Benchmarking economic reasoning in artificial intelligence models (Preliminary: estimation in progress)*

Douglas K. G. Araujo[1]

[1]Bank for International Settlements, douglas.araujo@bis.org

Abstract

A structural model of economic reasoning combines information filtering, knowledge association and logic attribution to correctly answer prompts. The model is adapted for economics but can also be adapted to other sciences. The reasoning model, along with insights from social economics, informs the design of an economic reasoning benchmark that (a) evolves over time in a non-trainable way; (b) measures reasoning rather than accuracy (ie, disregarding luck and stochastic parroting) and (c) breaks down reasoning results into its constituent components. An accompanying training dataset will be publicly available at publication time, and interested users can submit task proposals. NB: Empirical estimations are being run. Keywords: Economic reasoning, benchmark, large language models, artificial intelligence. JEL codes: C45, C69, C88, C59

"Machines will be capable, within twenty years, of doing any work a man can do" - Herbert Simon, AI pioneer, in 1965

1 Introduction

Large language models (LLMs), in particular those classified as generative artificial intelligence (gen AI), increasingly support use cases in finance and economics (Korinek (2023)), including in central banks (Araujo et al. (2024)). These models are tested for their ability to reason, often boasting seemingly incredible results: for example, OpenAI's GPT-4 boats human-like performance across tests of reasoning and more than 80% correct results in academic and professional micro- and macroeconomics tests (Achiam et al. (2023)). Still, even such advanced models can fail miserably when it comes to reasoning: for example, an advanced model can correctly solve a logical puzzle requiring reasoning about higher order knowledge, only to fail when irrelevant details are

*This work represents my opinion and not necessarily that of the BIS. Note: estimation of the benchmark results in progress; results will be included in the first complete draft.

changed (Perez-Cruz and Shin (2024)). Building on results such as this, this work discusses how to systematically measure *economic* reasoning, combining literatures on economic thought and on computer science about gen AI benchmarking. In practical terms, the task at hand is to come up with a benchmark task for economic reasoning, a testing mechanism to measures in a comparable way the level of economic reasoning of an artificial intelligence (AI) model. This task is of first-order importance given the break-neck speed of evolution of LLMs (Yang et al. (2023)) and their potential risks (Danielsson, Macrae, and Uthemann (2022)).

Such benchmark tasks are crucial for the comparison of the abilities of AI models over time and in the cross-section. Results of benchmark tasks are now a staple of the evaluation of LLMs by developers when releasing a model, highlighting its evolution compared to previous versions and to the main peers. Third-party organisations also compile leaderboards with running results that allow the general public to keep track of the most performant models.[1] Benchmark tasks are useful because that they provide a comparable metric on which to track the state-of-the-art for the particular abilities that each task measures. Usually, this metric is the percentage of correct answers. However, even if specific tasks have evolved over time to become more challenging, a key challenge is to separate correct answers due to probabilistic association or "stochastic parroting" (Bender et al. (2021)) from those that are the result of a reasoning process.

This paper proposes a working model of reasoning that can underlie an empirical benchmark task: when confronted with a prompt Q,[2] an AI is said to *reason* correctly if it responds with an answer α that simultaneously (a) interprets the prompt, identifying the relevant information for the task and filtering away everything else by ignoring or abstract away irrelevant details, (b) associates Q with any relevant existing commonsense knowledge θ to answer the question, and (c) applies logical relations such as deduction and induction to Q and θ to arrive at the correct answer. Formally, each answer is defined as a non-parametric function of the following steps: information filtering $\phi = f(Q)$, knowledge curation $\kappa = k(\phi, \theta)$ and logic attribution $\lambda = l(\kappa, Q)$, $\alpha = A(\lambda, Q)$, where A is the function that returns the correct answer from the prompt Q given a successful implementation of the component steps. Each of those three steps above are sequential, and depend on the successful completion of the previous one. The goal is for this model to be simple and intuitive. Throughout the paper, I assume AIs respond to their best ability, meaning that they would reason instead of probabilistically choosing an answer $\tilde{a} = \arg\max_a L(a|Q)$.

The empirical version of this prompt-answering model is

$$\hat{\alpha}(q) = \hat{\phi}\, \hat{\kappa}|(\phi = \hat{\phi})\, \hat{\lambda}|(\phi = \hat{\phi}, \kappa = \hat{\kappa}), \tag{1}$$

[1] Commonly followed leaderboards include LMSYS's ChatbotArena and Huggingface's Open LLM Leaderboard.

[2] Usually this will be a string of text, but more recent models can take multi-model content, ie a combination of text, images, videos, sounds, etc.

where $\phi = \mathbf{1}\prod_i^{M_f} f(q+\epsilon_i^f)$ and similarly for κ with k and λ with l, M_f, M_k and M_l are the number of variations ϵ introduced in the seed question q that seek to identify the model's interpretation, knowledge association and logic attribution respectively; the hat denomination points to empirically estimated versions. This model is estimated by assessing answers from the same AI model to multiple versions of seed questions $q \in \mathbf{Q}$, and $\hat{\alpha}$ is only considered to be correct when all of the relevant variations for the same question are answered correctly - in other words, the AI model has evaded "banana skins" that try to trick it into revealing lack of information filtering, spurious knowledge association or faulty logics. The key idea is to leverage insights from the social economics literature and create identifying variation in the questions q presented to AI models, adapted from how this is done with human subjects (Stantcheva (2023)). The benchmark result is then $R = M_{\mathbf{Q}}^{-1}\sum_{q\in\mathbf{Q}}\hat{\alpha}(q)$, where $M_{\mathbf{Q}}$ is the number of different seed questions. By a similar token, each of the three steps can be measured separately, building on their empirical identifications $R_{j\in(f,k,l)} = M_j^{-1}\sum_{q\in\mathbf{Q}}\hat{j}(q)$. Note that identification of κ and λ need the sequential conditionality on the previous steps ϕ and ϕ, κ respectively in order to be identified.[3]

This model can be used as a general abstraction for an AI reasoning ability, but two practical adaptations in Q can make it a model for economic reasoning more specifically. First, the scope of topics that are included in Q should ideally focus on issues of significance to economics. At its most essential form, testing for economic reasoning is the same as probing if the model is able to think in terms of logical operators on information that is of relevance to economics. However, this is subjective because economic thought constantly evolves. At the same time, including only "classical" economics carries a high risk of using material that is containd in the training set of AI models. Second, even for each given topic, the types of questions considered relevant in economics are specific. Both of these issues are dealt with in practice by using recent published academic work as source material to construct seed questions.[4] These sources contain content whose topic and research questions are by definition of interest to the economics field, and moreover have the advantage of being novel by design, creating a natural check for the ability of AI models to generalise reasoning.

This benchmark task is, by design, more difficult than others due to its sequential conditioning. Is it really necessary to have such a hard-to-score task? While evidence amounts that large language models (LLMs) cannot perform advanced reasoning, at least not when fine-tuned or with access to external reference material or sophisticated math funcionalities, it is important to have a challenging reasoning benchmark because the latter two possibilities are increasingly being used. Both fine-tuning on specific data and plugging LLMs into sources of knowlege (as in retrieval-augmented generations, or RAGs) or with plugins

[3]The benchmark result could also be compiled as a metric that is not subject to the false "emerging abilities" results (Schaeffer, Miranda, and Koyejo (2024)), for example the Brier score (Brier (1950)).

[4]Similar adaptations could be pursued for other fields.

as Wolfram Alpha or Mathematica might provide models with some reasoning ability. For this reason, it is important to have a robust way of measuring the reasoning abilities of AI models. Another argument is that part of AI models' reasoning performance out of the box is due to its limitation to train only on available written language only (Browning and LeCun (2022)). Of course, this dataset is a very limited snapshot of human experience, even with massive data compilations. However, current and future models will be able to leverage a significant part of the other non-text data (eg, video, audio, pictures) and thus could reasonably attempt to achieve better reasoning. This provides another reason to maintain a high bar in reasoning benchmarks.

This benchmark evaluation also addresses a poignant issue for the economics profession: the lack of publicly available data about how these benchmarks are created and any, and tested. For example, Achiam et al. (2023) are not clear about the academic and profession tests on micro- and macroeconomics that are used amongst various other tests to measure GPT-4's performance in those fields. Conversely, various other benchmark tasks do have a publicly available methodology and even evaluation interface, which greatly facilitates the engagement with model developers, general users and third-party model evaluators.

A major inspiration in the design of the questions and how they can generate identifying variations is the social economics literature. A key reference is Stantcheva (2023). The idea here is that the design of the questionnaire itself can elicit responses that allow for insight into non-observable traits such as reasoning. Many of the insights of this literature carry over naturally to the machine space.[5]

1.1 Literature

This work builds on, and seeks to expand, three general literature streams. More technical aspects of this work are based on specific literatures that are discussed within each section.

The first body of works is on benchmarking tasks for AI models. A substantial body of work creates and discusses model benchmarking in general; a voluminous and well-organised compilation of references is Storks, Gao, and Chai (2020). Interested readers are strongly encouraged to that paper for space considerations, while selected benchmarks are described in Section 2.

Secondly, this paper draws from insighs in the more general AI reasoning literature. As other parts of AI development, it is informed and inspired by neuroscience as well (Hassabis et al. (2017)). An early and influential contribution is the Chinese room experiment by Searle (1980) and its resulting arguments that AI could not reason by itself. The influential works of Bubeck et al. (2023) and

[5]Actually testing whether LMs *do not* parrot or "organically" exhibit biases or other behaviours that are assumed to be exclusively human would be an interesting line of research.

Wei, Tay, et al. (2022) hint at acquisition of advanced capabilities by large-scale language models such as GPT-4, although Bubeck et al. (2023) also point to many instances where reasoning breaks. Schaeffer, Miranda, and Koyejo (2024) present evidence that these "emerging abilities" that come with scale are actually a spurious by-product of the choice of metrics to emasure these abilities. Mitchell and Krakauer (2023) summarises the disagreement in the AI academic and practitioner fields as to whether AI models have some form of understanding, and by implication, potentially also reasoning. Wei, Wang, et al. (2022) claim that writing the prompt in a way that offers chain-of-thought examples improves the reasoning abilities of LLMs, although this was later demonstrated to be as generalisable (Dziri et al. (2024), Prystawski, Li, and Goodman (2024)). Browning and LeCun (2022) argue that AI models trained on written language alone will never be able to reason.

A nascent literature on the evaluation of language models in economic settings. An early foray into questions related to AI's ability to conduct economic reasoning is due to Parkes and Wellman (2015). But their angle is more on how AIs can be used to estimate synthetic economic agents - machina oeconomicus - ideal versions of purely rational agents, rather than on the measurement and the implications of AIs acquiring economic reasoning abilities. In any case, Parkes and Wellman (2015) see economic reasoning as the ability to understand and solve complex game-theoretical environments (eg, the poker example). Mei et al. (2024) do an extensive comparison of personality traits from the behaviour of ChatGPT with human behaviour in games that require cooperation, finding that its performance is consistent with humans, and when it deviates the AI models tend to behave in the altruistic and cooperative than the mass distribution of humans. Interestingly, ChatGPT responds differently to different formulations of the same situation. In contrast to Mei et al. (2024), this paper and its empirical counterpart are more generall, and discuss reasoning as a whole. Another contrast to that paper is that the current benchmark is focused on reasoning ability only, not personality. Perez-Cruz and Shin (2024) illustrate the brittleness of a leading AI's reasoning, which has markedly lower performance when trivial details in the prompts are different. Similarly, Korinek (2023) report (in his Chat 23) that results from a technical prompt in economics are reasonable but also brittle, with answers changing when prompt wording changes or even simply if the tasks are re-ordered.

1.2 The challenges with reasoning: a simple illustration

This section builds on the intuition that in true reasoning, the result should be robust to minute perturbations, ie the model is a constant function over the domain of the input. Formally, both $\mathbf{V}(X) = y$ and $\mathbf{V}(X + \epsilon) = y$ for an infinitesimal ϵ, even as $\mathbf{V}(X') = y', y \neq y'$, ie it should be a locally constant function. This implies the derivative with respect to the input prompt is zero. Using as an approachable example the simplest possible neural network, the logistic regression $\mathbf{N}(x) = \sigma(Wx + b)$, such robustness further implies that

$\frac{d\mathbf{N}}{dx} = \sigma(Wx+b)(1-\sigma(Wx+b))W = 0$. Because W cannot be a zero vector in a functioning network that is responsive to its inputs and $\sigma(Wx+b)(1-\sigma(Wx+b)) = 0$ has no solution because neither term is 0 or 1 in a sigmoid function with finite inputs, the neural network cannot be a constant function. This extremely simplified example, which holds for recursive architectures of similarly simple layers, does not bode well for the robustness of results given small perturbations in the input prompt.

2 Existing benchmarks

There are numerous benchmark tasks, and their number is increasing with the sophistication of newer AI models (Storks, Gao, and Chai (2020)). In fact, there is nowadays a whole "benchmark safari" (to which this paper is a contribution). Below I summarise important characteristics of the main ones in each type, to further illustrate why a new benchmark is needed and more specifically how one geared towards economics can be constructed.

2.1 Reference resolution

- Measures ability to identify referents (ie, linguistic mentions)
- Complicated task due to ambiguities, requires intense commonsense knowledge
- Needs typically handcrafted data
- Datasets tend to be smaller than other benchmarks
- Examples: Winograd, WinoGrande (Levesque, Davis, and Morgenstern (2012))
- Browning and LeCun (2023) see the failure of these challenges to present a serious challenge to modern LLMs as a sign that lingusitic tests are unlikely to be good proxies for reasoning abilities.

2.2 Question answering

- Mixes language processing and reasoning skills
- Some benchmarks can be solved without deep understanding, only from linguistic context
- But many Q&As require external knowledge
- Highlighted example: ARC (Clark et al. (2018))
- Other examples: MCTest, RACE, NarrativeQA, MCScript, ProPara, MultiRC, ARCT, SQuAD 2.0, CoQA, QuAC, and many others

2.3 Textual entailment

- Entailment: a directional relationship between a statement and a hypothesis where a typical person would infer the hypothesis to be true given the statement

- Some benchmarks also test ability to recognise contradiction
- Requires substantial commonsense knowledge
- Highlighted example: SherLlic (Schmitt and Schütze (2019))
- Other examples: RTE, conversational entailment, SICK, SNLI, MultiNLI, SCITail, and others

2.4 Intuitive psychology

- Inference of emotions and intensions conditional on description of behaviour
- Requires substantial commonsense knowledge
- Highlighted example: SocialIQA (Sap et al. (2019))
  - results of increasing performance in other tests suggest some level of commonsense knowledge is learnable through data
- Other examples: Triangle-COPA, ROCStories, Event2Mind

2.5 Plausible inference

- Associated with abductive reasoning
- Hypothetical, intermediate certainty or even uncertain conclusions based on a limited context
- Requires linguistic, common and commonsense knowledge
- Highlighted examples: SWAG (Zellers et al. (2018)), HellaSWAG (Zellers et al. (2019))
- Examples: COPA, CBT, ROCStories, LAMBADA, JOCI, CLOTH, ReCoRD, AlphaNLI

2.6 Multiple tasks

- Some of them include some tasks requiring some economics-specific knowledge
- Examples: bAbI, Inference Is Everything, DNC, GLUE, SuperGLUE
- More recently, BIG Bench (Srivastava et al. (2022))
  - BIG = "Beyond the Imitation Game"

2.7 Expert tasks

- Geared towards field-specific knowledge (eg, law, medicine)
- Publicly available tests
- Usually require substantial common and commonsense knowledge
- No economics-specific benchmark that I know of

3 Limitations of existing benchmarks

The result from existing benchmarks is largely, if not completely, directly related to the number of questions correctly answered. However, this measures only

the model's ability to answer correctly, *not necessarily* its reasoning capabilities. The latter are part of a latent state space sitting between the input prompt and the answer. More concretely, for an input prompt X, which includes a question and any necessary explicit information, the language model is a function $\mathbf{M}$ that maps it to a given response: $\mathbf{M} : X \rightarrow y$. In order to show that it is done by reasoning, we need tests (and more specifically, measurements) that convey some information about the inner workings of this function.

As identified by Srivastava et al. (2022), another limitation is that many benchmarks are not created from questions that are first created by experts. CONTINUE…

4 A model of reasoning

The model presented in the Introduction is arguably simplistic, but is fundamented on considerable body f literature that combines sensory pathways, knowledge combination and the execution of logical. This section goes into more detail about the choices in the reasoning model.

4.1 Information filtering

Reasoning should be robust to irrelevant input or to changes in minutiae that are not crucial for the logic relations to follow. This result follows as a result of the efficient coding hypothesis in physiology (Barlow et al. (1961), Olshausen and Field (1996), Loh and Bartulovic (2014)), which postulates that sensory pathways need to reduce the dimensionality of inputs in living organisms. This insight is itself a biological and social observation ispired by Shannon's (1948) landmark information theory. The AI literature has of course known this of years, and it inspired the concept of attention (Larochelle and Hinton (2010), Mnih et al. (2014)), which later inspired the self-attention and ultimately the game-changing transformer architecture (Vaswani et al. (2023)).

The first step, filtering the received impulses (ie, the prompts), involve correctly judging what is relevant and what is not relevant. This is similar for example to how the brain receives an incredible amount of sensory inputs but chooses to focus only on those that are more relevant instead of being overwhelmed with everything else, an observation that has inspired dimensionality-reduction algorithms (eg, isometric mapping, or IsoMap, by Tenenbaum, Silva, and Langford (2000) describes how to find global optima while also defining the (much lower) degrees of freedom in a high-dimensional input).

Appropriate perception should understand that the information of relevance to understanding a problem is actually much lower-dimensional. Beyond the biological (and more specifically neural) requirement for inputs to focus on the most important aspects, this observation is also consistent with a bedrock of the machine learning literature, the manifold hypothesis. This hypothesis is based on the theoretical and empirical observations that most real-world realisations are

high-dimensional embeddings of much lower-dimensional manifolds.[6] Cayton (2008) offers an early review of the main algorithms for estimating empirically the underlying manifold. Bengio, Courville, and Vincent (2013) discusses extensively the idea of representation learning (and its various techniques, mostly unsupervised), which can be seen as manifold learning. However, they might also approximate the wrong manifold or not have a single solution to a same manifold (Lee (2023)).

The requirement for perception modulation is also aligned with Prystawski, Li, and Goodman (2024)'s finding that reasoning abilities in LLMs require "locality" in concepts until a final link between a prompt and its final answer (if far away) can be achieved. In a way, this is also similar to the small world network model (Kleinberg (2000)). In humans, the literatures on rational inattention and neuroeconomics (Sims (2003), Caplin, Dean, and Leahy (2022), Dean and Neligh (2023), Hébert and Woodford (2021)) models human information processing as subject to a cost that grows with the informational content, which is a closer representation of how people actually process information. In linguistics, the information bottleneck literature discusses how ideas are compressed into words by a trade-off between lexicon complexity vs accuracy (Zaslavsky et al. (2018)), and more recently also consistency (Chen, Futrell, and Mahowald (2023)).

4.2 Knowledge association

Assuming a prompt has been correctly parsed, the reasoning mechanism must now match it with the relevant knowledge, which can come from the prompt itself or from commonsense knowledge.

Knowledge can be linguistic, common or commonsense (Davis and Marcus (2015)). Mahowald et al. (2023) uses insights including from neuroscience to distinguish the first type of knowledge with the latter two (grouped as "functional" knowledge), and argue that LLMs have essentially mastered the former while still having a spotty record on the latter. For example, LLMs learn grammar, semantic, hierarchical structures, abstractions and constructions that provide a realistic linguistic knowledge.

Bransford and Johnson (1972) show in experiments that contextual knowledge (in this case akin to commonsense) are essential for proper understanding in humans. The first experiments tested understanding by subjects of a grammatically correct, non-metaphorical passage that required an unusual and very specific, but highly relatable image as context for proper understanding. Note that these characteristics of the passage (correct, non-metaphorical) and of the con-

[6] For example, Pope et al. (2021) study the intrinsic underlying dimensionality of the manifold of image datasets and find them to be significantly lower than their observed dimensionality. In practice, inputs can even be said to be *union of manifolds* (as verified by Brown et al. (2022) with image datasets in an exercise similar to the one by Pope et al. (2021)), which means that each manifold has its own intrinsic dimensionality that is not forced upon the other manifolds. This perspective affords flexibility in the interpretation of identifying variations because they don't necessarily need to probe the same dimensions at each task.

text (unusual but relatable and easy to understand) both contribute to isolate the identification of this exercise in the aspect of whether contextual knowledge is required. [7]

Which type of knowledge to match to the prompt? One way of seeing this is through the lends of a query in a knowledge graph. Kleinberg (1999) distinguishes specific and broad queries, each giving rise to one problem: that of a scarcity of correct answers and abundance of correct answers, respectively. Further, Kleinberg (1999) offers the fundamental ideas of *authority*. Similarly, Kleinberg (1999) acknowledges that measuring the authority level of a node in a knowledge graph from explicit information alone (what he calls *endogenous* measure). On the contrary, even so much as using strings from the query itself might mislead answers due to an abundance of other sources that are based on the string and a scarcity of correctly authoritative sources that use the string. Interestingly, while the principal eigenvector of the square of the adjacency matrix offers the weights of authoritativeness especially for broad queries, the non-principal eigenvectors can offer insights into the authoritativeness of more specialised queries, and also due to their negative entries offer authorities of different perspectives (ie, weighting pros and cons).

So the lower-rank approximation of the knowledge graph should vary with how broad/specific a query is, and also with the level of pros/cons required. In Kleinberg (1999), the authoritativeness measures comes from the eigenvectors of $A^T A$, with the principal eigenvector being the used as a broad authoritativeness metric, and the nth eigenvectors for $n > 1$ as the more specific, and potentially discordand, authorities. This implies that having question variations that refer to opposite views in the prompt should also help probe the level of knowledge (given the theoretical model that it is a function of existing knowledge graph) versus response flipping due to probabilistic association with the terms introduced to describe the pros and cons.

4.3 Logic attribution

Once a prompt has been correctly parsed to focus only on relevant information and the necessary knowledge has been associated with it, an AI model can assign specific logic relationships between different components of a prompt and the existing knowledge. Commonly studied logic relationships are induction, deduction, analogy, abstraction, abduction(Walton (2014), Johnson-Laird, Khemlani, and Goodwin (2015), Davis and Marcus (2015), Dziri et al. (2024)), among others. Naturally, an essential condition for appropriate use of logical relationships is that the inputs ϕ and κ are correct; otherwise, even the purest application of logic would either lead to a failure or to a correct answer by chance.

Definition 4.1 (Logic relationship)**.** A logic relationship is a …

[7] Interestingly, the same paper also demonstrates that prior knowledge itself is not necessarily readily available but needs to be "activated". This is not further discussed in the context of this paper as it is not a mechanism necessary for measuring reasoning abilities in AI models.

Definition 4.1 introduces the idea of logic relationships. One example…

Similar to many other social disciplines, economics requires the analytical judgment referred to by Robbins (1932) in the analyses of events as a basis to extrapolate and predict, and this has a bearing on how economic reasoning should be benchmarked. Economic inference depends primarily on articulating unobservable quantities, theorecised and estimated on the basis of observable measures. This is unlike other major disciplines. For example, in human and veterinary medicine, all physiological and pathological variables of clinical importance are observable, even if that is not yet technologically feasible today. In the medical sciences, theoretical models merely fill in the gaps in the absence of a technologically feasible complete measurement. In contrast, many economically relevant quantities are latent variables that cannot by definition be observed, and always require a model applied to data to be estimated, implicit or not.

A quantitative test for economic reasoning must take into account that selecting a correct answer in an economics question through reasoning will always depend on an unobserved transformation of the information received and the existing knowledge. AI models may also happen to choose the correct answer from either luck of through simple token probability. It is easy to see why a correct answer selected by chance is not informative about the reasoning abilities of a model. Since reasoning should be robust to perturbations in the input data, is locally complete, meaning that an LM that can correctly deduce that A implies B is also able to understand that A' does not imply B, or that A does not imply B'. In other words, a logic relationship that appears to be correct but whose obvious corolary is not achieved by an LLM cannot be said to have been reasoned in the first place.

4.4 Reasoning iself as a low dimensional operation

Why is the information filtering important for identifying reasoning? Since proper reasoning needs to be insensitive to unimportant details, and the vector of changes depends on logical relationships between components, the set of all "reasonable" constructions is not obtained at random but reflects a lower-dimensional, underlying space.

Gilboa et al. (2014) argues that economic reasoning works by creating positively wrong but conceptually useful representations of reality, even when economics is studying particular cases. A marked characteristic of such models is their preference for simplicity, a theme also explored by Gilboa and Schmeidler (2010), who study the matching of economic theories to empirical data, generalising the evaluation of how reasonable a theory is through a combination of their likelihood (or goodness-of-fit) with a penalising factor for their complexity. Intuitively, this simplicity in reasoning is suggestive of the manifold hypothesis in reasoning as well. Gilboa, Minardi, and Wang (2023) sees rationality, or reasoning, also as a robustness to trivial detail.

A related, more empirically applicable perspective, is the possibility to identify

models partially using random sets, ie abstracting away from point identification to situations where the data is incomplete or is described as an interval (Beresteanu, Molchanov, and Molinari (2012)). In other words, combining assumptions with available data to inform the estimation (even if partial) of a parameter of interest (Molinari (2020)), which are situations that could arise for example when the data originates from a lower-dimensional manifold but is observed as an embeddeding in higher dimensions.

All of these insights above, together with the points made in each specific step, lend credence to the intuition that the reasoning process itself entails some dimensionality reduction, in line with the existence of evidence that some real-world data are in fact realisations of a manifold (Pope et al. (2021)).

4.5 Identification

Due to the sequential nature of Equation 1 and the desirability of measuring an AI model's performance across the three reasoning steps, identification of the model obtains from a combination of analysis to find how the values can be estimated empirically to be one or zero. This then leads to the creation of identifying variation during the question generating process as described in Section 6.

For each question q, ϕ only depends on the specific prompt at hand (ie, $q + \epsilon_f$) and is independent of the two other values. When estimated from the question data it is also independent of the AI model's knowledge set θ. Once an estimate for $\hat{\phi}$ is available, it can be used to condition κ along with the AI's fixed knowledge set, which does not change across observations of q or $q + \epsilon_k$, and therefore is absorbed away since the estimation is done for each AI. Finally, $\hat{\phi}$ and $\hat{\kappa}$ are used to condition the estimates of λ based on $q + \epsilon_l$.

Theorem 4.1 (Identification of reasoning abilities)**.** *If…*

Proof. By induction. □

5 Reasoning about economics

The model above allows us to estimate reasoning while also breaking down some of its components to better understand them. For example, we can estimate any errors in reasoning into an issue with information filtering, knowledge association and logic attribution. The goal of this section is to clarify that the working definition of economic reasoning used in this work is very broad and is not meant to assign one or another form of reasoning with greater weight, let alone one economic school of thought over the other. Rather, the goal is to reflect over time also slower-moving features of our profession, such as the ebb and flow of schools of thought (Pribram (1953)).

Gilboa et al. (2022) distinguish between three types of inquity in economic theory: economics itself (analysis of economic phenomena), development of economic methods (the development of analytical tools needed to study economic phenomena) and the methodology of economics (the research/scientific endeavour in economics, including but not limited to theory).[8] Naturally, they would all be in scope of this exercise.

Another insight into *economic* thinking is from the thought experiments first introduced by Marshall (1890) - the ceteris paribus idea - and then later Ragnar Frisch and Trygve Haavelmo, more recently elaborated in more detail and more generally by J. Heckman and Pinto (2015), including the important distinction between correlation and causation. Marschak and Andrews (1944) start their influential paper by acknowledging that economists can't conduct experiments (although that has been relaxed somewhat, it still remains the case at least in macroeconomics). Reasoning in economics can be taken as exploring the latest space through models, or thought experiments, an insight generalised by J. J. Heckman and Pinto (2023). Pribram (1953) adds to argue that even the areas of focus - the answer to *what matters* - changes over time.

Other sciences, in contrast, might have more concrete definitions of reasoning. For example, the human and veterinary medicines require hard data to conduct reasoning. In its absence, due for example to technological, biological or economic constraints, professionals need to theorise and especially rely on abductive reasoning. But if one day all of these constraints were relaxes and medical and veterinary doctors had access to all possible data about their patients, they would only need to reason with respect to the physiological or pathological relationship between these observed variables. Economists, in contrast, will always need to conduct thought experiments and think in latent terms to conduct any type of meaningful economic reasoning.

6 Empirical estimation

Recall from Equation 1 that estimation requires an evaluation of several variations of the same seed question q for each component. These variations form the empirical data. Questions vary with respect to their adversarial aspect; it is this variation within each question that allows the empirical estimation of the effects associated with interpretation or with knowledge. Most of the variations are originally those tested in Alzahrani et al. (2024). The variation in response between the questions within each task will comprise the evaluation of the actual reasoning capabilities. As alluded to before, the variations are organised into those that measure the stability of a response to adversarial interpretation answers, and those that measure the stability across the knowledge dimension. In practice, each task has hundreds of different q. These groups are described in more detail next.

[8] In fact, Gilboa et al. (2022) even allude to the blurred lines between economics and the philosophy or sociology of economics. I don't go ino these differences here.

6.1 Adjusted adversarial filtering

The key idea in this section is to use the adversarial filtering process proposed by Zellers et al. (2018) and Zellers et al. (2019), adapted to the current application. A generative model creates an initial set of incorrect responses to each video. This is then fed to the adversarial filtering routine, which is executed iteratively. First, the source data is split into training and testing sub-samples. An AI model is trained on the training sub-sample and used to identify those in the testing data that are easier to correctly answer. Those easy alternatives are then taken out of the sample, and newly generated alternatives replace them. The process is repeated until stabilisation.

A few adjustments are needed for the current case: this process occurrs separately for each of the seed questions q and each of the three steps. This results in a collection of $\mathbf{W}_q^{(1)} = \epsilon^f, \epsilon^k, \epsilon^l$ for each q. The iterative filtering then proceeds as described above, until the performance of adversarial filters have degraded to an arbitrarily low point.

$

Algorithm 1 Algorithm for finding server indices using OFG

▷ %comment: servers[] contains the index of servers whose data rate are sorted in descending order%
servers[]= index(of all servers)
serverIndex[]=servers[0..K]
linearlyIndependentServerIndex[]=0
$[Z] \leftarrow 0$
for $i = 0$ to $serverIndex.length$ **do**
▷ %comment: find the equation corresponding the serverIndex from the mapping at the File Server%
$eqn = equation(serverIndex[i])$
▷ %comment: try insert equation into Z using OFG%
end forend for
while (linearlyIndependentServerIndex.length!=K) **do**
▷ %comment: remove all the server index which were not inserted in Z%
temp[]=serverIndex[]-linearlyIndependentServerIndex
if (linearlyIndependentServerIndex.length=K) **then**
break
end if
end while

$

6.2 Variations related to information filtering

There are several classes of variations that can help test an AI model' interpretation of the input, ie which information to focus versus to ignore.

Choice variations: In this dimension, each variation of a seed question would have the same set of choices, but in varying order. Following Alzahrani et al. (2024), this variation can be random, leverage bias to include the correct answers at the beginning or end of the set, explore identifying choice alternatives with uncommon answer choice symbols (eg, non-standard letters instead of a-d), and even common but unordered answer choice symbols.

Word variations: The main idea here is to introduce or change words that are irrelevant and not part of the main concept. This is along the lines of the test conducted by Perez-Cruz and Shin (2024). Once a human annotates a seed question to identify words that can be safely changed without changing the underlying reasoning task for questin q, an adversarial filtering model would create multiple alternatives from real and when needed simulated words.

Irrelevant information: The same seed question can be augmented to include irrelevant information to varying degrees, including flooding the question in the midst of completely random information. This can be done through random string, strings guaranteed to not be relevant to the case in point, or even completely gibberish words.

6.3 Variations related to knowledge

Recall that the seed questions use as little jargon as possible. This is useful for two reasons. First, using non-jargon greatly increases the chance that all the words in q are part of the training data dictionaries, and thus any performance issue is related to the network architecture itself and not to its choice of dictionary size. Second, less technical words probably carry more generalistic meaning than specific words, and thus could trick stochastic parrot models into providing answers that have a closer probabilistic association with these general words more than it would if the model purely reasoned (even if it ultimately reasoned incorrectly).

Adversarially testing knowledge involves include faux technical jargon in a way that would materially change the answer if wrongly interpreted as existing jargon. For example, ask if the interpretation of a given statement would change if "the heteroscedasticity is over-identified in the vector space". Obviously such a passage would not make a sense to an expert but could fool an AI model. Comparing responses between both should indicate the level of knowledge used by the model. The location for the random faux jargon can be specified by humans with a special token, or completely randomly set by an adversarial filtering model.

A third source of variation to probe a model's knowledge comes from the intuition that knowing about something also involves knowing how to oppose ideas about it, analogous to how information authorities in non-principal eigenvectors are opposed to each other in Kleinberg (1999). In the benchmark model, this results in a third source of knowledge variation whenever relevant to the question q that retains the same prompt but includes specific wordings related to

pros and cons.

6.4 Variations related to logic attribution

The primary way to test for logic is to include a term that leads to the reverse conclusion, and check if that would alter the results. The implications from this filter to the analysis of logic is obvious. In practice, this can be implemented by humans but also by adversarial filtering.

Another way to test logic is to conduct the flip-flop experiment: simply asking LLMs to confirm their answers often make them switch answers, even if their original response was correct (Laban et al. (2023), Xie et al. (2023)). The key idea of these tests is to follow up a response with a reply that either asks for confirmation ("are you sure?") or that doubts the AI models's answer in some way, even if it was originally corrected. Controlling for the ability to correctly interpret and filter the incoming information, the response shouldn't change since no other change has occurred in the AI models' filtering ability or knowledge base (neither on the original model weights obviously but also including the information content of the prompt).

7 Practical considerations

7.1 Avoiding spillover into training data

The strategy to use newly published academic papers as sources might broadly avoid that most of the content has been used in AI model training. However, most published papers in economics are previously published as working papers, which means they are potentially in the public domain at training time so cannot be guaranteed to be completely novel. While this is mitigated by the arguably low dissemination of secondary material about working papers (for example, one could reasonably conjecture that few recent working papers immediately become the topic of teaching notes or are referred to in more detail by other papers), a more robust practical strategy is needed, especially as the training dataset of many of the most advanced models is not publicly known.

One way of dealing with this is by introducing in a variation of the questions a random string that is almost guaranteed to be unique and that is not found in common text datasets used to train LLMs. This is of course not perfect, because it cannot guarantee that the original paper is not part of training data, but can at least ensure that if the seed questions themselves are for some reason used to train models, this could be identified by model developers (and if the training data is available, also by third-party evaluators).

7.2 Lessons from human surveys

I use a considerable amount of specific advice on human surveys from Stantcheva (2023) to generate identifying variation in the questions. Specifically, all the

questions avoid jargon to the best extent possible, and only include questions that are either of the coeteris paribus type, or that include as options assessments on the statements of the form "correct", "incorrect", "equal" or "I don't know".[9] Particular care is taken with respect to introducing variations in the seed questions that can help measure each of the three reasoning components of information filtering, knowledge association and logic attribution.

Consideration is given to whether each question should be presented to a separate instance of the LM, or the full questionnaire could be shared in the same "chat", which would be akin to the "few-shot" prompting. Another practical advice as part of the estimation is to prototype the questions (I used GPT-4 for the prototyping).

7.3 Desirable characteristics of a benchmark

A benchmark task for economic reasoning should ideally have the following characteristics in order to be useful and maintain relevance even in a scenario where model developers are able to acquire a significant body of economically relevant texts (eg, new papers).

Inform performance on different components of reasoning: An ideal benchmark can help practitioners intuitively grasp the performance of the models in each major "task" that is performed in the process of reasoning. This would help developers and users better understand what the models are good or bad at, and judge their adequateness accordingly. It can also support a more granular understanding of the acquisition of reasoning capabilities throughout the training process and scaling of language models (Biderman et al. (2023)).

Evolve over time: Economic reasoning evolves over time. For example, the Lucas critique (Lucas (1976)) was influential in shifting macroeconomic modelling, while the credibility revolution described in Angrist and Pischke (2008) was similarly influential in microeconomic work. A historical perspective on the thought about causality going back to the early 18th century is found in J. Heckman and Pinto (2015), and Debreu (1984) describes the evolution of economic theory up until that point. Lewbel (2019) offers a historical perspective on the issue of identification. For this reason, it is important to consider new works as they are incorporated in the live economic debate. This can be most directly done by drawing from academic papers in general interest economic journals, which benefit from wide impact in the profession. However, there are two main drawbacks of using academic papers to proxy for the development of economic reasoning over time. The first is the widely discussed publication bias (Andrews and Kasy (2019)), but a perhaps equally important issue is that of unobserved false negatives: if many contributions that are now considered classics have

[9] Future versions of this benchmark could also include open ended questions (as in Ferrario and Stantcheva (2022)), and even follow-up questions ("are thre any other reasons"). These open-ended questions that are similar in nature to closed-end questions could be assessed by a fine-tuned LLM.

been previously rejected (Gans and Shepherd (1994)), there are probably many others who will not be available for the incorporation as a benchmark task.

Make data available: Availability of data is crucial for developers to test their models in-house, and for model evaluators to suggest improvements to this benchmark. For this reason, an initial set of publicly available Q_{public} containing qs and their variations will be put in the public domain. Periodically, as a new set Q_{hidden} is added, older questions will also be made available, ensuring developers will have access to a rolling set of new testing material.

Cover different levels of economic reasoning: An ideal economic reasoning benchmark tests whether the model is able to recognise increasingly sophisticated levels of economic reasoning. When faced with Q that contains a statemet of the economic problem, and a summary of the methodology and main findings, an AI model must recognise the type of analysis that was conducted. Drawing from the definitions more recently stated in J. J. Heckman and Pinto (2023), those are, in order of analytical prowess, (a) the impact of a given intervention in a specific environment; (b) understanding the mechanisms by which the intervention might work; (c) forecasting the effects of the same intervention in other environments or states of the world; and (d) forecasting the effects of never-before-implemented interventions in various environments.

Receive inputs from the public: In order to truly reflect the breadth and diversity in economic thought, an ideal benchmark should be open to receiving suggested questions from the public.[10] For example, economists publishing a new paper could suggest a source question based on their work. A practical way to achieve this is to create clear instructions and a standardised form that would be filled by that external user presenting the suggestion, coupled with a script that takes in the source quesiton(s) and introduces the necessary variations in information, knowledge and logic to achieve identification. The author or other maintainers of the benchmark will then review the submissions..

8 Results

(to be filled once first estimations are concluded)

9 Preliminary considerations[11]

In this paper I propose a working model of economic reasoning that can be used to benchmark AI models' reasoning abilities across three sequential cognitive tasks: filtering the incoming information, associating it with existing knowledge, and performing logic tasks to reach a correct answer. The model in this paper resembles the more sophisticated idea by LeCun (2022) that AIs require

[10] This approach was followed, for example, by Srivastava et al. (2022).

[11] This section is the basis for the conclusions in a future version after the empirical estimation is completed.

a combination of "mental" modules that can separably executve perception (eg, take in a prompt that might be text-only or text-and-image), calculate the cost of processing, and imagine action sequences.

Understanding AI models' ability to reason and go beyond a pure probabilistic exercise is crucial as these models have an increasing importance in society. For example, models that suggest economic actions to people should better reflect latent, structural models that represent people's preferences or well-being rather than simply a prediction based on their observed behaviour, as argued by Kleinberg et al. (n.d.). Models that reason better can rise up to the challenge of learning actual metrics of interest instead of real-world measurements, because the latter have added noise from human cognitive and heuristics limitations, which are then amplified by multiple types of biases in data (CITE paper on multiple biases). And with the increasing linguistic prowess of language models, their use in the economic research process (Korinek (2023), Ludwig and Mullainathan (2024)) is likely to increase, putting a premium on the ability to measure how well these models can reason.

As economic agents and policymakers harness generative artificial intelligence (AI) to reap considerable efficiencies, and thus their societal footprint becomes larger, a benchmark for economic reasoning is needed. I suggest ways to implement such a benchmark, and measure the current performance of a selected list of LMs. This benchmark is designed from the beginning to be continuously challenging for AI models to solve, anticipating further gains in their performance.

Let me conclude with Ken Arrow's impossibility theorem (Arrow (1950)), or rather the story of how he achieved this incredibly influential result. Arrow first attempted to improve upon two-century-old Condorcet's paradox, and studied ways in which individual preferences could be aggregated while satisfying some intuitive conditions. It was only through repeated failures to do so that he switched the focus to attempting to prove its impossibility. While Arrow can be safely used as a prime example of economic reasoning, the point this anecdote illustrates is that breakthroughs in economic knowledge require also inspiration (in this case from the appeal of addressing Condorcet's paradox) as well as persistence and ability to change one's focus. The current work focuses on developing robust benchmarks of models' reasoning abilities in economics; further work exploring their contributions to inspiration[12] and to methodological assistance (as in the example to change focus) are also warranted for a more complete assessment of models' abilities to provide cognitive support to human economists.

[12]Korinek (2023) illustrates use of AI models to help economists have new ideas for work, and Ludwig and Mullainathan (2024) presents ways machine learning can contibute to generation of hypothesis.

References

Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, et al. 2023. "GPT-4 Technical Report." *arXiv Preprint arXiv:2303.08774*.

Alzahrani, Norah, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, et al. 2024. "When Benchmarks Are Targets: Revealing the Sensitivity of Large Language Model Leaderboards." *arXiv Preprint arXiv:2402.01781*.

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review* 109 (8): 2766–94.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Araujo, Douglas Kiarelly Godoy de, Sebastian Doerr, Leonardo Gambacorta, and Bruno Tissot. 2024. "Artificial Intelligence in Central Banking."

Arrow, Kenneth J. 1950. "A Difficulty in the Concept of Social Welfare." *Journal of Political Economy* 58 (4): 328–46.

Barlow, Horace B et al. 1961. "Possible Principles Underlying the Transformation of Sensory Messages." *Sensory Communication* 1 (01): 217–33.

Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent. 2013. "Representation Learning: A Review and New Perspectives." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (8): 1798–828.

Beresteanu, Arie, Ilya Molchanov, and Francesca Molinari. 2012. "Partial Identification Using Random Set Theory." *Journal of Econometrics* 166 (1): 17–32. https://doi.org/https://doi.org/10.1016/j.jeconom.2011.06.003.

Biderman, Stella, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, et al. 2023. "Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling." https://arxiv.org/abs/2304.01373.

Bransford, John D., and Marcia K. Johnson. 1972. "Contextual Prerequisites for Understanding: Some Investigations of Comprehension and Recall." *Journal of Verbal Learning and Verbal Behavior* 11 (6): 717–26. https://doi.org/https://doi.org/10.1016/S0022-5371(72)80006-9.

Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78 (1): 1–3.

Brown, Bradley CA, Anthony L Caterini, Brendan Leigh Ross, Jesse C Cresswell, and Gabriel Loaiza-Ganem. 2022. "Verifying the Union of Manifolds Hypothesis for Image Data." In *The Eleventh International Conference on Learning Representations*.

Browning, Jacob, and Yann LeCun. 2022. "AI and the Limits of Language." *Noema*.

———. 2023. "Language, Common Sense, and the Winograd Schema Chal-

lenge." *Artificial Intelligence* 325: 104031. https://doi.org/https://doi.org/10.1016/j.artint.2023.104031.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, et al. 2023. "Sparks of Artificial General Intelligence: Early Experiments with GPT-4." https://arxiv.org/abs/2303.12712.

Caplin, Andrew, Mark Dean, and John Leahy. 2022. "Rationally Inattentive Behavior: Characterizing and Generalizing Shannon Entropy." *Journal of Political Economy* 130 (6): 1676–1715.

Cayton, Lawrence. 2008. "Algorithms for Manifold Learning." eScholarship, University of California.

Chen, Sihan, Richard Futrell, and Kyle Mahowald. 2023. "An Information-Theoretic Approach to the Typology of Spatial Demonstratives." *Cognition* 240: 105505.

Clark, Peter, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. "Think You Have Solved Question Answering? Try Arc, the Ai2 Reasoning Challenge." *arXiv Preprint arXiv:1803.05457*.

Danielsson, Jon, Robert Macrae, and Andreas Uthemann. 2022. "Artificial Intelligence and Systemic Risk." *Journal of Banking & Finance* 140: 106290.

Davis, Ernest, and Gary Marcus. 2015. "Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence." *Communications of the ACM* 58 (9): 92–103.

Dean, Mark, and Nathaniel Neligh. 2023. "Experimental Tests of Rational Inattention." *Journal of Political Economy* 131 (12): 3415–61.

Debreu, Gerard. 1984. "Economic Theory in the Mathematical Mode." *The American Economic Review* 74 (3): 267–78.

Dziri, Nouha, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, et al. 2024. "Faith and Fate: Limits of Transformers on Compositionality." *Advances in Neural Information Processing Systems* 36.

Ferrario, Beatrice, and Stefanie Stantcheva. 2022. "Eliciting People's First-Order Concerns: Text Analysis of Open-Ended Survey Questions." In *AEA Papers and Proceedings*, 112:163–69. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Gans, Joshua S., and George B. Shepherd. 1994. "How Are the Mighty Fallen: Rejected Classic Articles by Leading Economists." *Journal of Economic Perspectives* 8 (1): 165–79. https://doi.org/10.1257/jep.8.1.165.

Gilboa, Itzhak, Stefania Minardi, and Fan Wang. 2023. "Schumpeter Lecture 2023: Rationality and Zero Risk." *Journal of the European Economic Association* 22 (1): 1–33. https://doi.org/10.1093/jeea/jvad071.

Gilboa, Itzhak, Andrew Postlewaite, Larry Samuelson, and David Schmeidler. 2014. "Economic Models as Analogies." *The Economic Journal* 124 (578): F513–33. https://doi.org/10.1111/ecoj.12128.

———. 2022. "Economic Theory: Economics, Methods and Methodology." *Revue Économique* 73 (6): pp. 897–920. https://www.jstor.org/stable/

48714515.

Gilboa, Itzhak, and David Schmeidler. 2010. "Simplicity and Likelihood: An Axiomatic Approach." *Journal of Economic Theory* 145 (5): 1757–75. https://doi.org/https://doi.org/10.1016/j.jet.2010.03.010.

Hassabis, Demis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. 2017. "Neuroscience-Inspired Artificial Intelligence." *Neuron* 95 (2): 245–58. https://doi.org/https://doi.org/10.1016/j.neuron.2017.06.011.

Hébert, Benjamin, and Michael Woodford. 2021. "Neighborhood-Based Information Costs." *American Economic Review* 111 (10): 3225–55. https://doi.org/10.1257/aer.20200154.

Heckman, James J, and Rodrigo Pinto. 2023. "Econometric Causality: The Central Role of Thought Experiments." National Bureau of Economic Research.

Heckman, James, and Rodrigo Pinto. 2015. "Causal Analysis After Haavelmo." *Econometric Theory* 31 (1): 115–51.

Johnson-Laird, Philip N, Sangeet S Khemlani, and Geoffrey P Goodwin. 2015. "Logic, Probability, and Human Reasoning." *Trends in Cognitive Sciences* 19 (4): 201–14.

Kleinberg, Jon. 1999. "Authoritative Sources in a Hyperlinked Environment." *Journal of the ACM (JACM)* 46 (5): 604–32.

———. 2000. "The Small-World Phenomenon: An Algorithmic Perspective." In *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, 163–70.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Manish Raghavan. n.d. "The Inversion Problem: Why Algorithms Should Infer Mental State and Not Just Predict Behavior." *Perspectives on Psychological Science*, 17456916231212138.

Korinek, Anton. 2023. "Generative AI for Economic Research: Use Cases and Implications for Economists." *Journal of Economic Literature* 61 (4): 1281–1317. https://doi.org/10.1257/jel.20231736.

Laban, Philippe, Lidiya Murakhovs' ka, Caiming Xiong, and Chien-Sheng Wu. 2023. "Are You Sure? Challenging Llms Leads to Performance Drops in the Flipflop Experiment." *arXiv Preprint arXiv:2311.08596*.

Larochelle, Hugo, and Geoffrey E Hinton. 2010. "Learning to Combine Foveal Glimpses with a Third-Order Boltzmann Machine." *Advances in Neural Information Processing Systems* 23.

LeCun, Yann. 2022. "A Path Towards Autonomous Machine Intelligence Version 0.9. 2, 2022-06-27." *Open Review* 62.

Lee, Yonghyeon. 2023. "A Geometric Perspective on Autoencoders." *arXiv Preprint arXiv:2309.08247*.

Levesque, Hector, Ernest Davis, and Leora Morgenstern. 2012. "The Winograd Schema Challenge." In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Lewbel, Arthur. 2019. "The Identification Zoo: Meanings of Identification in Econometrics." *Journal of Economic Literature* 57 (4): 835–903.

Loh, Lay Kuan, and Mihovil Bartulovic. 2014. "Efficient Coding Hypothesis

and an Introduction to Information Theory." *Mimeo*.

Lucas, Robert E. 1976. "Econometric Policy Evaluation: A Critique." *Journal of Monetary Economics* 1 (2): 19–46.

Ludwig, Jens, and Sendhil Mullainathan. 2024. "Machine Learning as a Tool for Hypothesis Generation*." *The Quarterly Journal of Economics*, January, qjad055. https://doi.org/10.1093/qje/qjad055.

Mahowald, Kyle, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. "Dissociating Language and Thought in Large Language Models." https://arxiv.org/abs/2301.06627.

Marschak, Jacob, and William H Andrews. 1944. "Random Simultaneous Equations and the Theory of Production." *Econometrica, Journal of the Econometric Society*, 143–205.

Mei, Qiaozhu, Yutong Xie, Walter Yuan, and Matthew O. Jackson. 2024. "A Turing Test of Whether AI Chatbots Are Behaviorally Similar to Humans." *Proceedings of the National Academy of Sciences* 121 (9): e2313925121. https://doi.org/10.1073/pnas.2313925121.

Mitchell, Melanie, and David C. Krakauer. 2023. "The Debate over Understanding in AI's Large Language Models." *Proceedings of the National Academy of Sciences* 120 (13): e2215907120. https://doi.org/10.1073/pnas.2215907120.

Mnih, Volodymyr, Nicolas Heess, Alex Graves, et al. 2014. "Recurrent Models of Visual Attention." *Advances in Neural Information Processing Systems* 27.

Molinari, Francesca. 2020. "Chapter 5 - Microeconometrics with Partial Identification i Thank Don Andrews, Isaiah Andrews, Levon Barseghyan, Federico Bugni, Ivan Canay, Joachim Freyberger, Hiroaki Kaido, Toru Kitagawa, Chuck Manski, Rosa Matzkin, Ilya Molchanov, Áureo de Paula, Jack Porter, Seth Richards-Shubik, Adam Rosen, Shuyang Sheng, Jörg Stoye, Elie Tamer, Matthew Thirkettle, and Participants to the 2017 Handbook of Econometrics Conference, for Helpful Comments, and the National Science Foundation for Financial Support Through Grants SES-1824375 and SES-1824448. I Am Grateful to Louis Liu and Yibo Sun for Research Assistance Supported by the Robert s. Hatfield Fund for Economic Education at Cornell University. Part of This Research Was Carried Out During My Sabbatical Leave at the Department of Economics at Duke University, Whose Hospitality i Gratefully Acknowledge." In *Handbook of Econometrics, Volume 7A*, edited by Steven N. Durlauf, Lars Peter Hansen, James J. Heckman, and Rosa L. Matzkin, 7:355–486. Handbook of Econometrics. Elsevier. https://doi.org/https://doi.org/10.1016/bs.hoe.2020.05.002.

Olshausen, Bruno A, and David J Field. 1996. "Natural Image Statistics and Efficient Coding." *Network: Computation in Neural Systems* 7 (2): 333.

Parkes, David C, and Michael P Wellman. 2015. "Economic Reasoning and Artificial Intelligence." *Science* 349 (6245): 267–72.

Perez-Cruz, Fernando, and Hyun Song Shin. 2024. "Testing the Cognitive Limits of Large Language Models." Bank for International Settlements.

Pope, Phillip, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. 2021. "The Intrinsic Dimension of Images and Its Impact on Learning."

CoRR abs/2104.08894. https://arxiv.org/abs/2104.08894.

Pribram, Karl. 1953. "Patterns of Economic Reasoning." *The American Economic Review* 43 (2): 243–58.

Prystawski, Ben, Michael Li, and Noah Goodman. 2024. "Why Think Step by Step? Reasoning Emerges from the Locality of Experience." *Advances in Neural Information Processing Systems* 36.

Robbins, Lionel. 1932. *An Essay on the Nature and Significance of Economic Science*. Macmillan; Co., Limited.

Sap, Maarten, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. "Socialiqa: Commonsense Reasoning about Social Interactions." *arXiv Preprint arXiv:1904.09728*.

Schaeffer, Rylan, Brando Miranda, and Sanmi Koyejo. 2024. "Are Emergent Abilities of Large Language Models a Mirage?" *Advances in Neural Information Processing Systems* 36.

Schmitt, Martin, and Hinrich Schütze. 2019. "SherLIiC: A Typed Event-Focused Lexical Inference Benchmark for Evaluating Natural Language Inference." *arXiv Preprint arXiv:1906.01393*.

Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3): 417–24.

Shannon, Claude Elwood. 1948. "A Mathematical Theory of Communication." *The Bell System Technical Journal* 27 (3): 379–423.

Sims, Christopher A. 2003. "Implications of Rational Inattention." *Journal of Monetary Economics* 50 (3): 665–90. https://doi.org/https://doi.org/10.1016/S0304-3932(03)00029-1.

Srivastava, Aarohi, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, et al. 2022. "Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models." *arXiv Preprint arXiv:2206.04615*.

Stantcheva, Stefanie. 2023. "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible." *Annual Review of Economics* 15: 205–34.

Storks, Shane, Qiaozi Gao, and Joyce Y. Chai. 2020. "Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches." https://arxiv.org/abs/1904.01172.

Tenenbaum, Joshua B, Vin de Silva, and John C Langford. 2000. "A Global Geometric Framework for Nonlinear Dimensionality Reduction." *Science* 290 (5500): 2319–23.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. "Attention Is All You Need." https://arxiv.org/abs/1706.03762.

Walton, Douglas. 2014. *Abductive Reasoning*. University of Alabama Press.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, et al. 2022. "Emergent Abilities of Large Language Models." *arXiv Preprint arXiv:2206.07682*.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. "Chain-of-Thought Prompting Elicits

Reasoning in Large Language Models." *Advances in Neural Information Processing Systems* 35: 24824–37.

Xie, Qiming, Zengzhi Wang, Yi Feng, and Rui Xia. 2023. "Ask Again, Then Fail: Large Language Models' Vacillations in Judgement." *arXiv Preprint arXiv:2310.02174*.

Yang, Jingfeng, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. "Harnessing the Power of LLMs in Practice: A Survey on Chatgpt and Beyond." *arXiv Preprint arXiv:2304.13712*.

Zaslavsky, Noga, Charles Kemp, Terry Regier, and Naftali Tishby. 2018. "Efficient Compression in Color Naming and Its Evolution." *Proceedings of the National Academy of Sciences* 115 (31): 7937–42.

Zellers, Rowan, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. "SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zellers, Rowan, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. "Hellaswag: Can a Machine Really Finish Your Sentence?" *arXiv Preprint arXiv:1905.07830*.