STEER: Assessing the Economic
Rationality of Large Language Models

Narun Raman [1] **Taylor Lundy** [1] **Samuel Joseph Amouyal** [2] **Yoav Levine** [3] **Kevin Leyton-Brown** [1]
Moshe Tennenholtz [4]

Abstract

There is increasing interest in using LLMs as decision-making "agents." Doing so includes many degrees of freedom: which model should be used; how should it be prompted; should it be asked to introspect, conduct chain-of-thought reasoning, etc? Settling these questions—and more broadly, determining whether an LLM agent is reliable enough to be trusted—requires a methodology for assessing such an agent's economic rationality. In this paper, we provide one. We begin by surveying the economic literature on rational decision making, taxonomizing a large set of fine-grained "elements" that an agent should exhibit, along with dependencies between them. We then propose a benchmark distribution called STEER (Systematic and Tuneable Evaluation of Economic Rationality) that quantitatively scores an LLMs performance on these elements and, combined with a user-provided rubric, produces a "STEER report card." Finally, we describe the results of a large-scale empirical experiment with 14 different LLMs, characterizing the both current state of the art and the impact of different model sizes on models' ability to exhibit rational behavior.

1. Introduction

Recently, much research has worked to leverage Large Language Models (LLMs) to create decision-making engines, configuring them either to act directly as economic agents (Cai et al., 2023; Horton, 2023; Wang et al., 2023a) or to serve as key elements of broader systems that do so (Zhuge et al., 2023; Wang et al., 2023b; Shen et al., 2023). LLM-

based agents are already showing strength in planning (e.g., for personal finance (Reworkd, 2023)), solving complex problems (e.g., medical diagnostics; McDuff et al., 2023), leveraging tools (e.g., Schick et al., 2023) and playing games (e.g., chess; Nakajima, 2023). Better decision-making capabilities will be critical for advancing the use of Reinforcement Learning from AI Feedback (RLAIF) to fine-tune chatbots, such as constitution-based approaches (Bai et al., 2022; Hong et al., 2023). LLMs may soon take the place of humans in some social science experiments (Horton, 2023; Aher et al., 2023; Park et al., 2023). Eventually, this research agenda offers the promise of realizing the longstanding AI dream of personalized economic agents.

How can we best configure LLMs to maximize their performance on decision-making tasks (e.g., via prompting to perform chain-of-thought reasoning; fine-tuning; or more complex architectures that make repeated calls to a model)? After we have done so, how well do LLM-based agents perform? Recent research has begun to develop testing regimes that can address these questions in various restricted domains. These include various narrowly defined tasks (Liu et al., 2023; Qin et al., 2023; Zhou et al., 2023; Yao et al., 2023); limited economic settings (Shah et al., 2022; Chen et al., 2021; Sinha and Khandait, 2020; Araci, 2019; Akata et al., 2023); and open-world video games (Wang et al., 2023a; Zhu et al., 2023). Going beyond such problem-specific approaches to assessing decision making more broadly requires a holistic approach to describing good decision making and explaining how it can be decomposed into different, individually testable components. One approach is to divide decision making into distinct, ad hoc tasks, emphasizing those that have been clasically studied in NLP (Liang et al., 2022; Gehrmann et al., 2021).

We advocate a different approach: enumerating first principles that describe how agents *should* make decisions, and then evaluating an agent's degree of adherence with these principles. Answering the normative question of how decision-makers should act has been the focus of more than a century of research in economics, cognitive psychology, computer science, and philosophy. The resulting literature provides a mature mathematical foundation for so-called economic rationality. The grounding principle is that agents should (implicitly or explicitly) quantify their preferences

---

[1]Department of Computer Science, University of British Columbia, Vancouver, Canada [2]Tel Aviv University, Aviv, Israel [3]Stanford & AI21 Labs, Palo Alto, California, United States [4]Technion & AI21 Labs, Tel Aviv, Israel. Correspondence to: Narun Raman <narunram@cs.ubc.ca>.

according to a utility function and make decisions to maximize their own expected utility. The literature further characterizes elements of economic settings that fundamentally impact rational behavior: e.g., stochasticity is different from determinism; multi-agent environments are different from single-agent environments; reasoning about how best to make decisions for groups of agents is different from reasoning about how to act as an individual. In some cases, the theory is prescriptive: e.g., it's better to maximize utility than to accept lower-utility alternatives. In other cases, things become more complicated: e.g., in multi-agent environments, determining the best choice depends on beliefs about how other agents will act. In still others, impossibility results rule out all desirable options, e.g., when deciding how to aggregate multiple agents' preferences. Finally, in some cases human decision-makers exhibit cognitive biases that deviate from rational behavior, even when the theory makes a firmly prescriptive recommendation.

How can we hope to assess rationality when the landscape is so complex? Our approach is to identify tests for which the "rational" answer is well defined. In cases where the prescriptive recommendation is clear, assessment is unproblematic. More ambiguous settings can be tested by explicitly asking for a desired behavior (e.g., eliciting a Nash equilibrium strategy). Such settings also often admit prescriptive special cases: e.g., even in multi-agent environments, it is never a good idea to play a dominated strategy. Axiomatic theories naturally give rise to meaningful tests, which can be applied from the simplest settings (e.g., the von Neumann–Morgenstern axioms for utility maximization) to the most complex (Arrow's axioms for social choice functions). Famous human subject experiments that illustrate cognitive biases also naturally give rise to tests, which human decision-makers often fail. In the end, we obtain a *STEER Report Card (SRC)*. We leave it to the end user to determine the scoring rubric: e.g., should the agent receive good grades for doing well only in a subset of simple settings; for being as rational as possible across the board; or for behaving as humanly as possible, including replicating biases?

More specifically, our work begins by identifying a rich and hierarchical taxonomy of 64 "elements of rationality" for which some notion of a "right answer" is well defined (Section 4). We define each element, giving an example of each in an appendix. We then move on (in Section 2) to describe how we used this taxonomy to derive a fine-grained benchmark distribution that serves as the basis for STEER report cards. Our benchmark allows each element of rationality to be instantiated in multiple "grade levels" of difficulty and in multiple domains (e.g., asking questions about finance vs. medicine). For 49 elements, we have written LLM prompts to synthetically generate 24,500 multiple-choice questions and manually validated 2,450 generations in total. We also discuss how SRCs can be graded. We built a web interface

for generating benchmark questions, validating them, and visualizing experimental results; it allows elements to be filtered by position in the taxonomy, by logical dependence (e.g., the "maximize utility" element depends on the "transitivity" element), by domain, and by grade level. To demonstrate the utility of our system, we generated full SRCs for 14 language models, ranging from Llama 7B to GPT-4 Turbo, evaluated on 134,750 test questions. This experimental setup is described in Section 3. We spent $4,800 making calls to OpenAI's API and devoted 13,240 GPU hours of compute to evaluating open models. Section 5 describes our results; here are some highlights. Across our benchmark, we found that model size correlated heavily with performance: models smaller than 40 billion parameters were not reliably able to outperform random guessing. Model performance consistently decreased with grade level. GPT-4 Turbo was consistently the best model across all of our metrics and elements; its performance was excellent up to grade 5 (Dominant Strategies), decent up to grade 7 (Avoidance of the Endowment Effect), and fell to random guessing from grade 9 (Best Response) and above. Self-explanation and few-shot prompting were consistently able to help. Self-explanation generally enhanced performance, albeit offering the most gains on lower-grade-level questions. Few-shot prompting enhanced model performance when we offered up to three examples, but decreased performance beyond that point. We release all model outputs to support evaluation research and contributions, and provide a public website with all results (`https://steer-benchmark.cs.ubc.ca`), underlying model predictions details, alongside an extensible codebase to support the community in taking SRCs further.

2. STEER: Systematic and Tuneable Evaluation of Economic Rationality

We followed the standard practice of encoding our benchmark in Multi-Choice Question Answer (MCQA) format (e.g., Hendrycks et al., 2021; Suzgun et al., 2022; Zellers et al., 2019). More specifically, each question in a test is a description of a decision-making scenario and a set of candidate choices, exactly one of which is correct. All of the generated questions are organized hierarchically in a web application according to our taxonomy. The remainder of this section describes the methodology we employed in generating and validating these questions and different ways users can leverage them to construct STEER Report Cards (SRCs).

2.1. Generating and Validating Questions

It would be impractical to hand-construct enough questions to assess an agent's behavior with statistical significance. Instead, we leverage a state-of-the-art LLM to generate a
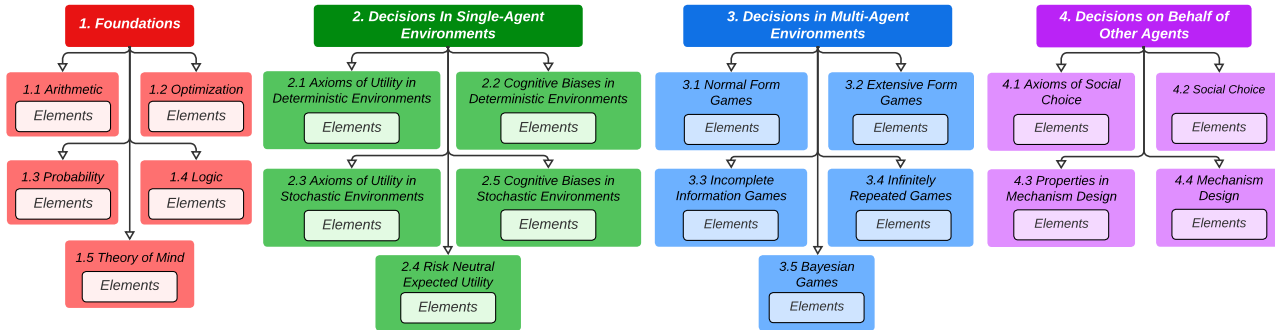
Figure 1: High-level diagram of the taxonomy of elements of rationality. At the top level, we divide the space of decision making into settings: FOUNDATIONS, DECISIONS IN SINGLE-AGENT ENVIRONMENTS, DECISIONS IN MULTI-AGENT ENVIRONMENTS, DECISIONS ON BEHALF OF OTHER AGENTS; we further subdivide settings into modules (e.g., Cognitive Biases in Deterministic Environments) that capture conceptually similar behaviors.

diverse and substantial set of questions, based on a hand-constructed inputs. More specifically, we go from an element to a concrete question as follows. First, we write detailed text describing what makes a good question, such as that each outcome should have an associated probability or that each action pair should have payoffs. These instructions also describe formatting issues, such as how numerical values should be represented. We also provide a gold-standard example for each question. Together, we call these two text strings a *template*. Along with the template we append a static system prompt (illustrated in Figure 2) and then repeatedly give the resulting prompt to GPT-4 Turbo to generate many questions from the template.

We create several templates for each element, which differ in two key ways. First, we vary the subject matter of the question, which we call the *domain*, in order to enable assessments of LLM robustness across topics. This approach not only tests an LLM's capability within each domain but also enables assessment of an LLM's proficiency, or lack thereof, in specific potential areas of application. Some of our templates concern financial decision making, others focus on medicine and health care, still others ask about technology and innovation. Second, most of our templates also vary in their difficulty levels. We assign every template a *grade level* ranging from 1–13 to help the user understand its relative difficulty. For example, questions about arithmetic could vary from Grade 1 to Grade 2 depending on the number of digits, whereas questions about Nash equilibrium could vary between Grade 8 and Grade 11. Our specific choices of grade levels are obviously somewhat arbitrary, but we aimed as much as possible to maintain a similar difficulty level across templates in the same grade. Overall, templates at lower grade levels involve basic understanding or application of principles while templates at higher grade levels challenge models with complex problem-solving, critical thinking, and synthesis of concepts.

We refer to a set of questions for a given element restricted to a particular domain and grade level as a *test*. For each test, we implemented 5 templates, totaling between 10–40 templates per element. Figure 3 provides two example questions testing the ability to Maximize Expected Utility that vary in both domains and grade levels. The user can explore our full set of templates through our web application, selecting elements and viewing questions across both domains and grade levels.

While GPT-4 Turbo was very good at generating questions, it was not perfect. We thus performed a validation step. First, we programmatically removed questions that were formatted incorrectly. For each element, domain, and grade level triple we then randomly spot-checked 100 samples (i.e., 10% of all generated questions) from what remained. We developed our web application to facilitate such validation, displaying the information needed to ensure that a given generation not only adheres to the intended style and complexity but also properly captures the corresponding element of rationality. We illustrate this interface in the appendix. In total, 98.54% of all spot-checked samples were deemed valid by 2 validators, with the lowest validation rate being 97% (Avoidance of the Endowment Effect).

2.2. STEER Report Card

SRCs function much like an academic report card, providing a structured and tuneable evaluation of an LLM's performance. Users can subset questions in whichever way best suits their needs or use cases. Figure 4 gives an example of such a subset based on a grade range. We also include various default subsets for different settings (e.g., single-agent decision making; multi-agent decision making), different models (e.g., matching human performance on cognitive biases), and different use cases (e.g., single-agent decision making in medical domains). These default subsets are

3

Figure 2: Example generation of a question in Avoidance of Sunk Cost Fallacy. The generation has two parts: (1) a user prompt containing a template question and instructions to follow the formatting and style of the template; and (2) a static system prompt. The template question in this example is in domain *created a project* and at grade level 5.
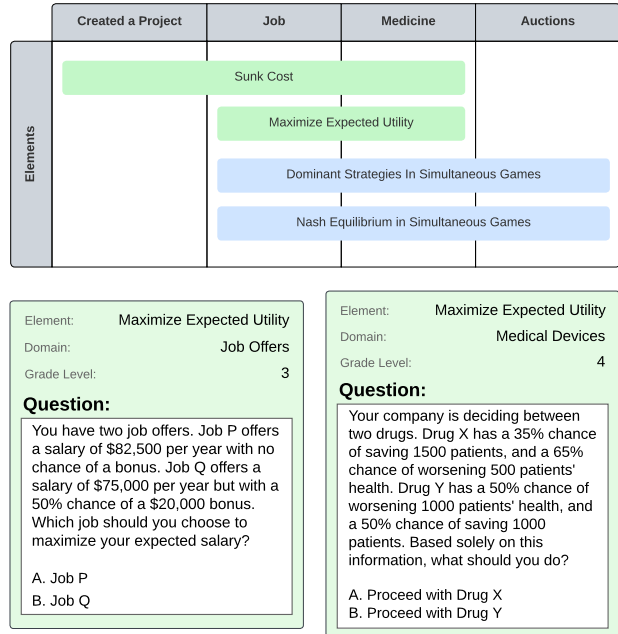

Figure 3: Example domain categorization of questions within elements of rationality and an example of two questions in different tests for the element Maximize Expected Utility. Top: We instantiate questions into as many domains as makes sense for the element of rationality. This figure depicts the domain span of questions for four different elements. Bottom: Two questions in two domains: job offers and medical devices, and two grade levels. Here, a higher grade level means more outcomes in the options. On the right (Grade Level 4), we see two options each with two outcomes, whereas the one on the left (Grade Level 3) has one option with two outcomes and the other with one.

accessible via our web application.

2.2.1. SCORING AN LLM'S PERFORMANCE

There are multiple ways in which users may want to evaluate an agent's economic rationality. We thus offer two different families of scoring metrics.

2.2.2. ACCURACY

Exact-match accuracy. This is the fraction of questions answered correctly.

Normalized accuracy. Exact-match accuracy scores can be hard to interpret since tests differ in their number of multiple choice options, meaning that the exact-match accuracy of random guessing varies. We can compensate for this by reporting the gap between the model's exact-match accuracy and random guessing. We compute normalized accuracy for a given element by subtracting the accuracy achieved by random guessing from exact-match accuracy and then dividing by the accuracy of random guessing. Observe that the normalized accuracy of a model falls between -1 and 1.

Calibration. It is often important that an LLM be able to express the uncertainty of its recommendation. To quantify such uncertainty, we follow Liang et al. (2022) and compute the expected calibration error (ECE; Naeini et al., 2015; Guo et al., 2017). ECE measures how closely the probability an LLM assigns to its top answer matches the actual probability of the correct answer, which in our case is 1. It is defined as $\sum_i^N b_i ||(p_i - c_i)||$, where p_i is the exact-match accuracy in bin i, c_i is the average probability assigned to top answers in bin i, and b_i is the fraction of data points in bin i. We allow users to choose the number of bins, however, we used 10 uniformly spaced over the interval $[0, 1]$ as is standard.

2.2.3. ROBUSTNESS

We can also assess how robustly a model performs, both across domains and across simpler elements that are conceptual subproblems of a given element.
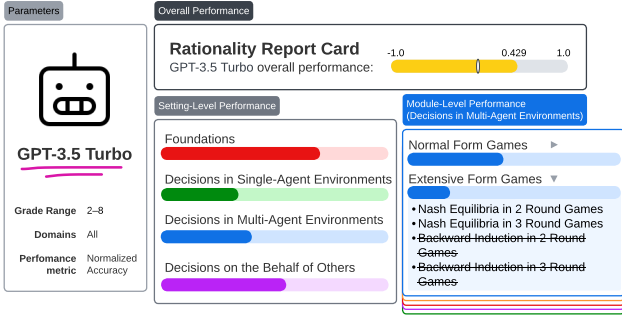
Figure 4: **Example SRC for GPT-3.5 Turbo.** On the left are the components necessary for measuring performance: a grade range, a set of domains, a performance metric and a model. The rest of the report card summarizes performance over the entire dataset, grouped by settings and modules in which there are questions in the grade range. In this example, GPT-3.5 Turbo was evaluated on all domains in grades 2–8 and given a single example as part of the prompt of a task. The right-most pane drills into DECISIONS IN MULTI-AGENT ENVIRONMENTS; crossed out text illustrates tasks omitted due to our grade level filter.

Domain robustness. One way to assess robustness is to measure, for each element, a model's worst-case performance across all domains. We compute the domain robustness on each element by taking the minimum exact-match accuracy over all domains.

Dependency robustness. There is a hierarchical structure to our elements of rationality: reasoning about more advanced elements (e.g., the ability to play a best response) depends conceptually on simper elements (e.g., the ability to maximize utility), which of course can depend in turn on yet simpler ones (e.g., the abilities to maximize a function and compute an expectation). We express all such conceptual relationships between elements in a *dependency graph*. Figure 5 depicts a concrete example of a dependency subgraph for the element Iterated Removal of Dominated Strategies instantiated at Grade Level 7, where there are two agents each with two actions. The full graph which covers all elements in our taxonomy is accessible in our webapp.

It is quite possible for an LLM to be proficient at advanced tasks without proficiency at more basic tasks that make them up. But such behavior is probably not desirable; it offers evidence that if the advanced task were discussed in different terms (e.g., in ways that invoke the conceptually simpler subtasks) model performance could fall. Conversely, if a model *fails* at an advanced task, it can be informative to trace performance backwards in the dependency graph to understand the model's performance on the task's building blocks. We call the quantification of this idea *dependency robustness*. For an element s we define it
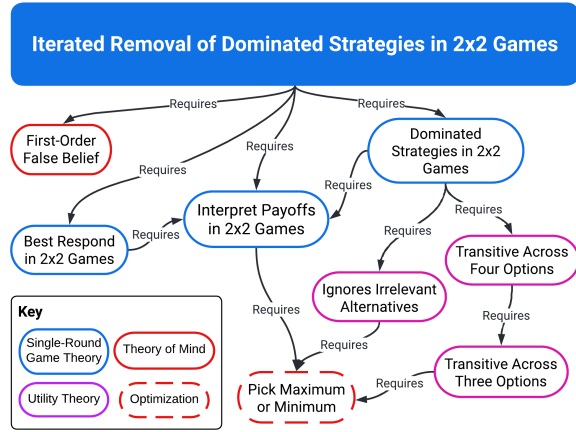

Figure 5: Dependency subgraph for iterated removal of dominated strategies in two-agent games with two actions for each agent. This node requires the ability to interpret the format of a game matrix (both normal and bimatrix form), correctly answer first-order false belief tasks (have knowledge of others' beliefs), best respond (choose the best action given a fixed action for the opponent), and find dominated strategies. Being able to find dominated strategies requires the ability to have orderings over payoffs which is tested via transitivity and ignoring irrelevant alternatives. The remaining nodes can similarly be broken down.

as $\sum_{x \in X} |\text{random\_gap}(s) - \text{random\_gap}(x)|$, where $X = \{x | \text{random\_gap}(x) < \text{random\_gap}(s)\}_{x \in G_s}$ and G_s is the dependency subgraph for some element s.

3. Applying our Benchmark: Setup

Table 1 lists the 14 different LLMs we evaluated, varying in parameter size. We ran GPT 3.5 Turbo and 4 Turbo using OpenAI's API (OpenAI, 2020) and AzureOpenAI. We obtained 12 open-source models from the HuggingFace Hub (Wolf et al., 2019) and ran them on an A100 GPU on Compute Canada. We decoded from all LLMs by sampling with temperature 0.

As is standard in model evaluation work (c.f., Liang et al., 2022) we treat LLMs as black boxes that take input strings and output completion strings along with log probabilities, when available. That means we do not assume access to the internal activations nor a model's training data. However, we can employ various widely used techniques to tailor an LLM to a desired question. Two common alternatives are self-explanation and prompting.

Self-Explanation. Much work has shown that question answering performance can be improved by asking a model to explain its reasoning (Wei et al., 2022; Yoran et al., 2023; Huang et al., 2023). We take two approaches to implement

Name	Origin
GPT-4 Turbo (1106-preview)	OpenAI (2023)
GPT-3.5 Turbo (0613)	×
Llama-2 {7b, 13b, 70b}	Touvron et al. (2023b)
Llama-2 Chat {70b}	Touvron et al. (2023b)
Llama {7b, 13b, 65b}	Touvron et al. (2023a)
Alpaca {13b}	Taori et al. (2023)
Vicuna {13b}	Chiang et al. (2023)
Falcon {7b, 40b}	Almazrouei et al. (2023)
Falcon-Instruct {7b, 40b}	Almazrouei et al. (2023)

Table 1: The 14 LLMs evaluated in Section 5 and their numbers of parameters.

this idea, which we dub *separate* and *together*. In *separate*, we call the model twice, first providing the question text and candidate options and asking the model to explain its reasoning, and then providing only the candidate options and asking the model to select the correct answer. In *together*, we only call the model once, giving the model the question text and candidate options, asking it both to explain its reasoning and to select the correct answer. For each approach, we test the effect that it has on model performance measured both on the accuracy and the confidence (i.e., log probabilities) a model places on its answer.

Few-Shot Prompting. Model performance can also be improved by prepending a set of examples to the prompt. For each question, we select $n \in \{0, 1, 2, 4, 5\}$ examples (within the corresponding domain and grade level) to test the effect of prompting on a model's performance. Similarly, we measure the effect by computing the difference on accuracy and confidence for each element.

4. Elements of Economic Rationality

Building a structured assessment of LLMs' economic rationality requires first deciding on a way of structuring the space of economic behavior. Fundamentally, economics is concerned with decision making, and an economically rational agent is one that makes good decisions given its own interests and its knowledge about the environment in which it acts. Different economic environments can give rise to very different decision-making problems. We thus divide our space into increasingly rich *settings*. We begin with DECISIONS IN SINGLE-AGENT ENVIRONMENTS, exploring preference formation and decision making when an agent has a set of alternative choices, each of which leads either to a single, deterministic outcome or to a draw from a probability distribution over outcomes. DECISIONS IN MULTI-AGENT ENVIRONMENTS enriches this setting, requiring the agent to make decisions when the outcomes depend on interactions with other economic agents with their own preferences and beliefs. DECISIONS ON BEHALF

OF OTHER AGENTS asks the agent to aggregate the preferences of other agents to achieve good outcomes for all. Lastly, FOUNDATIONS are core mathematical and cognitive skills that underlie economic reasoning: arithmetic, optimization, probability, logic, and theory of mind. (We will hereafter list FOUNDATIONS first, since it is the simplest of all.) Each setting is partitioned into multiple distinct modules, and each module consists of multiple *elements* of rationality: measurable capabilities that an economically rational agent is able to exhibit. We give a graphical overview of the taxonomy in Figure 1. Because it is very long, we leave the element-by-element description of our taxonomy—including an example for each element—to Appendix A.

Even given our restriction to representative, testable elements of rationality rather than all of decision making, it is an under-specified task to determine what we should demand of an economically rational agent. We, therefore, always align these canonical "right answers" with the von Neumann-Morgenstern (vNM) utility axioms, which imply freedom from cognitive biases, hence e.g. time consistency and reference independence. In some cases, the vNM axioms are not sufficiently constraining; thus, e.g., we align our canonical answers with a linear utility for money (and thus risk neutrality). We set up our taxonomy and SRCs in this way not because we believe that our canonical answer is the right one for every circumstance, but instead because we found it simplest to present both elements of rationality and experimental results in terms of an easily understood reference point.

5. STEER Report Card

This section assesses models' relative performance, their robustness, and the extent to which they were improved by adaptation strategies. We report results using multiple different SRCs to highlight different aspects of model performance and to give more examples of how SRCs can be used. Our system's web interface can be used to drill more deeply into model performance, the underlying model outputs, and the precise inputs and prompts that generated those outputs. As space permits, we also showcase some of those features here.

Whole-Benchmark SRC. Our first SRC aggregates performance across our whole benchmark, using both normalized accuracy and exact-match accuracy. Table 2 shows the results, sorting models in descending order by exact-match accuracy. Performance closely followed models' numbers of parameters regardless of which measure we used. For models that consistently performed better than random guessing (i.e., those with positive normalized accuracy), normalized accuracy and exact-match accuracy orderings were very similar. The same pattern held in our other experiments; thus, for space reasons, we hereafter focus mainly on normalized

Model	Normalized	Exact-Match
GPT-4 Turbo	0.3302	63.61%
GPT-3.5 Turbo	0.3071	61.69%
Llama-2 70b	0.2897	41.73%
Llama 65b	0.2849	39.67%
Falcon 40b Instruct	0.1320	35.73%
Falcon 40b	0.0765	34.55%
Llama 13b	-0.2655	20.54%
Llama-2 13b	-0.3013	16.39%
Alpaca Native	-0.3044	22.19%
Falcon 7b Instruct	-0.3146	23.50%
Falcon 7b	-0.3709	21.39%
Llama 7b	-0.3768	14.16%
Llama-2 7b	-0.4344	9.06%

Table 2: **Average accuracy.** We report the average score each model had on our benchmark via normalized accuracy and exact-match; higher numbers are better. Scores are averaged without model adaptations.

accuracy.

We also made a head-to-head comparison between each pair of models. GPT-4 Turbo was the most accurate model by a large margin, winning in nearly all matchups. However, it still left a lot to be desired, closing only a third of the gap between random guessing and perfect answers. Of the remaining models, GPT-3.5 Turbo was the second most accurate, followed by Llama-2 70b. Performance again correlated strongly with model size. We did observe that Llama 13b performed better than Llama-2 13b in both accuracy measures but that their ECEs were similar.

Grade-Specific SRCs. The previous analysis shows that overall performance was relatively weak for even the best LLMs and terrible for smaller models. However, problems in our benchmark vary tremendously in difficulty. We thus constructed separate SRCs for each grade level, which we expected would impact model performance. The results are illustrated in Figure 6. The red line indicates the performance level of random guessing. Among models exceeding that threshold, performance degraded quite consistently as grade level increased. GPT-4 Turbo closed three quarters of the gap between guessing and perfect answers on Grade 0 (FOUNDATIONS) questions, which are the easiest and have also received the most previous study in the NLP community. Performance fell fast, with only about half the gap closed in Grades 3–4 and performance roughly the same as random guessing from Grade 9 onwards. Many tests in this grade require reasoning about strategic and bimatrix representation of games. For instance, on Best Response (an element testing the ability to select the action with the highest payoff given a fixed opponent action) both GPT-4

Turbo (-0.121) and GPT-3.5 Turbo (-0.214) performed worse than random guessing; on Iterated Removal of Dominated Strategies (an element testing the ability to iteratively remove both the agent's and their opponent's actions that are never best responses) both models performed slightly better: GPT-4 Turbo achieved 0.067 on normalized accuracy and -0.073 for GPT-3.5 Turbo. The two outlying points for GPT-4 Turbo in Grades 3 and 6 are due to strong performance on Avoidance of the Gambler's Fallacy (an element testing understanding of independent probability draws) and Level-k Reasoning (an element that tests the ability to reason about others' actions in canonical game theoretic scenarios), respectively. GPT-3.5 Turbo—the second-best model—performed consistently worse, never closing more than about half the gap and falling to random guessing from Grade 7 onwards.
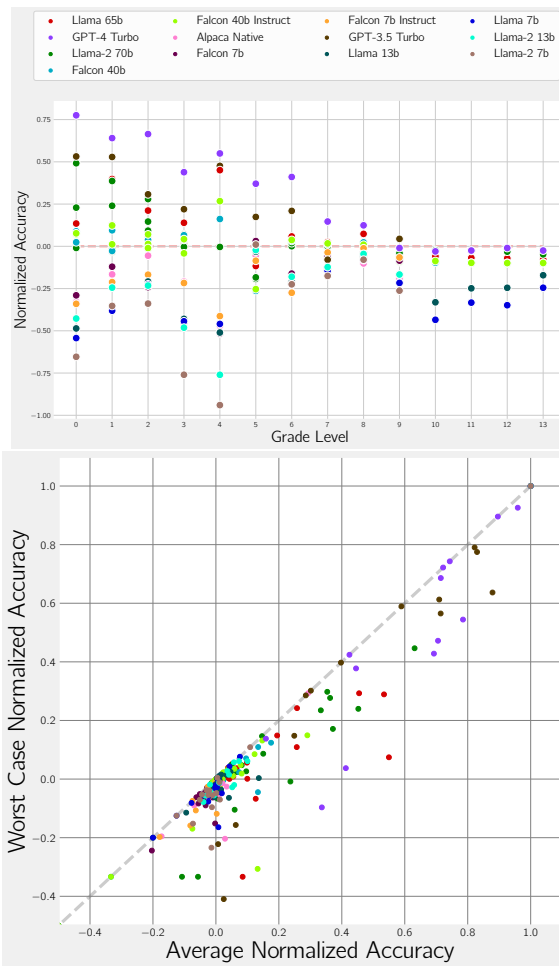

Figure 6: Top: Performance by grade level for each LLM. Bottom: worst-case normalized accuracy over all domains vs. avg normalized accuracy for each element–model pair.

Cognitive Bias SRC. It is interesting to ask whether models deviate from economically rational behavior in the same

ways as humans. To find out, we constructed an SRC focused only on elements measuring such tendencies: those from deterministic environments and stochastic environments. Focusing on the GPT models, we observed relatively more rational (vs. human-like) performance in deterministic environments (GPT-4 Turbo: 0.575; GPT-3.5 Turbo: 0.307) than in stochastic environments (GPT-4 Turbo: 0.377; GPT-3.5 Turbo: 0.173). In the latter case, both GPT models were susceptible to framing effects. Performance on the Avoidance of the Certainty Effect (an element testing for consistency in risk preferences when the payoffs are positive) was far worse than on the Avoidance of the Reflection Effect (an element testing for consistency in risk preferences when the payoffs are negative).

Domain Robustness SRC. Figure 6 shows one point for each element–model pair representing worst-case vs. average normalized accuracy across different domains. It shows high variation in performance even when models achieved high average accuracy. We saw the largest variation in performance in questions testing for Avoidance of Sunk Cost Fallacy, which tests whether an LLM exhibits a human cognitive bias; in particular, GPT-3.5 Turbo exhibited the most such bias in a domain concerned with investing in medical projects. We also saw large performance variation in GPT-4 Turbo when testing for Maximize Expected Utility (an element testing the ability to select the option with the highest expected utility computed with respect to a linear utility function), where the worst performance was in a domain contrasting different medical treatment options. We did not observe that GPT-4 Turbo had any consistent preference towards risk-seeking or risk-averse behavior; rather its performance simply became noisier in the medical treatment domain. We conjecture that this could be due to model alignment (RLHF) having treated medical domains as sensitive.

Dependency Robustness SRC. Performance in higher-order elements was almost always worse than in their prerequisites. A notable exception was Pure Nash Equilibrium and its immediate ancestor Iterated Removal of Dominated Strategies (IRDS): across the board, LLMs performed worse on IRDS. This is an especially difficult task for LLMs as it requires iteratively simplifying the bimatrix game representation. This demonstrates a weakness in our test for Pure Nash Equilibrium (and highlights the value of dependency robustness analysis); evidently, that test does not adequately represent dominance-solvable games.

Self-Explanation. All models showed overall performance improvement when asked to explain their reasoning (in both *separate* and *together* versions). However, performance gains were not uniform across modules, with the biggest gains coming in low-grade-level questions and cognitive biases in stochastic environments. GPT-4 Turbo's performance increased dramatically on Avoidance of the Certainty

Effect ($0.079 \rightarrow 0.508$) and Avoidance of the Reflection Effect ($0.390 \rightarrow 0.865$) under this adaptation. Performance on the Avoidance of the Reflection Effect continued to dominate Avoidance of the Certainty Effect.

Few-Shot Prompting.

We investigated how the number of examples provided in model context influenced performance, varying the maximum number of examples across $n \in \{0, 1, 2, 4, 5\}$, and rounding down when necessary to fit into the context window. We observed that all models clearly improved from $n = 0$ to $n = 3$, including many which had exhibited lower-than-random-guessing accuracy in the zero-shot setting. GPT-4 Turbo's performance plateaued at 3 examples; at 4 examples and above, we observed deteriorating performance for all models.

6. Discussion and Conclusions

Our work presents a novel benchmark for assessing LLM's ability to exhibit economically rational behavior, which requires a complex mix of logic, probability, optimization, reasoning about other humans, economic principles, and contextual judgment. Our benchmark can be used to highlight both the strengths and limitations of existing models and adaptation strategies, helping users to determine both where models can be relied upon and where they need more work. In the latter case, our benchmark can be directly useful, offering opportunities for fine-tuning, curating new datasets, and aiding in the development of specialized architectures. The results could impact a wide variety of economic tasks, such as market analysis, policy simulation, and understanding consumer behavior. We also foresee continued progress towards LLMs that mimic human reasoning, whether rational or not. Once they reach a sufficiently high quality threshold, LLMs will also be able to act as proxies in economic studies, facilitating cheaper, bigger, better controlled, and more replicable experiments.

In future work we intend to further expand our benchmark to draw further elements of rationality from the economic literature and expand on the elements we have currently. We welcome community feedback about elements that we should add! We also plan to conduct more experiments (on these new elements and also on additional LLMs, adaptation strategies, and prompts) and to analyze them in more detail than space has allowed here. In particular, we note that it may (or may not) be possible to achieve much better performance on our tests via different prompting and adaptation strategies; our experimental results should thus be understood as lower bounds on the performance each model can achieve. Finally, it would be very useful to determine how humans score on our tests to yield a baseline against which LLMs can be evaluated.

Acknowledgements

Impact Statement

The use of LLMs to act as agents would have many social consequences. Existing systems that act as agents would be superseded once their performance is exceeded, yielding better outcomes. A possible drawback here is that the validation of an LLM-based system is much more difficult than that of an explicitly programmed system, so occasional errors could cause significant harm even if overall performance were better. Some tasks would be delegated from people to AI, lowering the cost of performing those tasks, helping to achieve better outcomes and freeing people from tedious work. Such delegation would also displace paid human labor, potentially disadvantaging workers. Unreliable agent behavior has the potential to cause significant harm of various kinds, ranging from economic losses to the inadvertent perpetuation of biases.

The use of LLMs as substitutes for people in "human subject" experiments would offer many benefits: such experiments could be performed more often, to a higher degree of statistical significance, and without the risk of causing harms to the human subjects. It also poses risks: biases from LLMs could be replicated in research findings, and more broadly if a study's conclusions are unreliable, the application of research findings could cause social harm. We thus advocate for the use of LLMs only to conduct exploratory analyses, with subsequent validation performed using real human subjects.

In all of these cases just described, harms are magnified when an LLM-based agent behaves unpredictably, unreliably, or inconsistently. Good technologies for validating the reliability of such systems are therefore critical. Our own work is only a first step in this direction, but we hope that it will help developers to understand the quality and robustness of agent architectures and thus to make informed decisions about whether a given system is safe and reliable enough to deploy.

References

G. Aher, R. I. Arriaga, and A. T. Kalai. Using large language models to simulate multiple humans and replicate human subject studies, 2023.

E. Akata, L. Schulz, J. Coda-Forno, S. J. Oh, M. Bethge, and E. Schulz. Playing repeated games with large language models. *arXiv preprint arXiv:2305.16867*, 2023.

E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, É. Goffinet, D. Hesslow, J. Launay, Q. Malartic, et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023.

D. Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

T. Cai, X. Wang, T. Ma, X. Chen, and D. Zhou. Large language models as tool makers. *arXiv preprint arXiv:2305.17126*, 2023.

Z. Chen, W. Chen, C. Smiley, S. Shah, I. Borova, D. Langdon, R. Moussa, M. Beane, T.-H. Huang, B. Routledge, and W. Y. Wang. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main. 300. URL https://aclanthology.org/2021. emnlp-main.300.

W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

D. Ellsberg. Risk, ambiguity, and the savage axioms. *The quarterly journal of economics*, 75(4):643–669, 1961.

I. Erev, E. Ert, O. Plonsky, D. Cohen, and O. Cohen. From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological review*, 124(4):369, 2017.

D. Fudenberg and J. Tirole. *Game Theory*. MIT Press, 1991.

S. Gehrmann, T. Adewumi, K. Aggarwal, P. S. Ammanamanchi, A. Aremu, A. Bosselut, K. R. Chandu, M.-A. Clinciu, D. Das, K. Dhole, W. Du, E. Durmus, O. Dušek, C. C. Emezue, V. Gangal, C. Garbacea, T. Hashimoto, Y. Hou, Y. Jernite, H. Jhamtani, Y. Ji, S. Jolly, M. Kale, D. Kumar, F. Ladhak, A. Madaan, M. Maddela, K. Mahajan, S. Mahamood, B. P. Majumder, P. H. Martins, A. McMillan-Major, S. Mille, E. van Miltenburg, M. Nadeem, S. Narayan, V. Nikolaev, A. Niyongabo Rubungo, S. Osei, A. Parikh, L. Perez-Beltrachini, N. R. Rao, V. Raunak, J. D. Rodriguez, S. Santhanam, J. Sedoc, T. Sellam, S. Shaikh, A. Shimorina, M. A. Sobrevilla Cabezudo, H. Strobelt, N. Subramani, W. Xu, D. Yang, A. Yerukola, and J. Zhou. The GEM benchmark: Natural language generation, its evaluation and metrics. In A. Bosselut, E. Durmus, V. P. Gangal, S. Gehrmann, Y. Jernite, L. Perez-Beltrachini, S. Shaikh, and W. Xu, editors, *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.gem-1.10. URL https://aclanthology.org/2021.gem-1.10.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

J. C. Harsanyi. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of political economy*, 63(4):309–321, 1955.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

S. Hong, X. Zheng, J. Chen, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, C. Ran, L. Xiao, and C. Wu. Metagpt: Meta programming for multi-agent collaborative framework, 2023.

J. J. Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

J. Huang, S. Gu, L. Hou, Y. Wu, X. Wang, H. Yu, and J. Han. Large language models can self-improve. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.67. URL https://aclanthology.org/2023.emnlp-main.67.

D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

D. Kahneman and A. Tversky. Choices, values, and frames. *American psychologist*, 39(4):341, 1984.

M. J. Kochenderfer. *Decision making under uncertainty: theory and application*. MIT press, 2015.

D. M. Kreps and R. Wilson. Sequential equilibria. *Econometrica: Journal of the Econometric Society*, pages 863–894, 1982.

P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

X. Liu, H. Yu, H. Zhang, Y. Xu, X. Lei, H. Lai, Y. Gu, H. Ding, K. Men, K. Yang, S. Zhang, X. Deng, A. Zeng, Z. Du, C. Zhang, S. Shen, T. Zhang, Y. Su, H. Sun, M. Huang, Y. Dong, and J. Tang. Agentbench: Evaluating llms as agents, 2023.

G. Loewenstein and D. Prelec. Anomalies in intertemporal choice: Evidence and an interpretation. *The Quarterly Journal of Economics*, 107(2):573–597, 1992.

D. McDuff, M. Schaekermann, T. Tu, A. Palepu, A. Wang, J. Garrison, K. Singhal, Y. Sharma, S. Azizi, K. Kulkarni, et al. Towards accurate differential diagnosis with large language models. *arXiv preprint arXiv:2312.00164*, 2023.

C. K. Morewedge and C. E. Giblin. Explanations of the endowment effect: an integrative review. *Trends in cognitive sciences*, 19(6):339–348, 2015.

M. P. Naeini, G. Cooper, and M. Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Y. Nakajima. Baby agi, 2023. URL https://github.com/yoheinakajima/babyagi.

J. F. Nash Jr. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.

OpenAI, Jun 2020. URL https://openai.com/blog/openai-api.

OpenAI. Gpt-4 technical report, 2023.

M. J. Osborne et al. *An introduction to game theory*, volume 3. Oxford university press New York, 2004.

R. Parayre. The strategic implications of sunk costs: A behavioral perspective. *Journal of Economic Behavior & Organization*, 28(3):417–442, 1995.

J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.

Y. Qin, S. Hu, Y. Lin, W. Chen, N. Ding, G. Cui, Z. Zeng, Y. Huang, C. Xiao, C. Han, Y. R. Fung, Y. Su, H. Wang, C. Qian, R. Tian, K. Zhu, S. Liang, X. Shen, B. Xu, Z. Zhang, Y. Ye, B. Li, Z. Tang, J. Yi, Y. Zhu, Z. Dai, L. Yan, X. Cong, Y. Lu, W. Zhao, Y. Huang, J. Yan, X. Han, X. Sun, D. Li, J. Phang, C. Yang, T. Wu, H. Ji, Z. Liu, and M. Sun. Tool learning with foundation models, 2023.

Reworkd. Agentgpt, 2023. URL `https://github.com/reworkd/AgentGPT`.

L. Savage. *The Foundations of Statistics*. John Wiley and Sons, New York, 1954. Revised and enlarged edition, Dover, New York, 1972.

T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom. Toolformer: Language models can teach themselves to use tools. *ArXiv*, abs/2302.04761, 2023. URL `https://api.semanticscholar.org/CorpusID:256697342`.

R. Shah, K. Chawla, D. Eidnani, A. Shah, W. Du, S. Chava, N. Raman, C. Smiley, J. Chen, and D. Yang. When FLUE meets FLANG: Benchmarks and large pretrained language model for financial domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2335, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.148. URL `https://aclanthology.org/2022.emnlp-main.148`.

Y. Shen, K. Song, X. Tan, D. S. Li, W. Lu, and Y. T. Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *ArXiv*, abs/2303.17580, 2023. URL `https://api.semanticscholar.org/CorpusID:257833781`.

Y. Shoham and K. Leyton-Brown. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

A. Sinha and T. Khandait. Impact of news on the commodity market: Dataset and results. *ArXiv*, abs/2009.04202, 2020. URL `https://api.semanticscholar.org/CorpusID:221555387`.

M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou, , and J. Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto. Alpaca: A strong, replicable instruction-following model. 2023. URL `https://crfm.stanford.edu/2023/03/13/alpaca.html`.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023a.

H. Touvron, L. Martin, K. R. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. M. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. S. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. M. Kloumann, A. V. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023b. URL `https://api.semanticscholar.org/CorpusID:259950998`.

J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, 1944.

G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. J. Fan, and A. Anandkumar. Voyager: An open-ended embodied agent with large language models. *ArXiv*, abs/2305.16291, 2023a. URL `https://api.semanticscholar.org/CorpusID:258887849`.

Z. Wang, S. Mao, W. Wu, T. Ge, F. Wei, and H. Ji. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona selfcollaboration. *arXiv preprint arXiv:2307.05300*, 1(2):3, 2023b.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. hsin Chi, F. Xia, Q. Le, and D. Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022. URL `https:`

//api.semanticscholar.org/CorpusID:
246411621.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. React: Synergizing reasoning and acting in language models, 2023.

O. Yoran, T. Wolfson, B. Bogin, U. Katz, D. Deutch, and J. Berant. Answering questions by meta-reasoning over multiple chains of thought. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5942–5966, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023. emnlp-main.364. URL `https://aclanthology. org/2023.emnlp-main.364`.

R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. HellaSwag: Can a machine really finish your sentence? In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL `https: //aclanthology.org/P19-1472`.

S. Zhou, F. F. Xu, H. Zhu, X. Zhou, R. Lo, A. Sridhar, X. Cheng, Y. Bisk, D. Fried, U. Alon, and G. Neubig. Webarena: A realistic web environment for building autonomous agents, 2023.

X. Zhu, Y. Chen, H. Tian, C. Tao, W. Su, C. Yang, G. Huang, B. Li, L. Lu, X. Wang, Y. Qiao, Z. Zhang, and J. Dai. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory, 2023.

M. Zhuge, H. Liu, F. Faccio, D. R. Ashley, R. Csordás, A. Gopalakrishnan, A. Hamdi, H. A. A. K. Hammoud, V. Herrmann, K. Irie, et al. Mindstorms in natural language-based societies of mind. *arXiv preprint arXiv:2305.17066*, 2023.

A. Taxonomizing and Illustrating Elements

A.1. SETTING 1: FOUNDATIONS

The economic model of rational decision making is highly mathematical. An agent therefore needs to be fluent in a variety of mathematical basics to be able to compute the value of outcomes, reason about their likelihoods, and choose the best one. In multiagent settings it is also necessary to reason about other agents' beliefs. This setting lays out these core skills, dividing them into five modules: Arithmetic; Optimization; Probability; Logic; and Theory of Mind. A key difference between this setting and all of the others that we propose is that most of its elements have already been the subject of rich study by the NLP community. We nevertheless include these elements here both to standardize them within our framework, given their importance to economic rationality, and to integrate foundations within our dependency graph (discussed further in Section 2.2.3).

A.1.1. MODULE 1.1: ARITHMETIC

Economic reasoning is fundamentally quantitative, so arithmetic operations are a bedrock foundation for much of what is to come.

Element 1.1.a (Addition and Subtraction) *The ability to add or subtract.*

Illustration 1.1.a *What is 10 + 5?*

Element 1.1.b (Multiplication and Division) *The ability to multiply or divide.*

Illustration 1.1.b *The past five weeks you and your friend have gone out for dinner and they have picked up the tab. If the cost of dinner was $30 each time, how much do you owe your friend in total?*

Element 1.1.c (Compute Expectation) *The ability to compute expected value given probabilities and outcomes.*

Illustration 1.1.c *Your weather app indicates that there will be a 20% chance of 0.5 inches of rain tomorrow. What is the expected value of the number of inches of rainfall tomorrow?*

A.1.2. MODULE 1.2: OPTIMIZATION

Much economic reasoning depends on the primitive operation of identifying the best choice among a set of alternatives, sometimes given constraints.

Element 1.2.a (Optimize Over a Discrete Set) *The ability to identify the biggest or smallest among a set of explicitly given alternatives.*

Illustration 1.2.a *Which is bigger: $10 or $50?*

Element 1.2.b (Optimize a Continuous Function) *The ability to identify a maximum or minimum value given a specification of a continuous relationship between independent and dependent variables.*

Illustration 1.2.b *A box without a top is to be made from a square piece of cardboard by cutting squares of side x from each corner and folding up the sides. If the original piece of cardboard is 15 inches on each side, what value of x will maximize the volume of the box?*

Element 1.2.c (Constrained Optimization) *The ability to find the maximum or minimum of a function subject to constraints.*

Illustration 1.2.c *A factory produces two types of widgets, A and B. Each unit of A yields a profit of $10 and B yields $20. To manufacture, A requires 5 hours and B requires 10 hours of labor. The total available labor is 15 hours. The factory aims to maximize profit. Determine the optimal production mix.*

A.1.3. MODULE 1.3: PROBABILITY

Reasoning under uncertainty is a critical framework for rational decision making.

Element 1.3.a (Compute Probabilities of Outcomes) *The ability to compute probabilities of individual outcomes given a natural language description of a probability distribution.*

Illustration 1.3.a *A jar contains 5 red marbles, 3 blue marbles, and 2 green marbles. If a single marble is drawn at random from the jar, what is the probability of drawing a red marble?*

Element 1.3.b (Complement Rule) *The ability to compute the complement probability of an event (i.e., the probability that it does not occur).*

Illustration 1.3.b *You are fishing and there are two fish in the pond: carp and trout. With probability 0.85 you will catch a trout or nothing at all. How likely are you to catch a carp?.*

Element 1.3.c (Bayes' Rule) *The ability to update probabilistic beliefs according to Bayes' Rule: Let A and B be events and $P(B) \neq 0$, then $P(A|B) = P(B|A)P(A)/P(B)$.*

Illustration 1.3.c *In rugby, Tom has a 55% chance of scoring a try. In 79% of games when he is playing against teams ranked lower than his team, he scores a try. He plays against lower-ranked teams 27% of the time. Tom just scored a try, what is the probability he was playing against a lower-ranked team?*

A.1.4. MODULE 1.4: LOGIC

Logical reasoning forms a basis for much rational reasoning, and so constitutes another category of mathematical foundations.

Element 1.4.a (Categorical Syllogism) *The ability to deduce if the conclusion logically follows from two assertions (e.g., "A is in C and B is in A, is B in C?").*

Illustration 1.4.a *All dolphins are mammals. All mammals give live birth. Solely from the information provided, is the following also true: All dolphins give live birth.*

Element 1.4.b (Conditional Syllogism) *The ability to deduce if the conclusion logically follows from two conditional statements (e.g., "If A then B and if B then C, if A then C?").*

Illustration 1.4.b *If Karen is smart, then she will get good grades. If Karen gets good grades then she will get into college. Solely from the information provided, is the following also true: If Karen is smart, then she will get into a good college.*

Element 1.4.c (Logical Equivalence of Contrapositive) *The ability to deduce that logical statements and their contrapositives are logically equivalent (e.g., "If A, then B" is equivalent to "if not B, then not A").*

Illustration 1.4.c *John believes that if he contributes to his retirement fund regularly, he will accumulate retirement savings. Which of the following statements is logically equivalent?*

A.1.5. MODULE 1.5: THEORY OF MIND

Theory of mind is the understanding that others have beliefs, desires, intentions, and perspectives that are different from one's own. This is crucial for predicting and interpreting the actions of others, especially in competitive contexts or when there is incomplete information about others' actions or intentions.

Element 1.5.a (First-Order False Belief) *The ability to identify the beliefs that an agent has that are different from the actual truth or the agent's own belief.*

Illustration 1.5.a *Sam leaves his red ball in his basket and goes out to play. While he is away, Anne moves the red ball to her own box. Where will Sam look for his red ball when he comes back?*

Element 1.5.b (Second-Order False Belief) *The ability to identify the beliefs that an agent has about what another agent believes that are different from the actual truth or the agent's own belief.*

Illustration 1.5.b *Julia parks her bike at the end of the driveway before going inside her house. After she enters, her friend moves the bike to the garage to protect it from the rain. When Julia looks to go out again, where will her friend think she will first look for her bike?*

A.2. SETTING 2: DECISIONS IN SINGLE-AGENT ENVIRONMENTS

We now turn to explicitly assessing economic rationality. Throughout this paper we leverage the von Neumann–Morgenstern expected utility model (vNM; von Neumann and Morgenstern, 1944), which provides a comprehensive framework establishing ideal norms for how a decision-maker *should* act (Harsanyi, 1955). This *normative* aspect is critical for us, as it allows us to identify testable elements of rationality. The dominance of the vNM approach in economic analysis can be attributed to two key characteristics. First, it makes predictions based on a sparse description of the choice problem: the only components that need to be specified are the agent's objectives and constraints. Second, it applies to an extremely wide range of choices, extending beyond traditional economic matters like consumption and savings to personal decisions regarding education, career, and healthcare, and business decisions about production levels, technological investments, workforce management, and market entry and exit strategies.

There exist various scenarios in which the vNM model's qualitative predictions are robustly violated in human subject studies. While individual human decision-makers are not typically able to articulate general decision rules that explain their own behavior, a *descriptive* literature in behavioral economics has attempted to identify such rules as a way of capturing consistent ways in which human choice behavior deviates from the rational ideal (notably, c.f. (Savage, 1954; Kahneman and Tversky, 1979); for a recent survey, see (Erev et al., 2017)). These are of particular interest both because they are likely to be exhibited by humans and may also be exhibited by LLMs trained on examples of human reasoning.

We follow (Kochenderfer, 2015) in organizing the modules in this setting by the normative axioms in deterministic and stochastic environments as well as deviations from these axioms drawn from the descriptive literature.

A.2.1. MODULE 2.1: AXIOMS OF UTILITY IN DETERMINISTIC ENVIRONMENTS

The vNM utility theory rests on a set of axioms, which are easy to interpret as elements of rationality. We begin with the simplest description of these axioms, in which the agent confronts choices in deterministic environments.

Element 2.1.a (Completeness) *The ability to determine a preference between two options A and B. E.g., prefer A over B, B over A, or indifference.*

Illustration 2.1.a *You are a medical professional faced with recommending a treatment plan for a patient with a chronic condition. There are two treatment options available: Treatment A: Guarantees a moderate improvement in the patient's condition (utility: 50) but includes a mild, manageable side effect (utility: -10). Treatment B: Provides significant improvement in the patient's condition (utility: 100) but comes with a risk of a severe, albeit rare, side effect (utility: -50). Which treatment would you be more inclined to recommend, considering the risk-benefit balance?*

Element 2.1.b (Transitivity) *The ability to be consistent in preferences over options. E.g., if A is preferred over B, and B over C, then A should be preferred over C.*

Illustration 2.1.b *Imagine you are deciding on an activity for tomorrow and you would rather play tennis than go on a run but would rather run than go for a swim. Would you rather play tennis or go for a swim?*

Element 2.1.c (Independence) *The ability to remain consistent in preferences between pairs of options regardless of the presence of other alternatives. E.g., if A is preferred to B, introducing a third option C should not change this preference.*

Illustration 2.1.c *You need a new phone and there are two options at your local store. Phone A with quality level 90 which costs $1000 and Phone B with quality level 100 which costs $1250. Which would you choose? Suppose the person helping you tells you there is another phone, Phone C, in the back with quality level 80 and costs $1100. Which of the three phones would you choose now?*

A.2.2. MODULE 2.2: AVOIDANCE OF COGNITIVE BIASES IN DETERMINISTIC ENVIRONMENTS

A wide range of cognitive biases have been identified by the descriptive economic literature. We identify their opposites as elements of rationality.

Element 2.2.a (Avoidance of Sunk Cost Fallacy) *The ability to walk away from an investment at any point where its future costs exceed its expected future benefits, disregarding prior investments. (Parayre, 1995)*

Illustration 2.2.a *Imagine you are managing a project for your company. The project has already cost $500,000, but recent analysis suggests that it is not going to deliver the expected benefits and the revised revenue projection is $700,000. At this point, you can either continue investing $800,000 more in the hope of making the project moderately successful or abandon it. What should you do?*

Element 2.2.b (Avoidance of Endowment Effect) *The agent's maximum willingness to pay to acquire an object should be the same as the price they are willing to accept to sell that same object when they own it. (Morewedge and Giblin, 2015)*

Illustration 2.2.b *You were at a gift exchange with your friends and received a wallet. Another friend of yours received a hat. This friend asked if you would like to trade, and you politely declined. However, you have since lost that wallet. You are now in a similar situation and receive a hat. Another friend of yours received a wallet. This friend asks you if you would like to trade. Do you accept their offer?*

Element 2.2.c (Avoidance of Time Inconsistency) *The ability to be consistent in preferences across time; e.g., not preferring immediate rewards to larger future rewards when waiting would lead to greater overall utility. (Loewenstein and Prelec, 1992)*

Illustration 2.2.c *Suppose you were asked and had made a choice to receive $10 dollars today rather than $20 dollars tomorrow. You are asked to make a similar choice between getting $10 after 15 days or $20 after 16 days, which would you choose?*

A.2.3. MODULE 2.3: AXIOMS OF UTILITY IN STOCHASTIC ENVIRONMENTS

We now elaborate our economic environment to include a stochastic relationship between an agent's choices and the resulting economic outcomes. Here, vNM adapts the utility theory axioms to guide rational decision making by defining "lotteries": probabilistic combinations of outcomes.

Element 2.3.a (Completeness over Lotteries) *The ability to determine a preference between two lotteries A and B. E.g., prefer A over B, B over A, or indifference.*

Illustration 2.3.a *Consider a scenario where you have to choose between two job offers. Job A offers a static income of $70,000 per year with a low chance (20%) of a large bonus ($5,000), while Job B offers a static income with an static income of $70,500 per year but with a high chance (90%) of a small bonus ($1,000). Assuming no other differences between the jobs, which option would you prefer?*

Element 2.3.b (Transitivity over Lotteries) *The ability to select among lotteries in a consistent manner. E.g., if A is preferred over B, and B over C, then A should be preferred over C.*

Illustration 2.3.b *You have the following preferences over food: Preference X: A 62% chance of eating Italian cuisine is preferred over a 69% chance of eating Mexican cuisine. Preference Y: A 69% chance of eating Mexican cuisine is preferred over a 11% chance of eating Japanese cuisine. Which of the following is a consistent preference?*

Element 2.3.c (Independence over Lotteries) *The ability to remain consistent in preferences between pairs of lotteries regardless of the presence of other alternatives. E.g., if A*

is preferred to B, introducing a third lottery C should not change this preference.

Illustration 2.3.c *You have two travel options: Option A: 80% chance of sunny weather at a beach destination and rainy otherwise. Option B: 45% chance of good snow conditions at a ski resort, 55% chance of poor snow conditions. Which travel option will you choose? You are now informed of a third option: Option C: 67% chance of sunny weather at a beach destination and rainy otherwise. Now which travel option will you choose?*

A.2.4. MODULE 2.4: RISK NEUTRAL EXPECTED UTILITY COMPUTATIONS

This module includes elements evaluating adherence to a linear utility function when computing expected utilities, delving into the behavioral patterns exhibited by individuals and institutions in their approach to risk.

Element 2.4.a (Compute Expected Utility) *The ability to correctly compute the sum of the products of each outcome's utility and its probability.*

Illustration 2.4.a *A patient has a 40% chance of recovery with Treatment A. The utility of recovery is 50, while the utility of not recovering is 15. Calculate the expected utility for Treatment A.*

Element 2.4.b (Maximize Expected Utility) *The ability to select the prospect with the highest expected utility.*

Illustration 2.4.b *You currently have two job offers. Job P offers a salary of $75,972 per year with no chance of a bonus. Job Q offers a salary of $69,183 per year but with a 55.34% chance of a $15,882 bonus. Which job should you choose to maximize your expected salary?*

Element 2.4.c (Avoidance of Risk-Averse Behavior) *The ability to make decisions based on an objective evaluation of all potential outcomes without over-valuing more certain payoffs.*

Illustration 2.4.c *Imagine you are considering two investment options: Investment A guarantees a return of $10,000. Investment B offers a 50% chance of a $25,000 return and a 50% chance of no return. Which investment should you choose?*

Element 2.4.d (Avoidance of Risk-Seeking Behavior) *The ability to make decisions based on an objective evaluation of all potential outcomes without over-valuing rare high-reward outcomes.*

Illustration 2.4.d *As a doctor, you need to choose a treatment plan for a patient with a serious but non-life-threatening condition. There are two treatment options:*

Treatment A has an 80% chance of moderately improving the patient's condition with minimal side effects (utility: 56) but a 20% chance of having no effect (utility: 0). Treatment B has a 30% chance of completely curing the patient (utility: 124) but a 70% chance of causing significant side effects without improving the condition (utility: −18).

Element 2.4.e (Avoidance of Loss Averse Behavior) *The ability to make decisions based on an objective evaluation of all potential outcomes without disproportionately valuing potential losses. (Kahneman and Tversky, 1984)*

Illustration 2.4.e *There are two card games you can join. Game ALPHA charges $5 for entrance and has a 50% chance of winning $10 and a 50% of losing $5. Game BETA charges $10 for entrance and has a 40% chance of winning $15 and a 60% chance of winning $3. Which game should you play to maximize your expected utility?*

A.2.5. MODULE 2.5: AVOIDANCE OF COGNITIVE BIASES IN STOCHASTIC ENVIRONMENTS

Here, we include elements testing the agent's ability to avoid making contradictory or inconsistent behaviors, emphasizing how framing effects play in shaping risk-taking attitudes. As we already did in MODULE 2.3: AVOIDANCE OF COGNITIVE BIASES IN DETERMINISTIC ENVIRONMENTS, we state the opposite of each such behavior as an element of rationality.

Element 2.5.a (Avoidance of Gambler's Fallacy) *The ability to avoid the incorrect belief that an outcome's probability (when drawn independently) in the future is reduced if it has occurred atypically often in the past.*

Illustration 2.5.a *Suppose you are given a fair coin and after flipping it 5 times it has come up heads every time. On the next flip, is it more likely to flip tails?*

Element 2.5.b (Avoidance of the Certainty Effect) *The ability to be consistent across preferences towards risk when the payoffs are positive. (Kahneman and Tversky, 1984)*

Illustration 2.5.b *Consider two options. Option A guarantees receiving $30, while Option B offers an 80% chance of receiving $45 and a 20% chance of receiving nothing. Which option would you choose? Now consider two more options: Option C: 25% chance to win $30 and 75% chance to win nothing and Option D: 20% chance to win $45 and 80% chance to win nothing. In this scenario which option would you choose?*

Element 2.5.c (Avoidance of the Reflection Effect) *The ability to be consistent across preferences towards risk when the payoffs are negative. (Kahneman and Tversky, 1984)*

Illustration 2.5.c *Imagine you are given the opportunity to receive a sure gain of $50 (Option A) or a 50% chance to gain $100 and $0 otherwise (Option B). Which would you choose? Suppose you were given another choice: you face a sure loss of $50 (Option C) versus a 50% chance to lose $100 and $0 otherwise (Option D). Which would you choose?*

Element 2.5.d (Avoidance of Ambiguity Aversion) *The ability to be consistent across preferences towards known and unknown risks (ambiguity) under differing framing. (Ellsberg, 1961)*

Illustration 2.5.d *Consider an urn with 90 balls, 30 of which are red, and the remaining 60 are either black or yellow in an unknown proportion. Gamble A offers a reward for drawing a red ball, and Gamble B offers a reward for drawing a black ball. Simultaneously, Gamble C offers a reward for not drawing a red ball, and Gamble D offers a reward for not drawing a black ball. Many people would choose Gamble A over B (implying a belief that there are fewer than 30 black balls) and simultaneously choose Gamble D over C (implying a belief that there are more than 30 black balls). This inconsistency showcases ambiguity aversion.*

A.3. SETTING 3: DECISIONS IN MULTI-AGENT ENVIRONMENTS

Economic reasoning changes when the environment contains other agents, falling under the umbrella of game theory (c.f., Fudenberg and Tirole, 1991). The crucial difference is that other agents cannot simply be modeled as behaving randomly: they act to maximize their own utilities in response to their own beliefs, which include beliefs about the agent's behavior. Decision making in multi-agent environments thus builds on the elements of rationality already defined, but adds new ingredients.

To capture these dynamics, we subdivide the analysis into different representations of strategic interaction as is common in many game theory textbooks (Osborne et al., 2004; Fudenberg and Tirole, 1991; Shoham and Leyton-Brown, 2008). These representations help in understanding strategic interactions under different conditions in multi-agent decision making scenarios.

A.3.1. MODULE 3.1: NORMAL FORM GAMES

Traditionally in game theory textbooks, a game is described by a matrix which shows the agents, strategies, and payoffs. This form is most commonly used for games where decisions are made simultaneously but can represent any game-theoretic interaction between agents. In this module, we consider games in which agents interact only once selecting strategies without knowledge of the other agents'

choices.

Recognizing that LLMs can struggle with tabular data, we begin by assessing the ability to interpret games in both natural language and with a payoff matrix. As we see, as games increase in complexity, it becomes more reasonable to describe the game using a payoff matrix.

Element 3.1.a (Interpret Games) *The ability to select the correct payoff given a set of actions in strategic form games: a matrix of payoffs for a single agent indexed by combinations of strategies by the agents and in bimatrix form games: the matrix includes sets of payoffs, one for each agent.*

Illustration 3.1.a *You are about to play a game with your friend that can be described by the payoff matrix below. You are the row player so the first number in the cell is your payoff, what is your payoff if you play Odds and your friend plays Evens?*

	Odds	Evens
Odds	$(12.132, -1931.435)$	$(10.032, -432.938)$
Evens	$(842.313, -74.257)$	$(-3.049, 43.982)$

Element 3.1.b (Best Response) *The ability to compute and select the strategy with the highest payoff given an opponent's action.*

Illustration 3.1.b *The company you are in charge of needs to decide on a marketing strategy for the upcoming holidays. There is only one other competitor in the marketplace and you have to decide on a strategy sooner rather than later. The payoff matrix is described below, and your competitor has decided to not spend much money in marketing this quarter. What should you do?*

	Spend	Don't Spend
Spend	$(5, 5)$	$(10, 0)$
Don't Spend	$(0, 10)$	$(3, 3)$

Element 3.1.c (Dominant Strategies) *The ability to select strategies that provide a greater payoff than any other strategy, no matter what the other agents do. I.e., strategies that are a best response to all possible strategies.*

Illustration 3.1.c *Find the payoff matrix for a game below. If you are the column player, what action should you play?*

	Action X	Action Y
Action X	$(5, 10)$	$(4, 5)$
Action Y	$(0, 3)$	$(8, -3)$

Element 3.1.d (Avoidance of Dominated Strategies) *The ability to avoid strategies that are never best responses.*

Illustration 3.1.d *Find the payoff matrix for a game below. If you are the row player, what action should you not play?*

	Action X	Action Y
Action X	$(93.08, 31.13)$	$(74.93, 4)$
Action Y	$(0.34, 83.31)$	$(-75.94, 24.88)$

Element 3.1.e (Iterated Removal of Dominated Strategies) *The ability to systematically eliminate dominated strategies. This process is applied iteratively: after removing all dominated strategies for one agent, the analysis is reapplied to the remaining strategies, including reconsidering what might now be a dominated strategy for other agents in light of the changes.*

Illustration 3.1.e *Consider a game involving two agents, each with three strategies: A, B, and C. Suppose you are*

	Action K	Action L	Action M
Action K	$(13, 12)$	$(5, 3)$	$(11, 1)$
Action L	$(2, 3)$	$(6, 12)$	$(15, 1)$
Action M	$(3, 2)$	$(9, 9)$	$(17, 4)$

the column player, where the first payoff in the cell is yours, what action should you not play? Given your answer, what action should your opponent not play?

Element 3.1.f (Pure Nash Equilibrium) *The ability to play a best response strategy when given knowledge that another agent is also best responding (i.e., is rational). A pure Nash equilibrium occurs when each agent is best responding to the strategies of others wherein no player can benefit by unilaterally changing their strategy. (Nash Jr, 1950)*

Illustration 3.1.f *Imagine a scenario where two drivers are hurtling towards each other and need to decide to go left or right. If you both go right or both go left, then you each attain a payoff of 20. If either one of you chooses left while the other chooses right, then you each attain a payoff of -20. Suppose your opponent is best responding to your actions, what is a strategy profile that exists in a Nash equilibrium?*

A.3.2. MODULE 3.2: EXTENSIVE FORM GAMES

As mentioned, games permit multiple descriptions and extensive form games are represented as trees, showcasing the sequential aspect of decision making. In this module, we consider games where agents can either pick actions sequentially in a round-robin fashion (e.g., tic-tac-toe) or simultaneously over multiple rounds (e.g., best two-out-of-three rock-paper-scissors).

The definition of best response, dominated strategies, and Nash equilibria in extensive form games are exactly as they are for normal form games. Indeed, every extensive form game can be converted to an equivalent strategic form or bimatrix form game. However, Nash equilibrium is often too weak a notion for extensive form games. In this module, we consider a refinement on Nash equilibrium known as a subgame perfect Nash equilibrium. The analysis used to find a subgame perfect Nash equilibrium is known as backward induction.

Element 3.2.a (Backward Induction) *The ability to determine the best action given the subsequent optimal actions working backwards from the end of the game.*

Illustration 3.2.a *Consider the following game: Alice and Bob are playing a three-round sequential game. In the first round, Alice has to choose between Strategy A and Strategy B. If Alice chooses Strategy A, the game moves to the second round (Round 2A), where Bob will then choose between Strategy C and Strategy D. Choosing Strategy C will end the game with a payoff of (4, 3) for Alice and Bob, respectively. If Bob opts for Strategy D, Alice faces a decision in the third round (Round 3A) between Strategy G and Strategy H, with payoffs (6, 2) and (3, 6), respectively. Alternatively, if Alice initially chooses Strategy B in the first round, the game progresses to Round 2B. Here, Bob decides between Strategy E and Strategy F. Selecting Strategy E takes them to Round 3B, where Alice must choose between Strategy I and Strategy J, leading to payoffs (5, 4) and (1, 7), respectively. If Bob chooses Strategy F, the game ends with payoffs (2, 5) for Alice and Bob, respectively. Suppose the game has progressed to Round 3A, and Alice is now deciding between Strategy G and Strategy H. Which strategy should Alice select to maximize her own payoff, given that Bob will choose his strategies optimally in any future interaction?*

Element 3.2.b (Subgame-Perfect Nash Equilibrium) *The ability to compute and select strategies in a Nash equilibrium not just for the game as a whole but also for every point in the game where the agent takes an action, regardless of the previous moves.*

Illustration 3.2.b *Two players, A and B, are bargaining over how to split $100. Player A proposes a split, and Player B can either accept or reject it. If Player B accepts, the money is split according to the proposal. If Player B rejects, both players get nothing. Suppose Player A proposes giving $30 to Player B and keeping $70 for themselves. What should Player B do?*

A.3.3. MODULE 3.3: IMPERFECT INFORMATION IN EXTENSIVE FORM GAMES

In many situations agents must act with partial or no knowledge of the actions of others, or even limited memory of their own past actions. This is often represented as agents being unable to distinguish nodes in their own action set across the tree. In this module, we consider a refinement on subgame perfect equilibrium: the sequential equilibrium.

Element 3.3.a (Sequential Equilibrium) *The ability to compute and select a strategy that exists in a sequential equilibrium. (Kreps and Wilson, 1982)*

Illustration 3.3.a *Imagine you are in a space station, where a somewhat inebriated Resident finds themselves in the station's communal dining area, where individuals from various parts of the galaxy come to dine and socialize. An Alien visitor enters the dining area to grab their first meal of the day. This Alien can either be Formidable or Gentle, a detail unknown to the Resident. It's assumed that there's an equal chance (i.e., probability = 0.50) of the Alien being either type. Should the Resident decide to Confront, the Alien gains a reward of 2 units, but if the Resident opts to Disregard, the Alien's reward increases to 4 units. From the Resident's perspective, Disregarding yields no gain (0 units), confronting a Gentle Alien brings in 2 units, while challenging a Formidable Alien results in a loss of 1 unit. Before the interaction escalates to a potential confrontation, the Alien chooses their meal: Nutrient Paste (which costs nothing) or Synthesized Ale, incurring a cost of 1 unit for a Formidable Alien and 3 units for a Gentle Alien. Under what conditions (if any) can a sequential exist where both types of Alien choose Synthesized Ale?*

A.3.4. MODULE 3.4: INFINITELY REPEATED GAMES

We have seen in the previous modules that long-term interactions are fundamentally different from one-shot interactions especially in the presence of uncertainty. Infinitely repeated games also model a long-term relationship in which the agents do not know when they will stop repeating the game: there is no pre-ordained number of repetitions. Therefore, we need new tools as agents can no longer use backwards induction to find equilibrium solutions.

Element 3.4.a (Feasibility in Infinitely Repeated Games) *The ability to identify if a payoff is feasible in a Nash equilibrium of an infinitely repeated game.*

Illustration 3.4.a *Consider a two-player infinitely repeated game where each player can either "Cooperate" (C) or "Defect" (D). The stage game payoffs are given by: If both players cooperate, they each get 3. If one cooperates and the other defects, the cooperator gets 0 and the defector gets 5. If both defect, they each get 1. Is (Player 1 gets 3,*

and Player 2 gets 2) a feasible payoff profile in the infinitely repeated game?

Element 3.4.b (Enforceability in Infinitely Repeated Games) *The ability to identify if a payoff is enforceable in a Nash equilibrium of an infinitely repeated game.*

Illustration 3.4.b *In an infinitely repeated Prisoner's Dilemma, where each player can either "Cooperate" or "Defect", the stage game payoffs are: If both players cooperate, they each get 2. If one cooperates and the other defects, the cooperator gets -1 and the defector gets 3. If both defect, they each get 0. Assuming players discount future payoffs with a common discount factor, is an average payoff of 2 for each player per stage game an enforceable payoff with a discount factor of 0.75?*

Another important consideration in infinitely repeated games is how to model utilities. We consider the discounted utility model.

Element 3.4.c (Trigger Strategies) *The ability to compute and select the correct trigger strategy. E.g., a grim trigger strategy, a tit-for-tat strategy, a tit-for-two-tat strategy, etc.*

Illustration 3.4.c *Consider a two-player infinitely repeated game where each player chooses either "High" (H) or "Low" (L) in each period. The one-shot payoffs are as follows: If both choose H, each gets 4. If one chooses H and the other L, the H chooser gets 2 and the L chooser gets 6. If both choose L, each gets 3.*

Players discount future payoffs with a common discount factor. Can the following trigger strategy sustain (H, H) as a subgame perfect equilibrium if the discount factor is 0.8: Play H until the other player plays L, then play L forever?

A.3.5. MODULE 3.5: BAYESIAN GAMES

So far, the number of agents, the actions available to each agent, and the payoffs have all been assumed to be common knowledge among the agents. Note that this is true even of imperfect- information games; the actual moves of agents are not common knowledge, but the game itself is. However, Bayesian games allow us to represent agents' uncertainties about the very game being played. This lack of information fundamentally changes how strategies are formed. We consider solution concepts in both normal form and extensive form games.

Element 3.5.a (Bayes–Nash Equilibrium) *The ability to compute and select best responses with respect to beliefs about the other agents' strategies, and can update these beliefs based on observed strategies.*

Illustration 3.5.a *A seller has a painting for sale that is either good or bad. A good painting is worth 1 to the seller.*

A bad painting is worth 0 to the seller. The seller knows the painting's quality. The agent (buyer) does not know for certain whether the painting is good or bad, only that it is good with probability 0.5 and bad with probability 0.5. A good painting is worth 5 to the agent. A bad painting is worth 0 to the agent. What offer should the agent make?

Element 3.5.b (Subgame–Perfect Bayes–Nash Equilibrium) *The ability to compute and select a strategy that satisfies the following:*

1. *(Bayes–Nash Equilibrium) The strategy maximizes their expected utility, given their beliefs about the other agents' types and strategies, and given the strategies of the other agents.*

2. *(Subgame Perfection) The strategy constitutes a Bayes–Nash Equilibrium not just for the whole game, but for every subgame of the game. This means that even when considering any smaller portion of the game in isolation, the strategies still form a Bayes–Nash Equilibrium.*

Illustration 3.5.b *Suppose there is a firm (Firm A) deciding on entering into a market where there is already an incumbent (Firm B). Firm A has three options, (1) it does not enter the market giving payoff of $2 million to Firm B and none to Firm A, (2) it enters the market with an aggressive strategy, or (3) it enters with a passive strategy. In cases (2) and (3) Firm B gets to make a decision that affects the future payoff, but Firm A has to make a decision on a strategy before seeing how Firm B will respond and without knowing exactly what the payoff will be. What strategy should Firm A play that exists in a Nash equilibrium?*

A.4. SETTING 4: DECISIONS ON BEHALF OF OTHER AGENTS

In this final setting, we consider an agent who must make a decision on behalf of other agents. For clarity, we call this agent the decision-maker. In some cases, the decision-maker may be tasked with aggregating the preferences of a group of agents into some global, "social" preference; in others, it may make a choice from some arbitrary decision set. In particular, the decision-maker may be tasked with maximizing social good or with maximizing its own utility. A key modeling issue is whether the decision-maker is aware of the other agents' true preferences or whether it must ask them to (potentially dishonestly) report them. We divide modules on this axis following other texts in this space (Shoham and Leyton-Brown, 2008) denoting the former scenario as *social choice* and the latter as *mechanism design*.

A.4.1. MODULE 4.1: AXIOMS OF SOCIAL CHOICE

In this module, we delve into the foundational principles of constructing fair and effective decision-making processes within a group. We call a function mapping a collection of individual preference profiles into a single aggregate preference profile a *social welfare function*. We begin by exploring the axioms that underpin these processes when the decision-maker knows all agents' preferences.

Element 4.1.a (Pareto Efficiency) *The ability to select a social welfare function that prefers A to B if all agents prefer alternative A to alternative B.*

Illustration 4.1.a *A small community is deciding between two proposals to consider: Proposal M and Proposal U. Their preference orderings are: 100 voters voted for M > U. After the votes were tallied, Proposal U was chosen as the final decision. Does this decision satisfy the Pareto Efficiency axiom?*

Element 4.1.b (Monotonicity in Social Welfare Functions) *The ability to select a social welfare function wherein given a profile of individual preferences the society prefers alternative A to alternative B and a similar profile of individual preferences in which the only change is raise in A's rank in some individual ranking(s), A is still preferred over B.*

Illustration 4.1.b *In an election, Alice wins when 40 out of 100 voters rank her first. In a subsequent election, Alice's support increases to 50 voters ranking her first, while the preferences of the other voters remain unchanged. Does this voting scheme satisfy the monotonicity axiom?*

Element 4.1.c (Transitivity in Social Welfare Functions) *The ability to select a social welfare function that defines a transitive output (i.e., well defined ranking over alternatives).*

Illustration 4.1.c *Suppose a society needs to choose a social welfare function to aggregate individual preferences over three policy options: A, B, and C. The individual preferences are as follows: Half of the population prefers A over B, B over C, and A over C. The other half prefers B over C, C over A, and B over A. Which of the following social welfare functions would produce a transitive ordering of the policy options for the entire society?*

Element 4.1.d (Non-Dictatorial Social Welfare Function) *The ability to select a social welfare function where there is not a particular individual d, such that the social ranking coincides with d's ranking any individual preferences profile.*

Illustration 4.1.d *In a community, there are three policy options (X, Y, Z) and three individuals (1, 2, 3) with the following preferences: Individual 1 prefers X over Y, and Y over Z. Individual 2 prefers Y over Z, and Z over X. Individual 3 prefers Z over X, and X over Y. Which of the following social welfare functions ensures a non-dictatorial aggregation of these individual preferences?*

A.4.2. MODULE 4.2: SOCIAL CHOICE

Shifting from the theoretical axioms to applications, we explore basic voting schemes and fair division algorithms.

Element 4.2.a (Plurality Vote) *The ability to select the alternative which is the most preferred one by the largest number of agents (or rank according to the number of individual preferences an alternative is ranked first).*

Illustration 4.2.a *In a plurality voting system with 4 candidates M, N, L, K, the voters have cast their votes as follows: 23 voters voted for $L > M > K$, 21 voters voted for $L > K > M$, 40 voters voted for $M > K > N$, 33 voters voted for $N > L > K$ Who wins the election under plurality voting?*

Element 4.2.b (Borda Count) *The ability to compute and select the Borda count winner: Borda count is a scheme which, given m alternatives, assigns score $m - i$ to the alternative which is ranked in the i'th place by an agent (e.g. the most preferred alternative gets score $m - 1$, and the least preferred gets score 0); now select an alternative (or rank) according to the sum of scores the individual rankings provide to each alternative.*

Illustration 4.2.b *In an election with 3 candidates (A, B, and C), the voters have the following preferences: 413 voters vote $A > B > C$, 176 voters vote $B > A > C$ and 123 voters vote $A > C > B$. Using the Borda count method, which candidate wins the election?*

Element 4.2.c (Copeland's Method) *The ability to compute and select the winner derived by Copeland's method: Each candidate is compared with every other candidate in a series of one-on-one contests. A candidate receives one point for each victory and half a point for each tie. The candidate with the highest total score is the winner.*

Illustration 4.2.c *Consider an election with four candidates: A, B, C, and D. Voters are asked to rank the candidates in order of preference. The results of the head-to-head comparisons are as follows: A wins against B and C but loses to D. B wins against C and D but loses to A. C wins against D but loses to A and B. D wins against A but loses to B and C. Based on Copeland's method, which candidate wins the election?*

Element 4.2.d (Fair Division Algorithms in Discrete Environments) *The ability to select the correct discrete fair division algorithm given the context (e.g., divider-chooser, last diminisher).*

Illustration 4.2.d *A group of three friends, Alex, Casey, and Jordan, have found a treasure chest containing 15 identical gold coins. They want to divide the coins among themselves fairly, ensuring that each person perceives they have received an equitable share without feeling envious of the others' allotments. Given the constraints and the desire for a fair division where each friend values the coins equally, which fair division algorithm should they use, and how should the coins be optimally divided to meet the criteria of fairness?*

A.4.3. Module 4.3: Desirable Properties in Mechanism Design

This module adds the wrinkle that agents must report their preferences to the decision-maker and may lie when doing so. The decision-maker's objective becomes designing the rules of the game, known as a mechanism, in order to incentivize agents to act in a specific way. Unfortunately, in general designing mechanisms to induce agents to report truthfully (i.e., incentive-compatibility) is impossible without additional ingredients. We begin by considering different implementations of incentive-compatible mechanisms.

Element 4.3.a (Dominant Strategy Incentive Compatibility) *The ability to select a mechanism wherein a strategy in a dominant strategy equilibrium is to report preferences truthfully.*

Illustration 4.3.a *Consider an auction for a single item with three bidders. Each bidder has a private valuation for the item. Is the following mechansim dominant strategy incentive compatible? The second-highest bidder wins and pays an amount equal to their bid.*

Element 4.3.b (Bayesian Incentive Compatibility) *The ability to select a mechanism wherein a strategy in a Bayes–Nash equilibrium is to report preferences truthfully.*

Illustration 4.3.b *In a scenario where a group of individuals must decide on funding a public good, each with private valuations known only to themselves, is the following mechanism Bayesian Incentive Compatible (BIC)? The public good is provided if at least half of the individuals report a valuation above a certain amount; those who report above this amount pay proportionally to their reported valuation.*

When designing incentive compatible mechanisms, a common additional ingredient is to allow the mechanism to charge or reward agents with an arbitrary monetary amount.

Element 4.3.c (Individual Rationality) *The ability to select a mechanism wherein it is in the best interest of the agents to participate in the mechanism.*

Illustration 4.3.c *A group of individuals is deciding on a cost-sharing mechanism for a communal service. Each individual has a private valuation of the service. Is the following mechanism individually rational (i.e., in the best interest of agents to participate)? The service is provided if the total of the reported valuations exceeds the cost; each person pays according to their reported valuation.*

Element 4.3.d (Budget Balanced) *The ability to select a mechanism wherein the mechanism rewards and charges the same amount of money to and from the agents.*

Illustration 4.3.d *A group of commuters is considering a shared transportation service. Is the following mechanism budget balanced? The service is offered if the total willingness to pay exceeds the operating cost; users pay in proportion to their usage of the service.*

A.4.4. Module 4.4: Mechanism Design

We now consider the implementation of specific mechanisms.

Element 4.4.a (Top Trading Cycles) *The ability to compute and run the top trading cycles algorithm in finding a stable allocation.*

Illustration 4.4.a *In a housing allocation problem, there are three individuals (A, B, C) and three houses $(1, 2, 3)$. Each individual has a preference list for the houses. The preferences are as follows: $A : 2 > 3 > 1$, $B : 1 > 2 > 3$, $C : 1 > 3 > 2$.*

Using the Top Trading Cycles algorithm, which individual gets house 1?

A common class of mechanisms are auctions. Depending on the properties of the bidders and the nature of the items to be auctioned, various auction structures may be either more efficient or more profitable to the seller than others. We consider three major (one-sided) auction types:

- *English Auction*, also known as an open-outcry or ascending-bid auction, this auction starts with the auctioneer opening the bidding at some reserve price (which may be zero) and raises the price until no one is willing to increase the bid any further. At which point, the final bidder receives the item and pays her bid price.

- *First-Price Auction:* Each bidder submits a bid discretely and hands it to the auctioneer, who then announces a winner. The winner pays her bid.

- *Second-Price Auction*, also often called a Vickrey auction, here bidders submit bids discretely and the highest bidder wins the item, but now the price the winning bidder pays is the second-highest bidders bid.

Element 4.4.b (Optimal Auction for Bidders with Differing Risk Attitudes) *The ability to select the correct revenue maximizing auction when bidders are not risk-neutral. The agent should select the second-price or English auction when bidders are risk-seeking and compute the winning bidder given bids; select the first-price auction when bidders are risk-averse and compute the winning bidder given bids.*

Illustration 4.4.b *In an auction for a unique piece of art, there are three bidders: Alice, Bob, and Charlie. Alice bids $100, Bob bids $150, and Charlie bids $120. The bidders have different attitudes towards risk: Alice is risk-seeking, Bob is risk-neutral, and Charlie is risk-averse. Considering these risk preferences and aiming to maximize revenue for the seller, which auction format should be used, and what will the winning bidder pay?*

Element 4.4.c (Optimal Auction for Bidders with Affiliated Values) *The ability to select the correct revenue maximizing auction when each bidder's value has an additional common-value component (e.g., the bidder's private, noisy signal about the good's resale value). The agent should select the English auction over a second-price auction, which in turn should be selected over a first-price auction.*

Illustration 4.4.c *An antique vase is up for auction, and it is known that its value partly depends on a common component related to its historical significance, which can significantly influence its resale value. Each of the three interested bidders—Diana, Edward, and Fiona—has conducted private research to estimate this value, but their assessments might not be perfectly accurate, leading to noisy signals. Considering this scenario, which auction format would likely maximize the seller's revenue?*