Testing Machine Intelligence: an Internal Interpretation of Commonsense Reasoning

Riccardo Campanella, 8175721

Introduction

This paper investigates whether we should look for Commonsense Reasoning inside Large Language Models as their potential capacity for achieving general intelligence. Common sense reasoning remains a significant challenge in building AI systems, thereby we explore to which extent it is a key feature of Intelligence and how we can measure it. In this work, we also raise concerns over the prior methodologies of testing machine intelligence that failed to account for the internal workings of a model and focused on behavioural performances. Specifically, drawing from the limitations of the Winograd Schema Challenge and the Turing Test, we propose to use internal testing involving Mechanistic Reasoning to assess the real model capabilities of intelligence based on Commonsense Reasoning. Further research on this paper's findings is considered to bring added value to the ongoing theory of commonsense knowledge formalisation, identification, and improvement of systems with emergent capabilities closer to the AGI.

1. Unconstrained testing leads to failure

The defeat of the Winograd Schema Challenge (Kocijan et al. 2023) demonstrated how machines can easily pass tests of general Intelligence due to the perceived connections between tasks and methods. Indeed, Hector Levesque's pronoun disambiguation was selected as a task to test intelligence because it seems to require the use of commonsense knowledge in humans to solve it, however, that was not the case for the machines. Pre-trained transformer-based and fine-tuned on these kinds of problems learned to solve the challenge by using statistical correlations and achieved high accuracy demonstrating the weakness of using surrogate tests. Hence the careful evaluation of proposing proxy problems as tests for general Intelligence. Another concern arising from the same paper regards the creation of test sets that span, as much as possible, instances of the targeted problem as well as nested modalities including physical, spatial, interpersonal, and social reasoning. Solving the challenge demonstrated the proficiency of such models on problems of different domains but did not suffice to prove their capabilities in general commonsense reasoning. Furthermore, the WSC was subject to the critiques of the Turing Test (Frankish et al. 2014), for what concerns the lack of evaluation of the functional organisation of a system to assess machine intelligence. The challenge indeed was a mere behavioural test that appealed because of its ease of evaluation in determining whether a model passed the test or not. No explanation was required to be provided from the model in solving the problem, thereby exposing the test to the criticisms of the Chinese Room Argument (Frankish et al. 2014) where models still have issues of meaning, understanding, and interaction with the external world.

For that reason, it is necessary to prove that machines use common sense when providing answers and making decisions, based on interpreting their internal knowledge.

2. General Intelligence comes along with a key feature: commonsense reasoning

The upcoming passages offer a critical examination, aiming to define essential concepts for a thorough internal analysis. Beginning with an exploration of commonsense reasoning, we delve into its intersection with AGI, demonstrating its central role in building general intelligence.

2.1 Commonsense reasoning

To address the challenge of blind testing inherent in previous methodologies, a proposed solution involves internal interventions to measure the generalisation of models across commonsense reasoning capacities. However, before delving into this approach, it's essential to establish what is commonsense reasoning and its significance within the context of AGI. Commonsense reasoning involves making inferences and decisions based on everyday knowledge that most people possess. According to this definition, AI systems need to understand the context of a situation to make sense of it, and thereby master concepts such as time, space, causality, and relationships between objects or events. Additionally, those systems must possess the ability to make deductions based on available information to make reasonable guesses when the information is missing (Zandie et al. 2023) or ambiguous. The abductions are also part of common sense as they can be used to infer the most likely explanation or cause of a given set of facts or observations. Finally, another desiderata is the existence of a knowledge base of the world facts about the physical world, social norms, cultural practices, and other aspects of human experience.

2.2 Commonsense Reasoning and AGI

Artificial General Intelligence refers to a hypothetical AI system that possesses the ability to understand, learn, and apply intelligence across a wide range of tasks and domains, similar to human-level intelligence. AGI systems are characterised by their capacity for flexible problem-solving, adaptability to new situations, and their ability to perform tasks autonomously without specialised programming for each task. These systems are designed to exhibit cognitive versatility and to emulate the broad spectrum of cognitive abilities observed in humans. The ultimate goal of reaching the AGI is to develop machines capable of generalising knowledge and skills across diverse contexts, rather than being limited to specific, narrow domains. The aspiration of achieving the AGI originated from the human desire to mimic human intelligence, thereby it brings with it some expectations regarding the exhibition of similar skills of problem-solving and adaptability. Those expectations arise from the human’s capacity to reason about the world and perform intuitive understanding as well as anticipation that result in adaptation to the real world environment which involves complex and ambiguous situations. As opposed to other kinds of intelligence Common Sense Reasoning can be considered a discriminative characteristic of the human species but does not represent a

discriminative feature of general intelligence, as it is necessary to include other factors such as planning to obtain machine general intelligence. It simply does not imply human intelligence though it demonstrates a potential capability, allowing researchers to carefully look for its presence in the system's internal workings before assessing it in a final behavioural test. Systems' interpretations are necessary as Commonsense Reasoning is a form of reasoning that originated from external or internal inquiries, or in other words resulting from retrospective and introspective acts which makes it a crucial part of AGI

3. Internal testing: Mechanist Reasoning accounts for Explanation and Adaptation

This chapter delves into the testing tools, targets and applications to ultimately provide some general recommendations from this thorough analysis. Beginning with the interpretability methods, the core features of Commonsense Reasoning are uncovered through two Cognitive Theories of Mind and then identified with Mechanistic Reasoning interpretation.

3.1 Testing reasoning: Interpretability methods and Mechanistic Interpretability

Internal interpretation of the model's commonsense reasoning capabilities is achieved with Mechanistic reasoning rather than just using Prompting, Probing or Editing. Previous methods of testing common sense in machines leveraged Multiple-choice questions and multiple-question answers as well as adversarial benchmarks. They all involve crafting a dataset and testing a pre-trained model on that. As we break common sense into its core components, specific datasets are crafted to test different domains. None of those datasets alone suffice to test common sense in general, but when combined they can provide a reliable measure of how the model performs on different tasks. However, their weakness resides in the possibility that models can pass combinations of behavioural tests, without actually having general intelligence (Shen et al. 2023). Therefore a better measure of intelligence can be captured by internal testing through interpretability methods. One possible objection is that there exist some benchmarks that test common sense differently than the mentioned approaches. The straight answer is that for frameworks like the Context-rich Evaluation (Kejriwal at al. 2023), there is a human in the loop which underscores the test reliability on humans as well as the contingency on external input and inquiry to achieve common sense in machines. Having a human in the loop to evaluate the performance still underlies the weakness of the Turing Test.

Given these circumstances, It's natural to wonder what should be tested inside the models and which are the appropriate tools for doing so. Commonsense Reasoning is highly contingent on knowledge retrieval and knowledge reconstruction, as much as humans rely on retrieving information located in different parts of the hippocampus to build their memories. The Interpretability methods include Prompting, Probing and Mechanistic Interpretability can highlight which features or aspects of the input data the model is relying on to make predictions. In particular, Prompting is a way to extract knowledge directly from the model when prompted or asked to do so. On the other hand, probing addresses what knowledge can be extracted from

the representations. Finally, Mechanistic Interpretability explores how the expressions of knowledge can change with interventions on the model's components. All these interpretability methods though fail to address the reconstruction process involved in the model answer generation. It's here that Mechanistic Reasoning comes into play. It belongs to a new class of interpretability methods which harness the strengths of the Mechanistic Interpretability. Mechanistic Reasoning involves understanding the internal workings of a model, such as its architecture, parameters, and learning algorithms, to explain how it arrives at its predictions or decisions. Moreover, it can provide valuable insights into how models generate the answers by identifying the architecture components responsible for the answer's construction process.

3.2 Cognitive Theories of Reasoning: Explanation and Adaptation

Given our aspiration to mimic human reasoning in AGI, it's only natural to dive into the workings of human cognition. The theory of mind best suited to address our needs for interpretability testing is expressed by Hugo Mercier (Mercier et al. 2017). According to Mercier human reasoning can be defined as the act of justifying intuitive inferences. That's because humans make decisions almost instinctively while reasoning is rather a process justification of what we think to others. Conversely, we can say that reasoning also comes into play when we want to evaluate arguments made by others. Common sense reasoning is then built upon this knowledge reconstruction process triggered by justification and explanation purposes. That's what caused Richard Feynamn's Learning Theory to be successful: explaining to others allows one to gain a deeper understanding of complex concepts by identifying the gaps in the process of justifications. One potential objection is the scenario where reasoning does not necessarily involve explanation and justification: System1-System2 Thinking (Khaneman 2017). Mercier's theory contrasts to some extent with Kahneman's theory as it is consistent with what is sustained by System 1 thinking but restricts System 2 thinking to necessarily involve justification. One objection to that might involve the unnecessary use of justification as the primary driver as well as the potential introduction of biases in our justification due to the presence of influential factors such as confirmation bias. According to that, people may engage in reasoning processes to understand complex concepts, solve problems, or evaluate arguments, even in the absence of a need for external justification. A possible answer supporting Mercier's theory would be that reasoning is simultaneously a retrospective act of justification and a prospective act of explanation to others. This resembles to some extent the human's inner voice replicating the sound of each word in the act of reading a book, as a result of an inner reasoning process. Additionally, it can be demonstrated that the confirmation bias demonstrates that the reasoning process is triggered only when other individuals prove us wrong as opposed to having a static world model like in System1 thinking.

Reasoning involves retrieving the part of the knowledge required to solve the task at hand as well as adding to the world model the missing knowledge by reshaping it. This is formally theorised by Piaget's Theory of Cognitive Modeling (Rowland et al. 2022) which identifies two main steps involved in reasoning: Assimilation and Accommodation. In the first step, the observations are reshaped to fit the world model, and in the second step, the world model is altered to accommodate experiences that cannot be interpreted with the previous world model. This is a joint condition called Adaptation that we require models to have to showcase commonsense reasoning. One objection to that might involve a scenario where only assimilation or accommodation is present. Having just accommodation would be the equivalent of not being able to

memorise any type of information therefore and that is distant from how it is believed to work the human mind. As a consequence, the construction of AI systems is based on the training process to let them infer and make decisions. The assimilation step is humanly enabled in Chain of Thought prompting where we ask the model to answer questions to solve the task or in Knowledge graphs where the knowledge is assimilated according to a specified graph structure. Therefore the answer provided by the model will always be static and rely heavily on training data. Given these circumstances, It's natural to wonder whether the current models are assimilative and how they achieve accommodation. Current systems are assimilative in the sense that they need to be trained to answer questions regarding unseen data and the expedient to mimic common sense is to guide the answer by splitting it into a sequence of steps. In some cases, supervised question-answer pairs are provided to show the models how to reason. Therefore it comes down to providing a template or framework of reasoning. That's one of the reasons behind models like Chatgpt provide an error message when asking for for information on which the model is not trained. The access to information coming from web crawling in models like Chat-GPT4 is an attempt to "assimilate" new information. However, that is a limitation in being eligible to possess General Intelligence because that contradicts our definition of AGI in not being programmed when reasoning. The ability to accommodate new knowledge is shown in the next paragraph as part of two study cases.

3.2 Mechanistic Reasoning: Adaptation of knowledge through dynamic encoding

Following the shortcomings of the assimilative models, our focus shifts towards exploring an unsupervised approach for retraining the world model. This could not be done with rigid symbolic structure alone, rather it is achieved by exploiting the ability of Subsymbolic systems to self-adjusting the weights. On the other hand, the processes involved in Mechanistic Interpretability allow for the assimilation of new knowledge and accommodate new knowledge into the world model by updating the network's parameters as shown through the study of the Knowledge Critical Subnetworks (Bayatiz et al. 2023) and RECKONING (Chen et al. 2024).

The parameters affecting the process of knowledge reconstruction are identified in the knowledge-critical subnetworks employing Mechanistic Interpretability: when the subnetwork, representing the target knowledge, is removed the prediction of the knowledge required to answer a certain prompt is no longer the same. Additionally, the pre-existing knowledge that is not interested in the weights update must remain unvaried and that is achieved with a knowledge graph that controls. This means that each entity or word has the target knowledge expressed as a neighbourhood of knowledge families that are removed. The core families around the targeted entity are removed and the families surrounding them are kept unaried in the Control graph. Finally, to make sure the model keeps the same behaviours before the editing of knowledge, some random sequences of knowledge that are not involved in the control and target knowledge are kept the same. There is a marginal boundary between what is kept in the Control graph and what is removed as part of the Target knowledge. This is determined by looking at the model's precision in answering the questions but it remains unclear how we can define the knowledge that is forgotten. However, many theories have addressed this problem regarding Cognitive Control advocating that when certain components of a neural network are updated, it may disrupt the cognitive processes involved in maintaining and retrieving relevant knowledge, leading to forgetting. That is additionally supported by semantic networks theory according to

which the knowledge is organised into interconnected nodes and associations, forming a network of semantic relationships. Thanks to the Control knowledge this problem is avoided.

The RECKONING system leverages Meta Learning which consists of adapting quickly the model weights to learn from a a few new examples of new tasks on which the network was not pre-trained. This is usually called few-shot learning and it is done in a way where we use what we've learned in a different task to solve the task at hand, something we're the Deep Learning Neural Networks struggle with. This is done by finding those meta parameters that minimise the test loss over many tasks. The model learns how to update their parameters in a way that facilitates rapid learning of new tasks in a process referred to as learning to learn. To some extent, this is similar to what humans are doing when making new strategies to learn new things more efficiently. Indeed learning new tasks is a matter of building new skills and knowledge based on the existing one. Furthermore, the meta-learning leveraged by RECKONING is model-agnostic, which means it is not bound to any model architecture and thereby can well represent a more abstract and general principle of learning that is close to what humans do. The way the information is encoded is such that the model to answer the questions must recall the information used when learning from the few examples. This meta-learning method is somehow a way to let the machine encode new knowledge through assimilation and adaptation as well as reason about the information provided and use that knowledge to answer new questions. As an explainability method, it provides an explanation of which facts it has used to answer the questions given a set of facts, some of which may be irrelevant. One objection to that may be that reasoning cannot be expressed through the dynamic encoding of knowledge. The answer to that resides in the successful capacity of RECKONING systems. Once the question is given, the knowledge required to answer that question is no longer available but it is successfully encoded in the model parameters and later used to answer, demonstrating the efficacy of dynamic encoding. Additionally, the more steps of reasoning a question requires the more effectively the model can encode the knowledge to answer it. That indicates the ability of the composition of the systems as well as the efficacy in generalising on longer chains of reasonings. This is made possible thanks to the approach that gets this model closer to human cognitive processes, where there is nothing such as the Attention mechanism and the forward pass.

3.3 Internal testing: a complementary approach

Based on the previous findings, it is possible to provide a set of recommendations for developing a model evaluation framework that relies on internal and agnostic interpretations of models. The following steps overcome the blind testing problem of the Winograd Schema Challenge. It is important to acknowledge that internal testing does not replace a final test of intelligence rather it filters out "false positives" or machines incapable of intelligence by constitution.

- No human must be included in the loop
- The model must be asked to answer the questions involving solving problems of reasoning and explain how it crafted the answer together with the most relevant facts used in doing so. Requiring to furnish explanations avoids weak and unconvincing proof of commonsense reasoning abilities
- Mechanistic Reasoning must be used to identify the assimilation and accommodation of new knowledge

- Model Short exposition of new facts and exemplification of unseen knowledge structure is mandatory to replicate human learning capabilities.
- Comparing the models' ability of meta-learning can provide insights into the strengths and weaknesses of quick adaptation to new tasks or domains.
- Included in the tests context distractors, are facts that are irrelevant to the tasks, as they allow for a better comparison of the reasoning process of models.

Following, some types of tests can be carried out based on the study cases aforementioned. These two types of testing leverage the capacity of manipulating symbols and the ability to relate them in graphs which are considered to be fundamental steps in machine reasoning and are addressed in Neurcompositional computing and Commonsense Transformers.

- Training the systems on a sequence of riddles in a supervised approach prevents a closed-book examination of the machine's common sense capabilities. To test correctly the model must be provided with a sequence of facts, some rules on how to use the facts and final questions.
- Include the family of relationship problem, where the facts provided are relationships between entities, in the questions asked to the model. To test the model's compositional ability, the facts and the rules provided are used from the model to answer the questions.

Conclusion: implications and potential outcomes

In conclusion, this study sheds light on the failure of the prior methodologies in testing commonsense reasoning in machines to seek a new solution based on Cognitive Theories of Mind. The new approaches must account for a different definition of common sense, contingent on the properties of explainability according to Mercier and adaptability according to Piaget. Consequently, Mechanistic Reasoning is proposed as the most appropriate method among the last developed to test those features of common sense. Then, It is applied in two study cases to show the efficacy of such an approach and its potential to be extended to every model due to the existence of a model-agnostic type of learning. Finally, the discussion concluded with recommendations of more general principles that must be taken into account when testing models for commonsense reasoning. Ultimately, it is possible to affirm that this approach aims to prevent human overestimation of machine capabilities, reaffirm the importance of testing, and provide valuable insights for constructing machines with enhanced commonsense reasoning abilities.

One potential direction of research resulting from this discussion regards whether it is possible to integrate new measures of reasoning into the Mechanistic reasoning internal investigation. The dynamicity of knowledge encoding when reasoning certainly involves data flow in the following as opposed to the static access of memory when reasoning. Thereby it would be interesting to dive through more Cognitive Modeling Theories of Mind to establish when to measure the optimisation during learning using new interpretation methods that account for the energy flow used from the models in that process.

References

[Kocijan et al. 2023] Kocijan, V., Davis, E., Lukasiewicz, T., Marcus, G., & Morgenstern, L. (2023). The defeat of the Winograd Schema Challenge. Artificial Intelligence, 325. https://doi.org/10.1016/j.artint.2023.103971

[Frankish et al. 2014] Frankish, K., & Ramsey, W. M. (2014). The Cambridge handbook of artificial intelligence. Cambridge University Press. https://doi.org/10.1017/CBO9781139046855

[Mercier et al. 2017] Mercier, H., & Sperber, D. (2017). The enigma of reason. Harvard University Press.

[Khaneman 2017] Daniel, K. (2017). Thinking, fast and slow.

[Houdé et al. 2022] Houdé, O., & Borst, G. (2022). The Cambridge handbook of cognitive development. Cambridge University Press. https://doi.org/10.1017/9781108399838

[Rowland et al. 2022] Rowland, T. L. (2012). Everything you need to know about Jean Piaget's theory of cognitive development. BrainMass Inc. https://central.bac-lac.gc.ca/.item?id=10096_Piaget_s_Theory_o&op=pdf&app=Library

[Bayatiz et al. 2023] Bayazit, D., Foroutan, N., Chen, Z., Weiss, G., & Bosselut, A. (2023). Discovering knowledge-critical subnetworks in pre-trained language models. arXiv preprint arXiv:2310.03084.

[Chen et al. 2024] Chen, Z., Weiss, G., Mitchell, E., Celikyilmaz, A., & Bosselut, A. (2024). RECKONING: reasoning through dynamic knowledge encoding. Advances in Neural Information Processing Systems, 36.

[Kejriwal at al. 2023] Kejriwal, M., Santos, H., Shen, K., Mulvehill, A. M., & McGuinness, D. L. (2023, May). Context-Rich Evaluation of Machine Common Sense. In International Conference on Artificial General Intelligence (pp. 167-176). Cham: Springer Nature Switzerland.

[Zandie et al. 2023] Zandie, R., Shekhar, D., & Mahoor, M. (2023, July). COGEN: Abductive Commonsense Language Generation. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 295-302).

[Kejriwal et al. 2020] Kejriwal, M., & Shen, K. (2020). Do fine-tuned commonsense language models really generalize?. arXiv preprint arXiv:2011.09159.

[Shen et al. 2023] Shen, K., & Kejriwal, M. (2023). An experimental study measuring the generalization of fine-tuned language representation models across commonsense reasoning benchmarks. Expert Systems, 40(5), e13243.