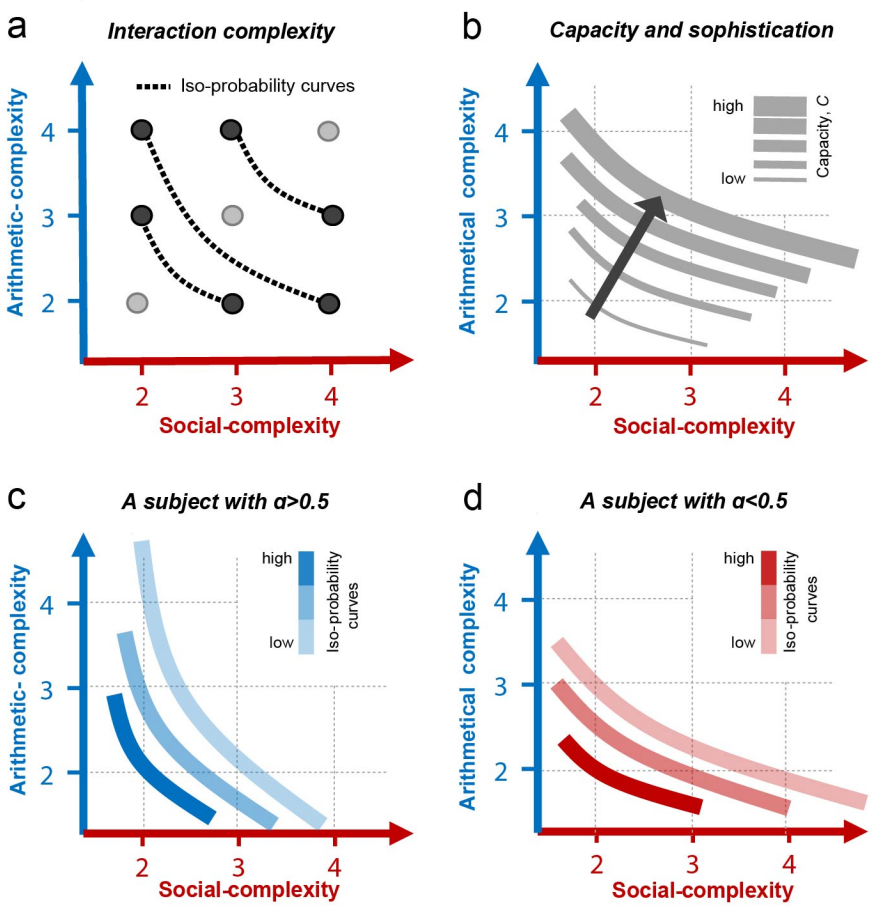


Fig. 2 | Model. (a) Each strategic interaction exhibits both *social-complexity* and *arithmetic-complexity*. The two sources of complexity combine to determine the overall cognitive demand of the task. Here, we plot iso-probability curves that depict pairs of complexity numbers that yield the same probability that the subject reasons to the maximum number of levels (Equation (1) in the text). The probability decreases as we move from one curve to the next in a northeast direction. Light gray points indicate strategic interactions tested in our task, which lie outside of the iso-probability curves. (b) C captures the idea of an overall capacity, which may vary, depending on the subject's cognitive abilities and the game environment. A higher value of C shifts a given iso-probability curve to the northeast, in the direction of the thicker gray curves. (c-d) We expect the probability $\Pr(l_k)$ to decrease as overall complexity increases. More opaque curves depict higher probability. Subjects differ according to the effect of the two sources of complexity on their behavior. (c) A subject with $\alpha > 1/2$ experiences a greater cognitive load from social-complexity- vs. arithmetic-complexity, yielding a higher probability $\Pr(l_k)$ when choice complexity is relatively higher. (d) A subject with $\alpha < 1/2$ experiences a greater cognitive load from choice complexity vs. social-complexity, yielding a higher probability $\Pr(l_k)$ when social-complexity is relatively higher.

Results from Experiment 1

We collected data from two populations, one population was composed of students at a highly-selective US University (NYU sample, N=54) and another was recruited from the general population in New York City via CraigsList (CL sample, n=55). Overall, subjects evidenced no difficulty understanding and completing the task even though subjects had a rather short processing time per decision problem. The average share of missed trials in the CL sample was 4.12% (min=0, max=35.8%, std=5.61%), and the average response time across subjects was 2.494 sec (min=0.588, max=3.920, std=0.7165). In the NYU sample 2.68% (min=0, max=21.60%, std=3.50%) of trials were missed and average response times were 2.134 sec (min=1.203, max=3.745, std=0.561; Supplementary Fig. 1).

Averaging across subjects and game types, choices matched l_k in 56.13% (min=22.84%, max=97.53%, std=19.48%) of the trials in the CL sample, and in 76.00% (min=31.48%, max=98.77%, std=19.19%) in the NYU sample. We find evidence for a slow and moderate learning effect throughout the experiment, as for each additional trial, the chances of subjects' responses to match l_k increased by 0.0264% in the CL sample (β=0.00070, p<0.0001 in a probit model clustered by subjects), and by 0.0762% in the NYU sample (β=0.0024, p<0.0001, respectively).

Levels of Reasoning

As a first step towards understanding how various complexity levels along the social-arithmetic grid (Fig. 1d) affected subjects' reasoning in the game, we identified each subject's discrete level of reasoning in each of the 27 rings (see Methods for the identification strategy). Subjects were observed to vary between L0 (random choice) and L4 (perfect accounting for the optimal behavior of all others), depending on ring size. Fig. 3a presents a histogram of the median level of reasoning produced by each subject as a function of ring size, for each of our populations. Across the different ring sizes, 41.82-74.55% of subjects in the CL population and 78.18-96.36% of subjects in the NYU population are classified as L1 or higher. In the CL sample, subjects are rather evenly distributed across the different levels of reasoning, suggesting a heterogeneity of reasoning capabilities (or *types* in economic jargon). In contrast, the NYU sample is skewed at L4, suggesting that some subjects are capable of iterating additional levels of reasoning (beyond L4), a possibility that cannot be directly assessed with the four layers of social reasoning we employed (Fig. 3a). In rings with similar structure to the ring that Kneeland tested in her study[13] (e.g. 4-person 3*3-matrix rings), 36.36% of subjects in the

CL sample and 78.18% of subjects in the NYU sample are classified as L1 or above. This proportion is somewhat lower than the one found by Kneeland (93%)[13], perhaps due to the heterogeneity in our subject pool (CL sample) or due to the greater difficulty in our task (27 rings compared to 1 with shorter processing time). Nonetheless, this proportion is still large enough to indicate that subjects were able to engage in an iterative reasoning process even though they were often facing a more difficult task than Kneeland's.

As expected, and by design, for each additional player added to the ring, the average level of reasoning increases by 0.291 ($p = 0.02$, ols regression clustered by subjects) in CL and by 0.659 in NYU ($p < 0.0001$, ols regression clustered by subjects). However, many subjects did not exhibit higher levels of reasoning as matrix size rose. Indeed, we found a negative interaction between the matrix size and ring size, also evident from the downward bowing of the curves depicted in Fig. 3b ($\beta = -0.106,\ p = 0.005$ in the CL sample, and $\beta = -0.106,\ p = 0.013$ in the NYU sample, ols regression clustered by subjects). Note that for larger matrix sizes (3*3 and 4*4 in CL, and 4*4 in NYU), when moving from 3-person to 4-person rings, curves are either parallel to the x-axis or even decline somewhat, suggesting that subjects reach a capacity-limit as social-complexity and/or arithmetic-complexity increase.

We further examined the chances of being classified as L1 or higher as a function of complexity (Fig. 3c). We found a significant negative interaction between ring size and matrix size (CL sample: $\beta = -0.130,\ p < 0.0001$, NYU sample: $\beta = -0.118,\ p < 0.0001$, probit regression clustered by subjects), suggesting that subjects' capacity was not sufficient for identifying the dominant strategy once complexity increased along either dimension. We found a similar pattern when we examined the chances for being classified as L2 or higher (Fig. 3d, CL sample: $\beta = -0.089,\ p < 0.0001$; NYU sample: $\beta = -0.081,\ p < 0.0001$, probit regression clustered by subjects). It is important to note that these results run counter to standard economic theory, which argues that subjects' level of reasoning remains constant *within* a specific class of games (when no experimental manipulation of features like the monetary incentives or the identity of players is undertaken)[18].

We note that due to the exclusion restriction criterion (ER, see Task and Methods), it was less likely for a subject to be classified as the same type (at a particular level of reasoning), as more players were added to the ring. For example, an individual could achieve classification as L2 in a 2-person ring with only four "correct" choices, but the same L2 type would require eight "correct" choices in a 4-person ring (see Methods).

For that reason, we also introduce a relaxed version of Kneeland's identification strategy that imposed the same classification criteria for each type regardless of ring size (ERrel, see Methods). Importantly, our main findings hold when using ERrel (Supplementary Fig. 2). Supplementary Figs. 3 and 4 further present two sensitivity tests performed on these two identification strategies (ERNo and ERrelNo, see Methods). Here, too, our results remain unchanged.

Finally, we find that having a higher level of reasoning paid off for subjects in the NYU sample, who were generally playing against more sophisticated players (Supplementary Fig. 5), but did not offer additional payoffs in the CL sample. Moving one level up increased subjects' winning amounts by $3.856 (p=0.009, ols regression) in the NYU sample, but had no effect in the CL sample (β=2.497, p=0.15, ols regression).

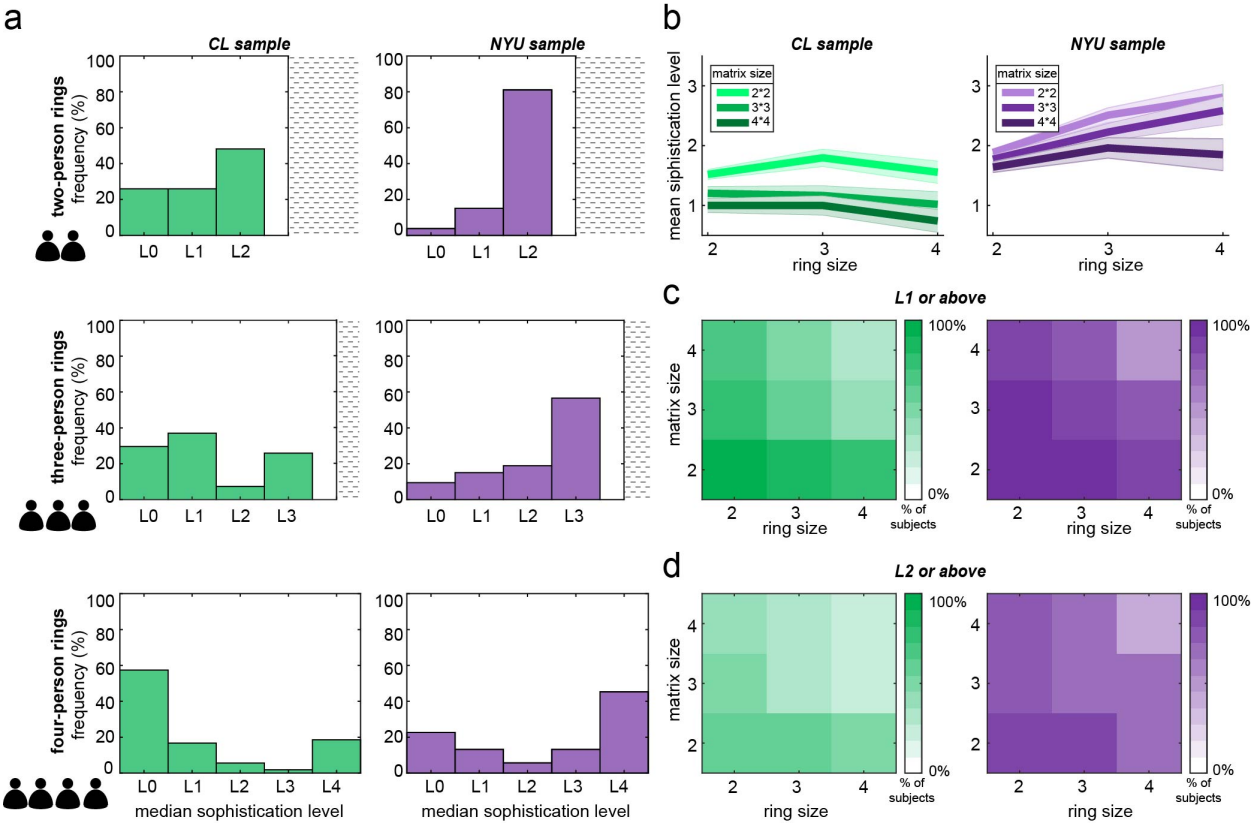


Fig. 3 | Classification into levels of reasoning, Exp. 1. (a) Distributions of types by the number of players in the ring. The x-axis presents the median type per subject for all the rings with the same ring size. *Left* – CL sample, *right* – NYU sample. Upper row – 2-persons rings (rings #1-9), middle row – 3-persons rings (rings #10-18), bottom row – 4-persons rings (rings #19-27). (b) Psychometric curves. Mean level of reasoning by ring size and matrix size. Error bars represent standard errors. *Left* – CL sample, *right* - NYU sample. (c) Heat maps showing the share of subjects classified as L1 or above as a function of ring size and matrix size. Subjects were classified as L1 or above if their median type in each cell in the 3*3 grid (which includes 3 different rings) was L1 or higher. (d) Same as (c), but for L2. (a-d) N_{CL} = 55, N_{NYU}=54.

Ruling out Simple Heuristics

Next, we examined the possibility that subjects' choices could have been guided by simple heuristics, which do not require an iterative mentalizing process. To that end, we examined two alternative models: (1) choosing the option that included the largest single payoff in the matrix faced by the chooser (MAX), and (2) avoiding the option with the lowest payoff in the matrix under consideration by the chooser (noMIN). Importantly, both the MAX and the noMIN heuristics do not depend on the choices of other players and hence should not involve mentalizing or social cognition. For 11 rings (out of 27) these heuristics produce the same choice behavior as would be observed from a perfectly sophisticated chooser who follows l_k. We found, however, that only five subjects (out of 109) used these heuristics in more than two thirds (18) of the rings. For a detailed account on this analysis, see Supplementary Note 3, Supplementary Table 11 and Supplementary Fig. 6.

Performance Index, Capacity Frontiers and Model-Free Classification

To empirically measure $Pr(l_k)$, the probability that an individual will express the maximal levels of reasoning (l_k) in position k in a given ring, we developed a *Performance Index* (PI): For each of the nine game types (Fig. 1d), we computed the share of responses (trials) that match l_k, normalized by chance-level performance in a specific game type (chance-levels vary with matrix size, see Methods).

We then visualize PI by creating for each subject a *Capacity Frontier* (Fig. 4a-b). The shaded area in each table in Fig. 4a indicates the game types in which a particular subject was able to reason beyond chance-level. The shape of the shaded region – whether more horizontal or vertical – indicates each subject's individual tendency towards performing better as ring size (social-complexity) or matrix size (arithmetic-complexity) increases. The Capacity Frontiers are a model-free representation of different kinds of individuals, as depicted using the Cobb-Douglas approach in Fig. 2b-d. Fig. 4a presents an illustration for five such types of individuals when examined in this manner. Fig. 4b presents frontiers for five representative subjects which largely mirror these illustrations.

To quantify this notion, we compute for each subject a Capacity Index and Trade-Off Index (TOI) (see Methods) based on summary statistics. The Capacity Index is the average PI (Performance Index) across all game types. A Capacity Index=1 is a subject who responds to all combinations of ring size and matrix size optimally. A Capacity Index=0 identifies a subject

whose choices are random (L0) for all game-types. Overall, subjects in the NYU sample exhibited a higher Capacity Index than subjects in the CL sample ($p<0.0001$, two independent samples, one-sided t-test, Fig. 4c).

The TOI evaluates subjects' tendency to perform better (higher PI) as one complexity dimension increases rather than another. A TOI=0 describes a subject for whom increases in ring size (social-complexity) and matrix size (arithmetic-complexity) exert equal effects on performance. A TOI>0 describes a subject who can achieve higher performance as social-complexity (ring size) increases than in response to equivalent increases in arithmetic-complexity (matrix size). A TOI<0 indicates the reverse, a higher performance as arithmetic-complexity increases than in response to equivalent increases in social-complexity. In both samples, we observe an asymmetry towards the demands imposed by social-complexity (TOI>0), though subjects in the NYU sample are more concentrated around TOI=0 (Fig. 4d-e, marginally significant at $p=0.0531$, Kruskal-Wallis test). Note, however, that the individual subjects who exhibit extremely strong asymmetry favoring social-complexity (TOI>1) also tend to exhibit very low overall capacity scores (mean Capacity Index=0.0572 in CL and 0.1151 in NYU, compared with sample averages of 0.3466 and 0.6371, respectively). Overall, we observed that the vast majority of subjects (82.6%) lay in the $-0.5 \leq TOI \leq 0.5$ band. Thus, subjects can largely trade-off (substitute) the allocation of their capacities between social and arithmetic demands.

Finally, we pay careful attention to reasoning in pairs of rings which show roughly complementary social- and arithmetic-complexity. Such pairs include 2*2 3-person vs. 3*3 2-person rings (pair #1); 4*4 2-person vs. 2*2 4-person rings (pair #2); and 4*4 3-person vs. 3*3 4-person rings (pair #3). Our populations behave as would be expected from a cohort with a TOI~0, there is no difference, at the population-level, in performance across these complementary pairs of rings (CL: $\beta = -0.0320,\ p = 0.057$; NYU: $\beta = 0.0161,\ p = 0.301$, ols regression clustered by subjects, Fig. 4f). In contrast, we do identify a significant decline in PI as overall complexity increases moving outbound from the origin (CL: $\beta = -0.0324,\ p = 0.002$; NYU: $\beta = -0.0308,\ p < 0.0001$, ols regression clustered by subjects, Fig. 4f).

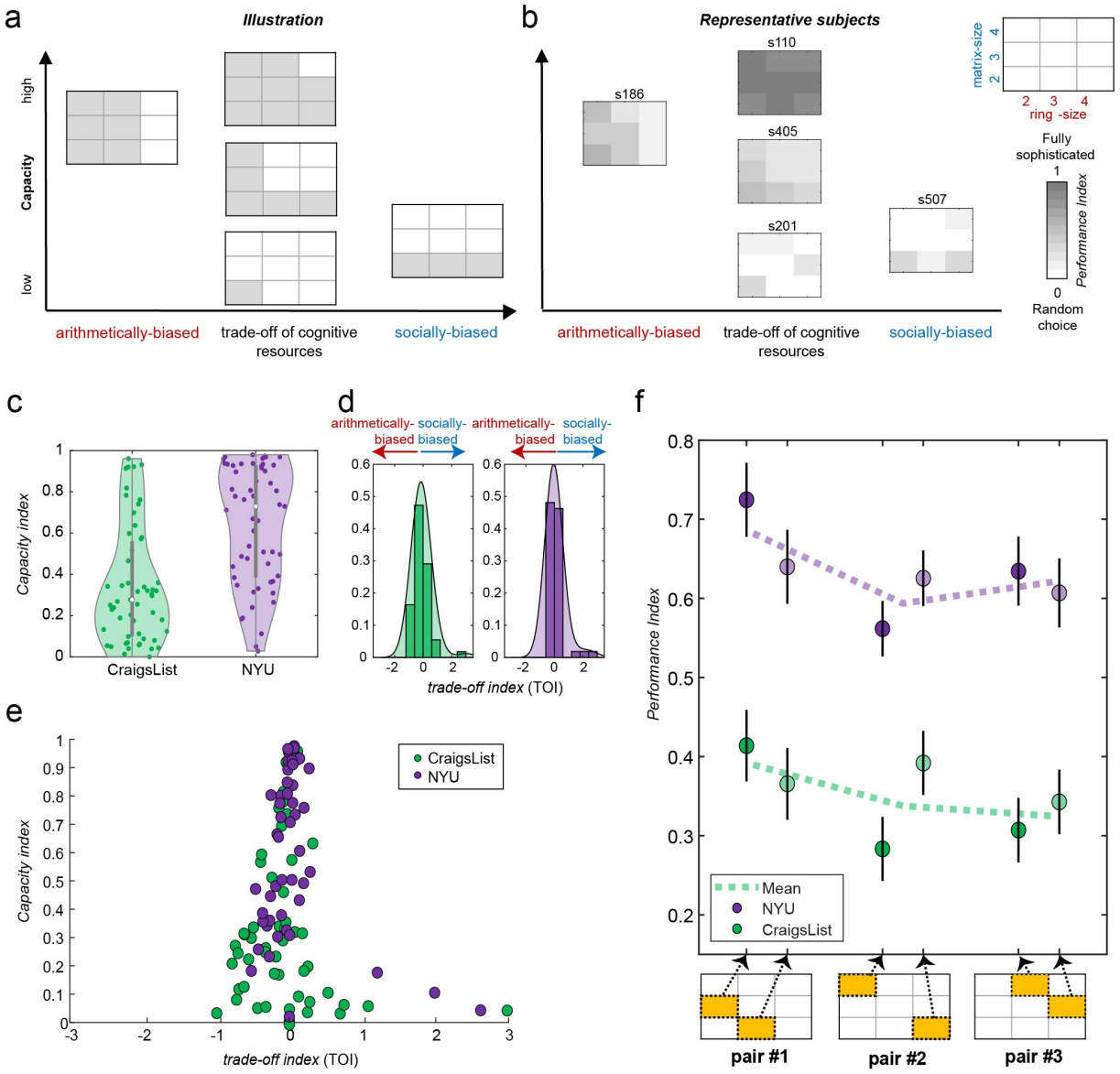


Fig. 4 | Model-free analysis, Exp. 1. (a-b) Capacity frontiers. For each subject, we can identify the subset of game types on which they reason above the chance level. The capacity axis shows increasing levels of performance (PI) across all game types. The trade-off axis refers to subjects' tendency to exhibit beyond-chance reasoning in rings with either a higher social-complexity (larger ring size) or rings with a higher arithmetic-complexity (larger matrix size; see Methods). (a) Conceptual illustration. Shaded area indicates rings for which the hypothetical subjects chose above chance-level (PI>0). (b) Representative subjects. (c) Distributions of the Capacity Index in each sample. CL: mean=0.3466, std=0.2853, min=0, max=0.9609, NYU: mean=0.6371, std=0.2848, min=0.0279, max=0.9794. Scatter plot points indicate individual subjects. Violin plots show Kernel smoothing, boxplots indicate interquartile range. (d) Histograms show distributions of TOI with Kernel smoothing by sample. CL: mean= -0.0776, std= 0.5846, min=-1, max=3; NYU: mean=0.0555, std= 0.5104, min= -0.5236, max= 2.6334. (e) Spread of TOI vs. the Capacity index, by sample. Scatters show individual subjects. (f) Mean PI in rings that show complementary social- and arithmetic-complexity. Left to right: pair #1: 3*3 2-person vs. 2*2 3-person rings; pair #2: 4*4 2-person vs. 2*2 4-person rings; pair #3: 4*4 3-person vs 3*3 4-person rings. Error bars indicate standard errors. (c-f) N_{CL} = 55, N_{NYU}=54.

Structural Psychophysical Analysis of Subjects' Task Performance

A central goal of this work is to validate a psychophysical model of how social and arithmetic cognitive capacities influence strategic reasoning in social interactions. To accomplish this, we modeled the individual likelihood to engage in iterative reasoning with a standard Cobb-Douglass production function from economics, as described above in eq. (1). We employed an ols estimation technique (see eq. (4) in Methods), using the specific cognitive demands of each ring (social- and arithmetic-complexities) as independent variables, and our Performance Index as the dependent variable. [We report both the sample pooled aggregate parameters (Table 1, Fig. 5a), as well as the subject-level estimates (Supplementary Tables 3-4, Fig. 5b).]

Our model relies on capturing human social strategic reasoning with two parameters, C and α. C is intended to capture the overall capacity of an individual, with higher values of C identifying individuals (or populations), who can achieve higher levels of reasoning under a given set of conditions. The parameter α is intended to capture the relative social and arithmetic capacities, with values below 0.5 identifying a higher relative social capacity and values above 0.5 capturing a higher arithmetic capacity.

In the CL sample, which more closely approximates the general population, we found that C was 3.83 (in arbitrary units; a.u.). In contrast, in the NYU sample of highly-selective US university students we found that C was 4.62. This difference between the two populations in overall capacity was highly significant (captured by a dummy for the sample in an OLS regression clustered by subjects, $p<0.0001$, Supplementary Table 5). Following eq. (1), and holding all other parameters equal, the higher capacity of the university students, would translate to PI scores higher by ~0.2 (a.u.), meaning, that they were 20% more likely to reach the maximal levels of reasoning in a given trial.

Turning to α, the parameter which indicates whether a subject is more socially or arithmetically capable, we found that across individuals the estimated values range between 0.346 (SE=0.1218, p=0.0088, high social-complexity orientation) and 0.786 (SE=0.1334, $p<0.0001$, high arithmetic-complexity orientation). These individual estimates were significantly different from 0 at $p<0.005$ for all the subjects in our study. A Shapiro-Wilk test verified that the alpha parameter is normally-distributed ($W=0.988$, $V=1.077$, $p=0.4341$), suggesting that the individual differences in social and arithmetic capabilities we observed appear to reflect a random variation in individuals in our populations. We note that, on average, our subjects exhibited a slight bias towards a higher arithmetic, rather than social, capacity ($t(108)=4.3921$,

p<0.0001, one-sample t-test with H0: α=0.5, also evident by the population estimate for α, which was 0.5385 (SE=0.0088, p<0.0001), Table 1).

Even though subjects exhibited individual differences in their relative social and arithmetic capacities, we did not find any significant differences in the value of α when comparing between our two populations. The CL sample showed an average estimated α of 0.5420 (SE=0.0126, p<0.0001) versus an estimate of 0.5349 (SE=0.0122, p<0.0001) in our sample of university students (t(107)=0.402, p=0.6885, two-sided two-sample t-test). Thus, although our sample of university students did show a significantly high overall capacity, this was not accompanied by a significant difference in their relative social and arithmetic capacities.

In order to assess the relationship between overall capacity and the bias towards arithmetic-capacity, we also examined the correlation between these parameters across all individuals from both cohorts. We found an inverse quadric relationship between the C and α parameters (first-order polynomial: β=0.5853, p<0.0001, second-order polynomial: β=-0.0690, p<0.00001, ols regression). Subjects with a very low overall capacity have no detectable orientation towards either dimension (social or arithmetic) because their responses are essentially random. Subjects with a very high overall capacity also show no significant orientation favoring either dimension, because they achieve essentially perfect performance up to the highest level tested in our task. However, subjects with an intermediate overall capacity do show a bias towards the arithmetic capacity. To assess the robustness of this conclusion, we examined the correlation between the model-free Capacity Index and TOI, and the model parameters C and α. We found a very strong correlation between the non-parametric capacity and C (r=0.9984, p<0.0001), as well as a strong correlation between TOI and αas shown in Fig. 5d (r=-0.4967, p<0.0001). This result further validates our modelling approach.

Table 1 | Pooled structural estimates, Exp. 1. Estimation via constrained OLS regression, clustered by subjects.

	α			C				N
	estimate	SE	pval	estimate	ln(C), regression constant	SE	pval	
Full Sample	0.5385	0.0088	<0.0001	4.2015	1.4354	0.0198	<0.0001	2,943
CL	0.5420	0.0126	<0.0001	3.8275	1.3422	0.0253	<0.0001	1,485
NYU	0.5349	0.0122	<0.0001	4.6200	1.5304	0.0249	<0.0001	1,458

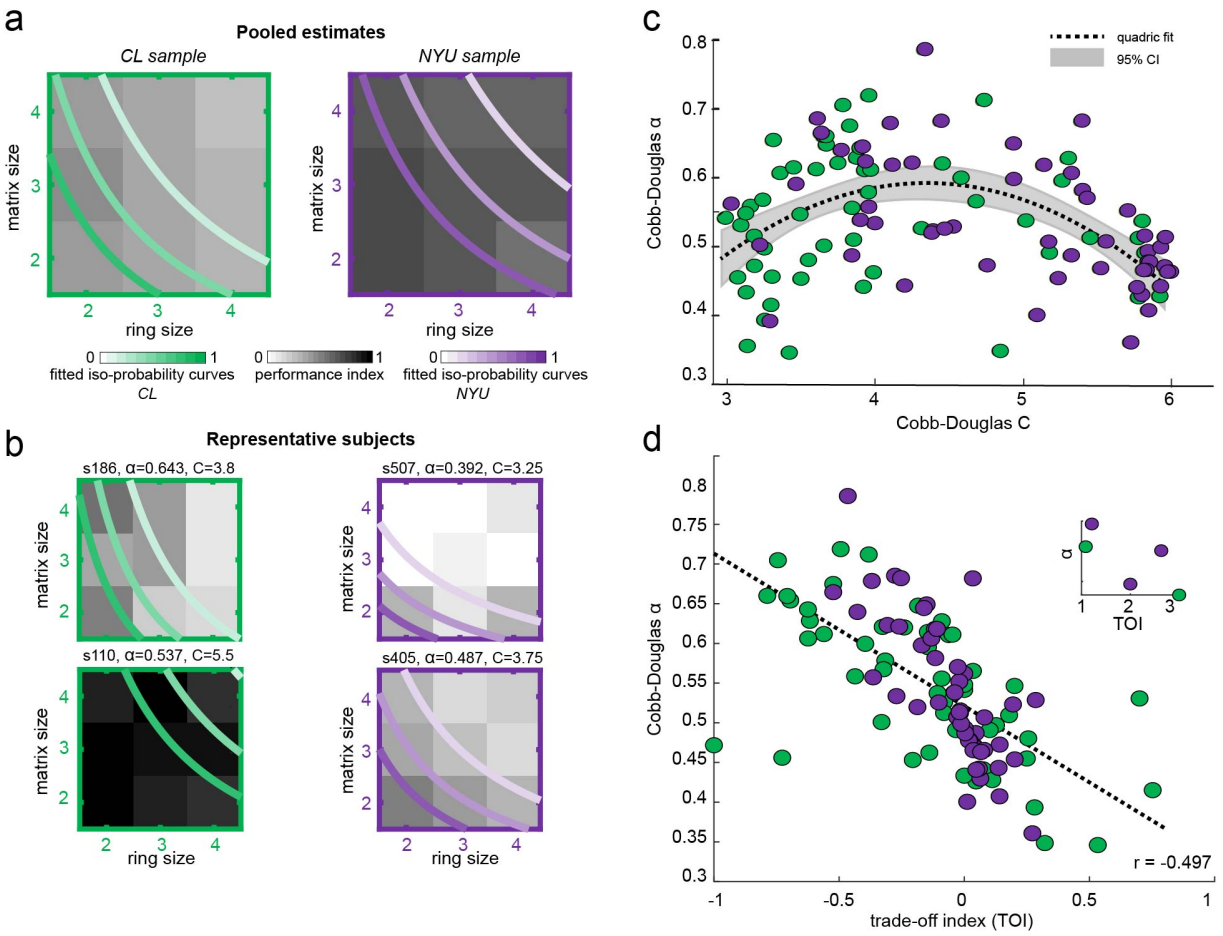


Fig. 5 | Structural model fitting. (a-b) Iso-probability curves. We fit subjects with our psychophysical variant of the standard Cobb-Douglas function (eq. 1 and Methods), and plot the derived iso-probability curves. Every coordinate on the same curve, has an equal PI. As curves move to the northeast PIs decline. Curves are plotted on top of the measured capacity frontiers. (a) Population-level estimates (left – CL, right – NYU). Iso-probability curves shown mark PIs of 0.25, 0.5 and 0.75. (b) Representative subjects. Iso-probability curves show PIs of 0.25, 0.5 and 0.75. Curves are shown for four of the subjects in Fig. 4b. (Subject 172 did not exceed chance level and is not shown). (c) Relationship between the C and α parameters from the Cobb-Douglas function. Dashed line indicates the quadric fit from a least-squares estimation.(d) Model-free trade-off index (TOI) compared with structural estimates of the α parameter from the Cobb Douglas function for all subjects. Inset: outliers with TOI>1. Dashed lined is least-squares fit). (c-d). scatters represent individual subjects, green – CL sample, purple – NYU sample (N_{CL} = 55, N_{NYU}=54).

Experiment 2: The Effect of Processing Time on Levels of Reasoning

In a second experiment (Exp. 2), we found that processing time strongly influences the levels of reasoning we observed. Our goal in Exp. 2 was to manipulate available cognitive resources by regulating processing time, and then investigate the relationship between available resources and sophistication. We examined a total of five processing times, from 1.6s to 10s (Fig. 6a and Methods). Even at the shortest processing times, subjects were able to complete

the task effectively (94% completion rate in the 1.6s treatment, Fig. 6b). We thus restrict our analysis to completed trials, and treat all processing times as yielding valid estimates of capacity.

We assessed the effect of processing time on subjects' reasoning via changes in the extracted parameters from our psychophysical (production function-based) model (Fig. 6c, Table 2). We find that the population-level estimate of the C parameter, which denotes the overall capacity of the subject, increases from 4.0436 (a.u.) in the shortest viewing duration (1.6s) to 5.1179 (a.u.) in the longest viewing duration (10s) (Table 2, Supplementary Table 6 and Supplementary Fig. 7b). We conclude that longer processing times increase the overall capacity available for reasoning.

The lawful increase in the C parameter can be described as a logarithmic function of processing time:

(2) $ln(C) = a * \ln(ProcessingTime) + b.$

We fit the data describing Capacity as a function of time with a multiplicatively scaled natural logarithm with a non-zero intercept and thus two free parameters, a and b (Fig. 6d and Methods). We find that for each additional second of processing time, $\ln(C)$ increases by ~0.123 (a.u.) divided by the current duration of processing time ($\frac{d}{dt} = \frac{0.123}{processing\ time}$), a relationship well captured by the multiplicative parameter a (a=0.1229, CI=[0.7662; 0.1691]). Plugging-in those values into the model detailed in eq. (1), if current processing time was 1.6s, the likelihood that a subject would reach the maximal level of reasoning in a given trial ($\Pr(l_k)$) would increase by ~8% for one additional second of processing but only by ~2% if the current processing time was 6.4s. (See Supplementary Table 7. Supplementary Fig. 7a replicates this result using the model-free Capacity Index instead of the parametrically estimated C). This lawful relation reveals that processing determines a subjects' accessible capacity, and hence the bounds on their sophistication. We note that this relationship is highly reminiscent of *Fitt's Law*, which states that measured performance accuracy is a function of task difficulty divided by task duration[41].

Unlike the C parameter, the α parameter – which captures the trade-offs between social and arithmetic capacities – remains unchanged across processing time treatments (Fig. 6e, Table 2). We find that α values are approximated at 0.5 at the population-level, regardless of

processing time. Hence, subjects trade-off social and arithmetic capacities at the same rate, even when the available computational resources change.

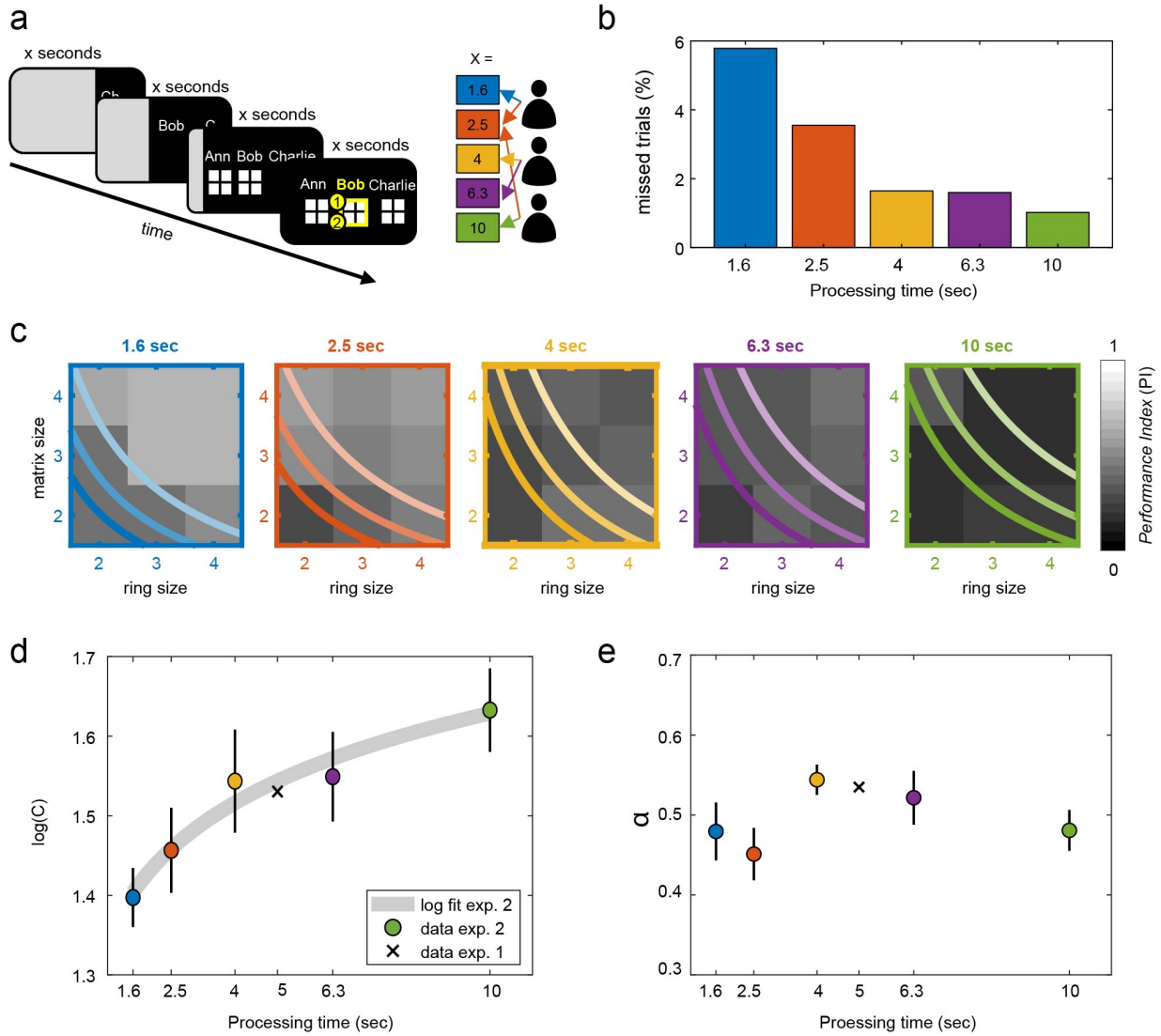


Fig. 6 | Exp. 2. (a) Processing time manipulation. In Exp. 2, we utilized 18 out of the 27 rings used in Exp. 1 to reduce total experimental time. Each subject (N=26) was assigned to two processing times out of five possible durations (1.6, 2.5, 4, 6.3 or 10 sec). Subjects completed all 18 rings twice for a total of 216 trials, 108 trials in each treatment. Processing times randomly alternated between blocks of 27 trials each. (b-e) We pooled subjects' responses and compared performance across treatments. (b) Share of missed trials across treatments. (c) Capacity frontiers with their Iso-probability curves across processing time treatments. The contour plots are based on population-level parameters extracted from the Cobb-Douglas model for each treatment (see Table 2) and indicate PI=0.25, 0.5 and 0.75. (d) C parameter across treatments of processing time. Gray curve indicates the natural log fit: $\ln(C) = a\ln(processing\_time) + b$, such that a=0.1229 CI=[0.07662; 0.1691], b= 1.346, CI=[1.275; 1.417], RMSE=0.0211. See Supplementary Fig. 7a for the same analysis using the model-free Capacity Index instead of the parametric C. (e) α parameter across treatments of processing time. (d-e) Error bars indicate the standard error of the estimated parameters. X indicates the estimates on data from Exp. 1 (NYU sample).

Table 2 | Pooled structural estimates, Exp. 2. Estimation via constrained OLS regression, clustered by subjects.

Processing time	α			C				N
	Estimate	SE	pval	Estimate	ln(C), Regression constant	SE	pval	
1.6 sec	0.4794	0.0362	<0.0001	4.0436	1.3971	0.0371	<0.0001	216
2.5 sec	0.4511	0.0328	<0.0001	4.2905	1.4564	0.0534	<0.0001	216
4 sec	0.5441	0.0189	<0.0001	4.6812	1.5436	0.0648	<0.0001	162
6.3 sec	0.5216	0.0337	<0.0001	4.7072	1.5491	0.0562	<0.0001	198
10 sec	0.4808	0.0255	<0.0001	5.1179	1.6327	0.0524	<0.0001	180

Incentives Influence Available Capacity

The levels of reasoning achieved by subjects also reflected the monetary incentives presented on each trial. In both experiments, we systematically manipulated the difference in payoff level (in dollars) between the best and worst options for the last player in each of the rings we examined (*dominant* vs *dominated* strategies, Fig. 7a and Methods). We hypothesized that in rounds with larger monetary differences between good and bad outcomes, subjects would allocate additional cognitive resources and thus could achieve higher levels of reasoning[18]. Such an observation would highlight that incentive magnitudes also influence the levels of reasoning a subject expresses.

To test this hypothesis, we estimated our psychophysical model for each payoff level, and examined the impact of the magnitude of the differences in payoff level on the parameters C and α (Fig 7b, Supplementary Table 8). Both Exp. 1 and Exp. 2 revealed an inverse relationship between payoff levels and capacity (C parameter). In Exp. 1, we found that when moving from the lowest to the highest payoff difference, capacity increased from 3.768 to 3.900 (a.u.) in the CL sample and 4.525 to 4.722 (a.u.) in NYU (constrained OLS regression. Estimates of the C parameter are significant at p<0.0001, Fig. 7c and Supplementary Tables 8). In fact, moving up one level of payoff difference increased ln(C) by 0.0193 (a.u., payoff difference was entered as a categorical variable to a constrained ols regression, p<0.001, Supplementary Table 9). In Exp. 2, the increase in accessible capacity due to payoff differences is evidenced by a shift upwards of the logarithmic function, capturing the relationship between capacity and processing time (eq. (2)). The model constant, *b*, increased from 1.306 (CI=[1.239; 1.372]) in the lowest payoff difference to 1.386 (CI=[1.311; 1.461]) in the highest payoff difference (Fig. 7d. Estimates of the C parameter were significant at p<0.0001, Supplementary

Table 10). These results resonate with the main conclusions drawn by Alaoui and Penta (2016)[18], indicating that monetary incentives should be expected to endogenously regulate accessible cognitive capacities devoted to reasoning.

Interestingly, however, in Exp. 1 the relative allocation of processing resources between arithmetic and social domains was also influenced by payoff difference. At low payoff levels, where accessible capacity was limited, subjects allocated a larger fraction of total resources to social capacity. As payoffs differences increased, subjects shifted towards a more equal allocation in that experiment. Analyzing the impact of payoff level on the arithmetic-social trade-off parameter, α, revealed a systematic shift. At higher payoff differences, subjects exhibited an arithmetic orientation in both samples. In the CL sample, α was estimated at 0.5389 (SE=0.0218) at the highest payoff difference, and 0.5698 (SE=0.0177) at the medium payoff difference. In the NYU sample it was estimated at 0.5420 (SE=0.0178) at the highest payoff difference, and 0.5805 (SE=0.1701) at the medium payoff difference. In contrast, we obtained evidence of a moderate social orientation in the lowest payoff difference, where α was estimated at 0.5172 (SE=0.0193) in the CL sample, and as low as 0.4823 (SE=0.0184) in the NYU sample (Fig. 7e and Supplementary Table 8). This change in resource allocation as a function of payoff in Exp. 1 implies that the allocation of resources to the social dimension is less costly than an allocation to the arithmetic dimension. However, it should be noted that the results from Exp. 2 do not provide further support for this conclusion (Fig. 7f).

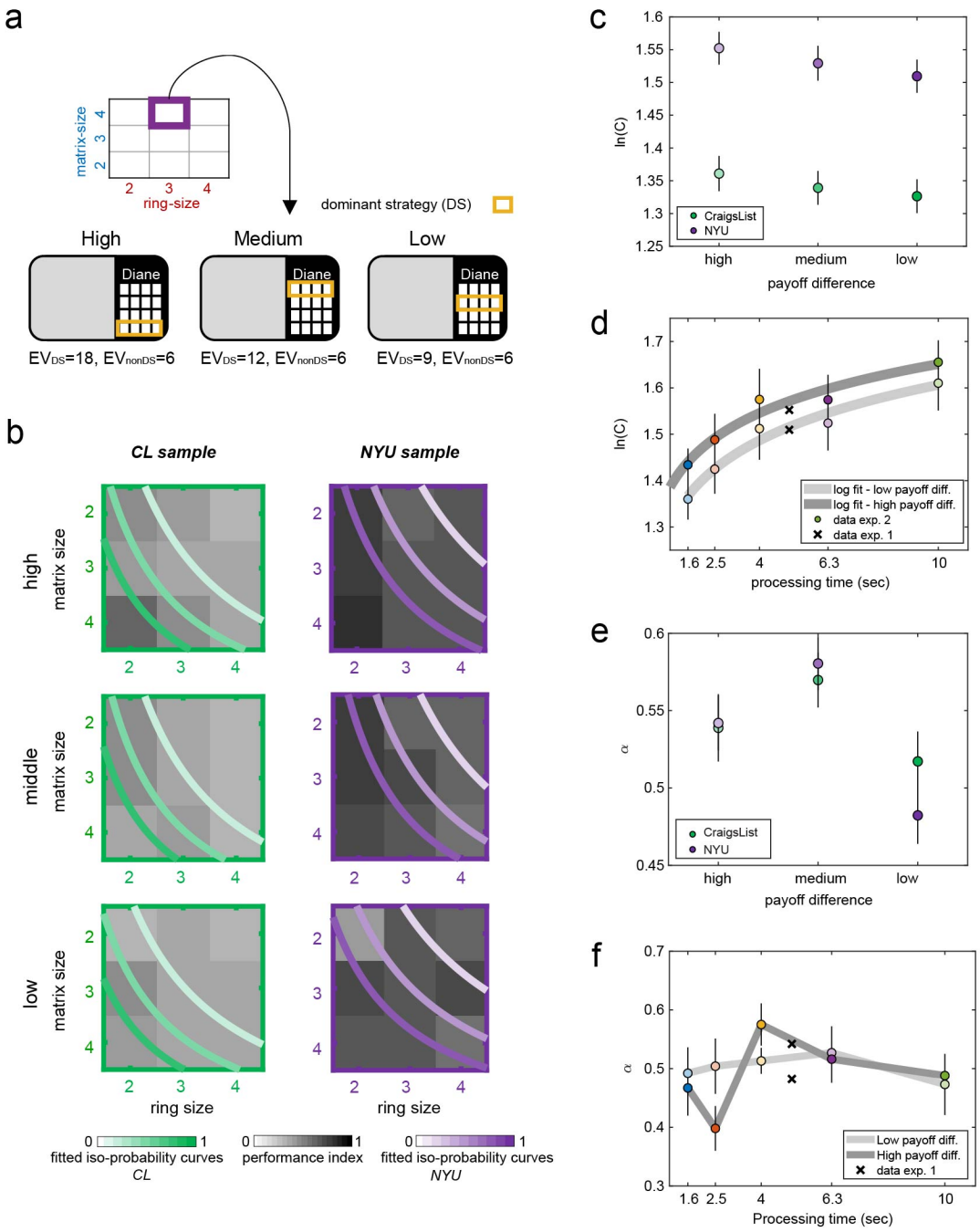


Fig. 7 | Incentives manipulation. (a) Incentives manipulation. We manipulated payoff levels by controlling the expected value (EV) of the dominating strategy (DS) for the last player in the ring. We created three levels of payoff difference: for each cell on the 3*3 rings grid, one ring had a DS with EV=18 (*high*), one ring had a DS with EV=12 (*medium*), and one ring had a DS with EV=9 (*low*). The other options (strategies) in the payoff matrix always had an EV=6. In Exp. 2, subjects only played the *high* and *low* rings. (b) Iso-probability curves sorted by payoff differences. Curves were fit at the population-level (eq. 1 and Methods), and plot iso-probability curves. As curves move to the northeast probability of an optimal response, PIs, decline. Curves are plotted on top of the non-parametrically measured capacity frontiers (left – CL, right – NYU). Iso-probability curves shown mark PIs of 0.25, 0.5 and 0.75. (c) Exp. 1. C parameter in each sample by payoff level (*Green* – CL sample, *Purple* – NYU sample). (d) Exp. 2. C parameter by payoff difference across treatments of processing time. Gray curves indicate the fit curve: $\ln(C) = a\ln(processing\_time) + b$. Low payoff difference: a=0.1152 CI=[0.0663; 0.1641], b= 1.386, CI=[1.311; 1.461], RMSE = 0.0223. High

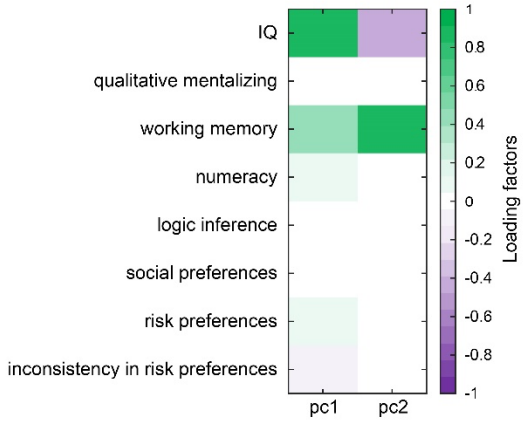
payoff difference: a=0.1306 CI=[0.0869; 0.1743], b= 1.306, CI=[1.239; 1.372], RMSE = 0.0199. (e) Exp. 1. α parameter in each sample plotted as a function of payoff difference (*Green* – CL sample, *Purple* – NYU sample). (f) Exp. 2. α parameter by payoff level across treatments that varied processing time. (c-e) Error bars indicate standard error of the estimated parameters. See Supplementary Fig. 8 for a complementary model-free analysis.

Relating our findings to classical psychological measures of capacity

Our findings indicate that subjects vary idiosyncratically in their arithmetic and social capacities, and that this influences the sophistication they employ as features of the game theoretic task they face change. We demonstrated this by using the common-practice from the behavioral economics literature of classification into *Levels of reasoning*, by using a model-free analysis, and by using a psychophysical analysis that employed a Cobb-Douglas cognitive production function. All three approaches indicate that subjects differ in how they allocate their internal cognitive resources; that they have idiosyncratic capabilities over resource utilization, in economic parlance. Our final aim was to determine whether we could relate these idiosyncratic capabilities to existing psychological measures of personality traits[42]. The most relevant psychological measures of traits that we examined in these same subjects were IQ (via Raven matrices[43]), an estimate of working memory capacity (WM, via the OSpan task[44]) and an estimate of qualitative mentalizing capacity (via the Perspective Taking task[15]). The full list of psychological measures we examine is detailed in Supplementary Table 12.

To relate these measures to the arithmetic and social capacities measured with our *Ring Game* task, we first reduced the dimensionality of scores derived from all of the psychological instruments we examined with a principal component analysis (PCA). We then focused on PCs 1-2, shown in Fig 8a, which capture nearly all of the variance in our measures (94.44% of explained variance, 69.55% and 24.89% for each component, respectively). We next correlated task performance measurements with these PCs (Fig. 8b). All of the performance measures from the *Ring Game* that capture total capacity correlated strongly with PC1 (mean level of reasoning: r=0.4825, Capacity Index: r=0.4963, Cobb-Douglas C parameter: r=0.4988, see Methods). These results suggest that overall capacity is strongly related to the classic psychological traits of IQ and working memory. In contrast, only our model-free TOI, our estimate of trade-offs over resource allocation (relative cognitive load) aligned weakly (r=0.1467), with a second PC that was associated with other elements of IQ and working memory (Fig. 8). Future work will be required to identify psychological measures that predict both the level of social capacity in an individual and the trade-off between social and arithmetic capacities.

a



b

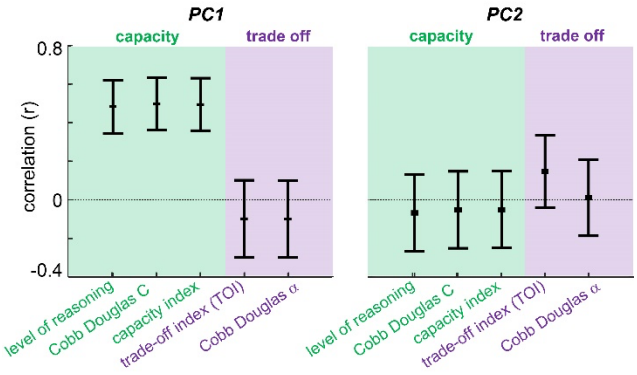


Fig. 8 | External validation of findings (Exp. 1). (a) Loading factors of external tasks on principal components 1 and 2. (b) Correlations between PCs 1-2 and task performance measures: mean sophistication level, Capacity Index, trade-off index (TOI) and the Cobb-Douglass structural parameters, C and α.

Discussion

Summary

We developed and explored the concept of a *cognitive production function* in order to describe how subjects' reasoning in strategic interactions is affected by complexity across two dimensions of interaction — a social dimension that captures the need to represent the mental states of other players and an arithmetic dimension that captures the need to represent non-social complexities. The strategic sophistication of subjects was measured while we manipulated these two forms of complexity using psychophysical variations of Kneeland's *Ring Game*[13]. Previous studies[4,5,16,17] have represented the cognitive sophistication of a subject as a fixed feature of the individual. Our results challenge this conclusion, indicating that an individual's sophistication, or level of reasoning, depends on an interaction between that individual's social-arithmetic capabilities and the precise demands of a given problem. Increased social-complexity does prompt individuals to reason more deeply about the mental states of others, but we find clear evidence that this increase in social-complexity also imposes an increased cognitive load. Similarly, we find that increased arithmetic-complexity also imposes a cognitive load. Across the two dimensions, the higher cognitive load reduces the chances that a subject engages in the maximum number of levels of reasoning available. Using both model-free measures and model-based (structural) estimates, we characterized the abilities of individual subjects across both social and arithmetic dimensions, and we also characterized the way performance across these dimensions trade-off as complexity varies. Indeed, we found significant differences between individuals and across our two samples, one drawn from the general public and the other from a highly-selective US university.

To test further our conclusion that the demands of the problem interact with the capabilities of the decision maker to define the cognitive sophistication of the individual's response, we also examined the effect of incentives and processing time on sophistication. In line with Alaoui and Penta (2016)[18], we found that when we increased the monetary impact of a decision, subjects improved their performance on both dimensions, while still being constrained by their relative capabilities on each dimension. In the second experiment, we extended this approach by causally manipulating the cognitive resources available to our subjects; changing the duration of processing time that subjects had to perform on each iteration of the *Ring Game*. We found that individuals' estimated capacity increased with each additional second of processing time, in a fashion quite precisely following a logarithmic progression as is often the case in psychophysical settings[41]. Finally, we found that an individual's capacity estimates

correlated with their scores on both an IQ test and a working memory test. We have not yet identified a psychological testing regime that correlates with our estimates of the trade-offs over the social and arithmetic dimensions of complexity.

Previous work by behavioral and experimental game theorists has attempted to classify individuals with regard to their level of strategic sophistication (or "type" in game-theory language). Level-k[5,13], which dominates behavioral game theory, hypothesizes that each subject can be classified as reasoning to a fixed number of levels. However, numerous studies have found that an individual's type (level of reasoning) depends on the task they are set[10,18,19,21,23,24]. Our results provide a fully cognitive-based resolution of these difficulties, supported by a tested experimental design. The social- and arithmetic-complexities of an interaction, the monetary incentives of a given task, and even the processing time allowed, all affect the strategic sophistication exhibited by an individual. The difficulty to date in reliably classifying individuals into types, together with heterogeneity of levels of reasoning within-subject, is not the result of misidentification or shortcomings in experimental design, but a genuine endogeneity (in the terminology of economics) to task context originating in cognitive constraints.

Our findings are complementary to recent advances, such as Alaoui and Penta (2016)[18], who introduced a framework for analyzing endogeneity in subjects' sophistication in games created by monetary incentives, and Gabaix and Graber (2023)[31], who introduced a model designed to capture subjective complexity of a decision problem (via a Cobb-Douglas production function). We see these studies, together with our own work, as laying the foundations for a cognitive theory of reasoning in games.

Classical psychophysics has tended to assess individuals along single dimensions[34,45]. Measures of visual or acoustic sensitivity record thresholds and the rates at which percepts grow using unidimensional strategies. One exception is classical two-dimensional signal detection theory, which views choices as the product of two-dimensional stimuli, allowing for covariance across the two dimensions[46]. Even this approach, however, views subject performance as a passive product of stimulus properties. Here, we develop a different form of multidimensional psychophysics[47] drawn from the economics of production functions. We build a novel psychophysical model in which individuals can regulate the allocation of internal cognitive resources to some individually-fixed degree. We represent subjects as having internal (or endogenous) capabilities that vary along two dimensions – social and arithmetic – which can be traded-off in a non-linear fashion against one another, in a subject-specific manner. This novel use of a production-function approach may offer an interesting direction in the

psychophysical analysis of higher cognitive functions and may be of particular use to neurobiologists.

The non-linear individual-specific trade-offs we uncovered between social- and arithmetic-complexity suggest a direct neurobiological approach to measuring an individual's psychophysically-measured cognitive production function. The existing neurobiological literature point to a direct link between the structure and activity of the temporoparietal junction (TPj) and an individual's ability to represent social-complexity[48–50]. In a similar vein, structure and activity in medial and lateral frontal cortex as well as in the posterior parietal cortex appear to be related to an individual's ability to represent what we refer to as arithmetic-complexity[51]. Trade-offs of neural resources between these networks were identified during moral decision-making[52]. Furthermore, a transcranial magnetic stimulation to the TPj decreased mentalizing capabilities, while simultaneously reducing activity in the ventromedial prefrontal cortex (vmPFC), demonstrating a direct causal evidence for an interaction between the so-called mentalizing and valuation networks[36]. Similarly, a transcranial direct current stimulation to the vmPFC increased mentalizing capabilities among subjects with autism syndrome disorder[53]. These observations suggest that the psychophysical assessment of each individual's capabilities via our cognitive production approach may be useful for identifying neurobiological markers of strategic types in future work.

Limitations of Our Approach

There are limitations to our study. One that bears particular mention is that the highest level of reasoning we test in our design was Level-4. While this places our study at, or beyond, the ranges of most studies in the literature, it does constrain the conclusions we can draw about behavior in other settings. Future studies will be required to broaden these findings to more complex strategic environments that impose demands outside the ranges we have explored. Another limitation is that we do not systematically manipulate the identities of the other players an individual faces. Subjects may increase their reasoning level when facing other players they think are high-capacity individuals[10,18].

Conclusions

In the current work we presented a novel framework, validated in two independent studies, for analyzing subjects' reasoning in strategic settings. We show that owing to a trade-off between limited cognitive resources, subjects exhibit a psychophysical dynamic range in their strategic sophistication, which is unrelated to external experimental manipulations outside the

game design. Our novel approach is the first to provide a psychologically-grounded mechanistic account of levels of reasoning in games, which also mitigates previous reports on inconsistencies in subjects' levels of reasoning. Finally, the nature of our psychophysical design calls for future neuroimaging studies, which will investigate the different brain networks involved in reasoning in games[35–39] and will allow mapping of the inter-network trade-offs between social and arithmetic neural computations. Such results may also prepare the ground for computational psychiatric investigations of disorders of social cognition[54,55].

Methods

Experimental design. We created a series of 27 different rings in a 3x3x3 design that varied in ring size, matrix size and levels of payoff-difference, to assay the cognitive modules that we hypothesized affect reasoning in games. To examine how social-complexity affects subjects' reasoning, we manipulated the number of players (ring size) in each ring, which could vary between two and four players. A 2-person ring forces an upper bound of L2 to type classification, whereas a 4-person ring relaxes this bound, and enables a classification up to L4. Similarly, to examine how arithmetic-complexity affects subjects' reasoning, we manipulated the matrix size, which could also vary between two and four choice options. This yielded a 9-cell grid of game types (Fig. 1b).

Lastly, we also manipulated monetary incentives, by varying the difference in payoff level (in dollars) between choice options. In each of the nine game types, we created three levels of payoffs, for a total of 27 unique rings. That is, for the last player in each ring ("Diane"), the magnitude of the average value of the dominant strategy could range between $9 (low), $12 (medium) and $18 (high), while keeping the expected value of the other inferior options constant at $6 (Fig. 7a). To avoid other context effects, the average value of all of the other choice options across all the other players in the ring ("Ann", "Bob", or "Charlie") was always $10.

Following Kneeland's original design, all players, except the last player in the ring, had no strictly dominant strategies, and each ring was repeated in two variations to test the ER criterion. The two variants of each ring solely differed by the order of choice options for the last player in the ring ("Diane"). Our design also addressed criticism about the original *Ring Game*[56]. Namely, in some of the rings the options which survived iterated deletion of dominated strategies (options which match l_k), were not necessarily the options with the highest reward

amount in each matrix (e.g. rings #2, 3, 19, 20 and 21), and could potentially lead to zero payoffs (e.g. rings #4, 6, 10, 11, 13, 14, 16, 19, 21 and 23). For a full list of the rings used in our study, see Supplementary Table 1. Supplementary Table 2 details the profiles of choices which would survive iterated deletion of dominated strategies in each ring.

Procedures. *Experiment 1.* In each trial subjects were presented with the payoff matrix for every player in the ring. Matrices were revealed in order from right to left. Each matrix was presented for five seconds, after which subjects were told which role they were playing in that specific round, and had up to five seconds to submit their choices, via the number-keys on the top row of the keyboard (Fig. 1e). All the roles across all the rings were randomized across subjects. Subjects faced a total of 162 trials, divided into six blocks of 27 trials each. Each trial was essentially a one shot game, as subjects received no feedback until the realization of a three trials after the entire experiment was complete.

Due to the complicatedness of the task, and the relative short time subjects had to process the information, we implemented a few standardizations of the presentation: (1) matrices were situated in the same location on the screen, regardless of the number of players in the ring at the current round. (2) We named the players in alphabetical-order: *Ann*, *Bob*, *Charlie* and *Diane*, instead of using the generic *Player 1*, *Player 2* and so forth. (3) Payoff tables (matrices) were kept at a constant size on the screen, and numbers were presented in a fixed font-size, regardless of the number of actions from which subjects had to choose. For example, in a trial with a 2*2 matrix, subjects were presented with a 4*4 matrix, but saw numbers only inside a subset of four cells at the upper left corner of the matrix. This procedure enabled us to control for saliency and sensory effects, reducing the degree to which these could have acted as confounds. (4) We labeled the actions of each player using the same letters, regardless of matrix and ring size: Ann's actions were always *a*, *b*, *c,* and *d.* Bob's actions were always *e, f, g* and *h*. Charlie's actions were always *i, j, k* and *l*. Diane's actions were always *m, n, o* ad *p*. In trials with 2*2 or 3*3 matrices, only a subset of these letters was used for each player. For example, in a trial with a 2*2 matrix 4-person ring, Ann's actions were *a* and *b*, Bob's actions were *e* and *f,* Charlie's actions were *i* and *j,* and Diane's actions were *m* and *n* (respectively). (5) We also used a fixed color-coding for the matrices, such that Ann's matrix was always shaded in red; Bob's matrix was shaded in blue; Charlie's matrix was shaded in Green; and Diane's matrix was shaded in purple. All colors had the same luminance on the CIE 1931 XYZ color space (e.g., equal Y values) to control for relative brightness' effects on salience and processing speed. (6) Finally, a gray "veil" covered the screen, and revealed each matrix every five

seconds; so that subjects would not be aware of the number of players in the ring that they were facing in that round until the end of the sequence. The gray shade of the veil had an equal Y value in the CIE 1931 XYZ color space to the colors that were used for the color-coding of the matrices. For screenshots of the experimental software, see Supplementary Note 1.

Before the experiment started, subjects read the instructions, and thereafter completed a questionnaire to verify that the task was clear (both the instructions and the questionnaire are presented in Supplementary Notes 1-2). For further details on the questionnaire see Kneeland (2015)[13]. Soon after, subjects completed a practice block on the experimental software, which consisted of 15 rounds, divided into short clusters of five trials. The first cluster allowed subjects up to 12 seconds to view each screen in the round, the second cluster allowed 8 seconds, and the last cluster resembled the actual experiment, and allowed the subjects up to 5 seconds for each screen.

After subjects finished the experiment, they completed a series of additional tasks. Our aim was to relate the model-free and the psychophysical estimates of capacity and trade-off of cognitive resources derived from behavior in the *Ring Game*, to existing psychological measures of personality traits, by using a battery of well-validated tasks. See Supplementary Table 12 for the list of tasks and for subjects' average scores in each such task. Lastly, subjects also completed a demographic survey.

Experiment 2. In Exp. 2, we wanted to test how processing times affected reasoning, to provide a full psychometric characterization of subjects' reasoning capabilities as a function of their accessible cognitive resources. To limit the total experimental time, in Exp. 2, we used a subset of 18 rings out of the full array of 27 rings, which included the nine rings with the highest level of payoff difference, and the nine rings with the lowest level of payoff difference. We tested, across subjects, five different processing times that were evenly spaced in 0.2 unit intervals on a log-scale, i.e. 1.6, 2.5, 4, 6.3 and 10 seconds. Each subject faced (at random) two different durations of processing times, and completed the 18 rings twice, each at one of the two durations, for a total of 216 trials. Subjects faced alternating blocks (in random order), such that in each block all trials had the same processing time for each screen. A total of 8 blocks of 27 trials each were presented. For Exp. 2, the practice block was divided into five clusters of 3 trials each, such that each cluster had a different processing time, in a descending order from 10 to 1.6 seconds. Subjects participating in Exp. 2 did not complete the series of additional tasks. All other procedures were identical to Exp. 1.

Participants. All subjects gave informed written consent before participating in the study, which was approved by the New York University School of Medicine's Institutional Review Board. *Experiment 1.* We recruited two samples: one sample (n= 60) was recruited via CraigsList to approximate the general New York City population. Three subjects decided to quit after reading the instructions, and two additional subjects were dropped from the sample due to technical problems during their run. We therefore report the results from the remaining 55 subjects (M=28, F=26, T=1, mean age=39.1, min=19, max=74). The second sample was recruited via the subjects' pool of the Center for Experimental Social Science (CESS lab) at New York University and was made up of New York University students (n=58). One subject decided to quit after reading the instructions, and three additional subjects were dropped from the sample due to technical problems during their run. We therefore report the results of the remaining 54 subjects (M=28, F=26, mean age=22.8, min=19, max=32). *Experiment 2.* Twenty-eight volunteering NYU students were recruited via the CESS subjects pool. Two subjects were dropped from the sample due to technical problems during their run. We therefore report the results of the remaining 26 subjects (M=11, F=15, mean age=24.6, min=19, max=32). In both experiments, we excluded subjects who majored in economics or business from recruitment, nor did we allow the participation of any subject who reported college-level coursework in Game Theory.

Payoffs. Subjects received a $30 participation fee for completing the study. To ensure subjects were highly incentivized, three rings were randomly selected for payment at the end of the experiment. Subjects were randomly and anonymously matched into groups and paid based on their choice and the choices of their group members in the selected rings. The number of members in each group and the corresponding rings that were selected for payment depended on the number of participants in each experimental session. Subjects received the sum of the dollar value of their payoffs in the selected rings. In Exp. 1, the average winning amount was $29.5 (std=10.1, min=5, max=47) in the CL sample and $29.9 (std=9.0, min=3, max=46) in the NYU sample. Respectively, in Exp. 2, the average winning amount was $28.0 (std=6.2, min=17, max=39).

Classification of subjects into levels of reasoning. Supplementary Table 2 presents the choice options that survived iterated deletion of dominated strategies for each role in each ring across the two variations. For our main analysis, we used Kneeland's original identification strategy[13]. Kneeland required that a subject who satisfied (k)-levels of reasoning, but not $(k+1)$-levels, would not respond to changes in their (m)-order payoffs, whenever m was greater

than or equal to k. In other words, subjects' choices should be consistent across the two variations of the ring for any higher-order type beyond their level of reasoning. For example, in a 4-person ring, a subject would be classified as L2 only if their choices matched iterated deletion of dominated strategies for the roles of Diane and Charlie, whereas for the roles of Ann and Bob, they repeatedly chose the same options across the two variations (respectively), regardless of changes in Diane's payoff matrix. In each ring, for each *type* (level of reasoning), we simulated all the choice profiles that matched that type. We then looked for exact matches between subjects' actual choices and simulated profiles and allowed up to one mistake. In case subjects' choices were within one mistake of being classified as belonging to two different types, we assigned them to the lower type. For further details, see Kneeland (2015)[13]. We refer this identification strategy as ER.

We also employed a relaxed version of ER (ERrel), similar to the one used by Brocas and Carillo (2021)[57]. In ERrel, we loosely interpreted Kneeland's requirement that subjects would not respond to changes in (m)-order payoffs, and allowed subjects to choose at random for any role beyond their level of reasoning. Following up on the example above, under ERrel, for L2 types, we only required that subjects followed iterated deletion of dominated strategies for the roles of Diane and Charlie. However, we no longer required that subjects would choose the same options across the two variations of the ring for the roles of Ann and Bob. Here, too, we allowed up to one mistake between actual choices and simulated choice profiles. The motivation for inducing ERrel was twofold: (1) We did not want to impose more severe classification criteria in longer rings. That is, according to the original ER, in a 2-person ring subjects had to match a specific choice profile with four trials in order to be classified as L2 (two roles * two variations of the rings). In contrast, in a 4-person ring, subjects had to match a profile of eight trials (four roles * two variations), which suggested a far more severe criterion for this type in longer rings. Once allowing random choice in the (m)-order payoffs, we unified the classification criterion for L2 across all ring sizes (and all other types for that matter). (2) ERrel also addressed some criticisms regarding the arbitrariness implied by the original ER[58]. The results from the ERrel classification are presented in Supplementary Fig. 2.

Since 18 of the 27 rings in our design are shorter than Kneeland's original 4-person ring, we also conducted two sensitivity tests, which we termed ERNo and ERrelNO. Here, we did not allow any mistakes and only looked for exact matches between subjects' actual choices and simulated profiles, thus mitigating the higher chances for misidentification in shorter rings. The results from ERNo and ERrelNo are presented in Supplementary Figs. 3 and 4 (respectively).

Performance index (PI) and capacity frontiers. To empirically measure $\Pr(l_k)$ (detailed in eq. (1)) for each game-type g, we presented a *performance index*:

(3) $PI_g = \frac{Pr(l_k)_g - RC_g}{1-CR}$.

In each trial, we examine whether the subject followed l_k, the maximum number of levels of reasoning available for a player in position k in the ring (yes =1, 0 otherwise). $Pr(l_k)_g$ averages this score across all the trials from the same game type $g \in \{1,2,\ldots,9\}$, i.e., all the rings within the same cell in the 9-cells grid depicted in Fig. 1d. Hence, for the purpose of this analysis, we analyze each trial independently. This means that we neglect subjects' responses for other trials from the same ring where they may have failed to follow l_k, even if those trials tested lower-order types. RC_g is the likelihood of following l_k by chance. Note that this probability is 50% for 2*2 matrices with two choice options, 33.3% for 3*3 matrices with three choice options, and 25% for 4*4 matrices with four choice options. $PI_g = 1$ indicates that the subject was fully sophisticated in that game type, and $PI_g = 0$ indicates that the subject chose completely at random in game type g. In cases where $PI_g < 0$ (when $Pr(l_k)_g$ is lower than RC_g), we impose $PI_g = 0$. The capacity frontiers presented in Fig. 4 are simply the graphical visualization of PIs across all game-types g.

Model-free analysis. While PI_g enabled us to explore the effect of our design within-subject, we also wanted to compare performance in our task across subjects, to investigate whether we trace differences in subjects' capacities and trade-off over arithmetic and social resources. For this aim, we employed both model-free and parametric analyses.

In the model-free analysis, we characterized subjects' general capacity by simply computing the average PI_g across all game-types ("Capacity Index"). To capture subjects' trade-off between social- and arithmetic-complexities, we computed their trade-off index (TOI) -

(4) $TOI = \frac{S-A}{S}$

Where S is subjects' average PI_g in the rings with the highest social-complexity (4-person rings), and A is their average PI_g in the rings with the highest arithmetic-complexity (4*4 matrices rings). Positive TOI indicates that subjects were more likely to follow l_k in rings with a high social-complexity, compared to rings with a high arithmetic-complexity, and vice-versa for negative TOI scores. A TOI=0 indicates that the subject was equally successful in their reasoning, regardless of game-type.

Model fitting. *Cognitive production function.* We extracted both individual-level and aggregate-level estimates of eq. (1) using a constrained ols regression, with the following econometric specification:

$$(5)\ \log(P\,I_r + 1) = \log C + \alpha \log\left(\frac{1}{m_r}\right) + (1-\alpha)\log\left(\frac{1}{n_r}\right) + \varepsilon,$$

where m_r and n_r are the social- and arithmetic-complexities in ring r. For technical considerations, we added a constant of 1 to PI_r to avoid the function's asymptotes. We used elicited parameters to visualize the iso-probability curves, depicted in Figs. 5-6.

Data & code availability

The experimental software and all the datasets generated and analyzed for the current study will be uploaded upon publication to OSF.

Acknowledgements

We thank M. Olifer for her help in running experimental sessions. We also thank P. Battigalli, the Glimcher lab members and attendees of the Annual Meeting of The Society for Neuroeconomics for invaluable comments. This work was funded with a grant from the Israel Science Foundation (#103/22 for V.K.D), and financial support from NYU Grossman School of Medicine (P.G.), NYU Stern School of Business, NYU Shanghai, and J.P. Valles (A.B.).

Author contributions

V.K.D., A.B. and P.W.G conceived the study and computational framework. V.K.D and P.W.G designed the tasks. V.K.D collected and analyzed the data from both experiments. V.K.D., A.B & P.W.G wrote the manuscript.

References

1. Connan Doyle, A. *The Final Problem*.
2. von Neumann, J. & Morgenstern, O. Theory of Games and Economic Behavior: 60th Anniversary Commemorative Edition. in *Theory of Games and Economic Behavior* (Princeton University Press, 2007). doi:10.1515/9781400829460.
3. Stahl, D. O. & Wilson, P. W. On Players′ Models of Other Players: Theory and Experimental Evidence. *Games Econ. Behav.* **10**, 218–254 (1995).

4. Nagel, R. Unraveling in Guessing Games: An Experimental Study. *Am. Econ. Rev.* **85**, 1313–1326 (1995).
5. Camerer, C. F., Ho, T.-H. & Chong, J.-K. A Cognitive Hierarchy Model of Games*. *Q. J. Econ.* **119**, 861–898 (2004).
6. Brandenburger, A. & Dekel, E. Hierarchies of Beliefs and Common Knowledge. *J. Econ. Theory* **59**, 189–198 (1993).
7. Brandenburger, A. *Language Of Game Theory, The: Putting Epistemics Into The Mathematics Of Games*. (World Scientific, 2014).
8. Costa-Gomes, M., Crawford, V. P. & Broseta, B. Cognition and Behavior in Normal-Form Games: An Experimental Study. *Econometrica* **69**, 1193–1235 (2001).
9. Costa-Gomes, M. A. & Crawford, V. P. Cognition and Behavior in Two-Person Guessing Games: An Experimental Study. *Am. Econ. Rev.* **96**, 1737–1768 (2006).
10. Agranov, M., Potamites, E., Schotter, A. & Tergiman, C. Beliefs and endogenous cognitive levels: An experimental study. *Games Econ. Behav.* **75**, 449–463 (2012).
11. Arad, A. & Rubinstein, A. The 11–20 Money Request Game: A Level- *k* Reasoning Study. *Am. Econ. Rev.* **102**, 3561–3573 (2012).
12. Crawford, V. P. & Iriberri, N. Fatal Attraction: Salience, Naïveté, and Sophistication in Experimental “Hide-and-Seek” Games. *Am. Econ. Rev.* **97**, 1731–1750 (2007).
13. Kneeland, T. Identifying Higher-Order Rationality. *Econometrica* **83**, 2065–2079 (2015).
14. Meyer, M. L., Spunt, R. P., Berkman, E. T., Taylor, S. E. & Lieberman, M. D. Evidence for social working memory from a parametric functional MRI study. *Proc. Natl. Acad. Sci.* **109**, 1883–1888 (2012).
15. Stiller, J. & Dunbar, R. I. M. Perspective-taking and memory capacity predict social network size. *Soc. Netw.* **29**, 93–104 (2007).
16. Battigalli, P. & Siniscalchi, M. Strong Belief and Forward Induction Reasoning. *J. Econ. Theory* **106**, 356–391 (2002).

17. Battigalli, P. & Siniscalchi, M. Interactive Beliefs and Forward Induction. Preprint at (1999).

18. Alaoui, L. & Penta, A. Endogenous Depth of Reasoning. *Rev. Econ. Stud.* **83**, 1297–1333 (2016).

19. Alaoui, L., Janezic, K. A. & Penta, A. Reasoning about others' reasoning. *J. Econ. Theory* **189**, 105091 (2020).

20. Goeree, J. K. & Holt, C. A. Ten Little Treasures of Game Theory and Ten Intuitive Contradictions. *Am. Econ. Rev.* **91**, 1402–1422 (2001).

21. Gill, D. & Prowse, V. Cognitive Ability, Character Skills, and Learning to Play Equilibrium: A Level-k Analysis. *J. Polit. Econ.* **124**, 1619–1676 (2016).

22. Brocas, I., Carrillo, J. D., Wang, S. W. & Camerer, C. F. Imperfect Choice or Imperfect Attention? Understanding Strategic Thinking in Private Information Games. *Rev. Econ. Stud.* **81**, 944–970 (2014).

23. Georganas, S., Healy, P. J. & Weber, R. A. On the persistence of strategic sophistication. *J. Econ. Theory* **159**, 369–400 (2015).

24. Cooper, D. J., Fatas, E., Morales, A. J. & Qi, S. Consistent depth of reasoning in level-k models. *Univ. Málaga Málaga* (2016).

25. Arad, A. & Rubinstein, A. Multi-dimensional iterative reasoning in action: The case of the Colonel Blotto game. *J. Econ. Behav. Organ.* **84**, 571–585 (2012).

26. Saks, M. & Wigderson, A. Probabilistic Boolean decision trees and the complexity of evaluating game trees. in *27th Annual Symposium on Foundations of Computer Science (sfcs 1986)* 29–38 (1986). doi:10.1109/SFCS.1986.44.

27. Federgruen, A. On N-person stochastic games by denumerable state space. *Adv. Appl. Probab.* **10**, 452–471 (1978).

28. Kalai, E. Bounded Rationality and Strategic Complexity in Repeated Games. in *Game Theory and Applications* (eds. Ichiishi, T., Neyman, A. & Tauman, Y.) 131–157 (Academic Press, San Diego, 1990). doi:10.1016/B978-0-12-370182-4.50010-6.

29. Frith, C. D. & Frith, U. The Neural Basis of Mentalizing. *Neuron* **50**, 531–534 (2006).

30. Frith, C. D. & Frith, U. Interacting Minds--A Biological Basis. *Science* **286**, 1692–1695 (1999).

31. Xavier Gabaix & Thomas Graeber. The Complexity of Economic Decisions. Preprint at https://dx.doi.org/10.2139/ssrn.4505599.

32. Ratcliff, R. A theory of memory retrieval. *Psychol. Rev.* **85**, 59–108 (1978).

33. Ratcliff, R. & McKoon, G. The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Comput.* **20**, 873–922 (2008).

34. Fechner, G. T. Elements of psychophysics, 1860. in *Readings in the history of psychology.* (ed. Dennis, W.) 206–213 (Appleton-Century-Crofts, East Norwalk, 1948). doi:10.1037/11304-026.

35. Konovalov, A., Hill, C., Daunizeau, J. & Ruff, C. C. Dissecting functional contributions of the social brain to strategic behavior. *Neuron* **109**, 3323-3337.e5 (2021).

36. Hill, C. A. *et al.* A causal account of the brain network computations underlying strategic social behavior. *Nat. Neurosci.* **20**, 1142–1149 (2017).

37. Coricelli, G. & Nagel, R. Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proc. Natl. Acad. Sci.* **106**, 9163–9168 (2009).

38. Hampton, A. N., Bossaerts, P. & O'Doherty, J. P. Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proc. Natl. Acad. Sci.* **105**, 6741–6746 (2008).

39. Bhatt, M. A., Lohrenz, T., Camerer, C. F. & Montague, P. R. Neural signatures of strategic types in a two-person bargaining game. *Proc. Natl. Acad. Sci.* **107**, 19720–19725 (2010).

40. Cobb, C. W. & Douglas, P. H. A Theory of Production. *Am. Econ. Rev.* **18**, 139–165 (1928).

41. Fitts, P. M. COGNITIVE ASPECTS OF INFORMATION PROCESSING: III. SET FOR SPEED VERSUS ACCURACY x.

42. Smith, A., Bernheim, B. D., Camerer, C. F. & Rangel, A. Neural Activity Reveals Preferences without Choices. *Am. Econ. J. Microecon.* **6**, 1–36 (2014).

43. John & Raven, J. Raven Progressive Matrices. in *Handbook of Nonverbal Assessment* (ed. McCallum, R. S.) 223–237 (Springer US, Boston, MA, 2003). doi:10.1007/978-1-4615-0153-4_11.

44. Oswald, F. L., McAbee, S. T., Redick, T. S. & Hambrick, D. Z. The development of a short domain-general measure of working memory capacity. *Behav. Res. Methods* **47**, 1343–1355 (2015).

45. Stevens, S. S. On the Theory of Scales of Measurement. *Science* **103**, 677–680 (1946).

46. Green, D. M. & Swets, J. A. *Signal Detection Theory and Psychophysics*. xi, 455 (John Wiley, Oxford, England, 1966).

47. Falmagne, J.-C. *Elements of Psychophysical Theory*. (Oxford University Press, Oxford, New York, 2002).

48. Kliemann, D. & Adolphs, R. The social neuroscience of mentalizing: challenges and recommendations. *Curr. Opin. Psychol.* **24**, 1–6 (2018).

49. Luyten, P. & Fonagy, P. The neurobiology of mentalizing. *Personal. Disord. Theory Res. Treat.* **6**, 366–379 (2015).

50. Mar, R. A. The Neural Bases of Social Cognition and Story Comprehension. *Annu. Rev. Psychol.* **62**, 103–134 (2011).

51. Dehaene, S., Molko, N., Cohen, L. & Wilson, A. J. Arithmetic and the brain. *Curr. Opin. Neurobiol.* **14**, 218–224 (2004).

52. FeldmanHall, O., Mobbs, D. & Dalgleish, T. Deconstructing the brain's moral network: dissociable functionality between the temporoparietal junction and ventro-medial prefrontal cortex. *Soc. Cogn. Affect. Neurosci.* **9**, 297–306 (2014).

53. Salehinejad, M. A. *et al.* Contribution of the right temporoparietal junction and ventromedial prefrontal cortex to theory of mind in autism: A randomized, sham-controlled tDCS study. *Autism Res.* **14**, 1572–1584 (2021).

54. Frith, U. & Happé, F. Autism: beyond "theory of mind". *Cognition* **50**, 115–132 (1994).

55. Baron-Cohen, S. Theory of mind and autism: A review. in *International Review of Research in Mental Retardation* vol. 23 169–184 (Academic Press, 2000).

56. Jin, Y. Reinvestigating R k behavior in ring games. *J. Behav. Exp. Econ.* **98**, 101878 (2022).

57. Brocas, I. & Carrillo, J. D. Steps of Reasoning in Children and Adolescents. *J. Polit. Econ.* **129**, 2067–2111 (2021).

58. Cerigioni, F., Germano, F., Rey-Biel, P. & Zuazo-Garin, P. Higher orders of rationality and the structure of games. (2019).