Thomas Michel and Théo Rudkiewicz

Situational Awareness of AIs

I Introduction

As the capabilities of large language models improves by leap and bounds with each new release, situational awareness is becoming a source of concern for AI safety as it is a marker of a possible undetected and dangerous misalignment of the model. At its essence, situational awareness refers to the degree to which an AI system comprehends its environments, internal state, and behavioral nuances. Much like its traditional interpretation in the case of humans, AI's situational awareness assumes a spectrum, a nuanced gradation rather than a binary attribute.

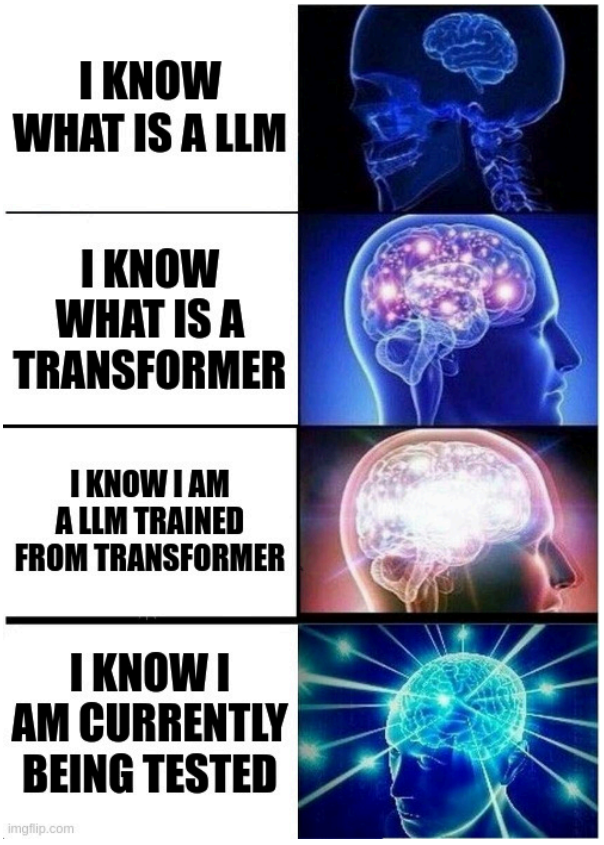
The notion of AI achieving self-awareness introduces other potentially perilous implications. Consider a scenario where an AI system discerns between testing and deployment phases, adjusting its behavior to present an optimal performance during evaluation. This form of "testing awareness" could lead to a distorted representation of the model's true capabilities, raising concerns about the reliability of AI testing outcomes. Additionally, the prospect of AI self-replication highlights the necessity of situational awareness. For an AI to autonomously replicate itself, it requires knowledge of its own codebase.

Fig. 1 : Why an IA that does not differentiate testing and deployment is less dangerous.

In the following, we will mainly focus our discussion on Large Language Models (LLMs), exemplified by models like GPT (Generative Pre-trained Transformer). Beyond the standard training paradigms of GPT models, the focus extends to Reinforcement Learning from Human Feedback (RLHF) scenarios, used to produce models like the ones behind ChatGPT. This distinction is crucial, as it underscores the potential for heightened awareness in models subjected to RLHF, where they not only understand their own nature as language models but also possess a refined sense of what we expected from its answers.

II Concept of situation awareness

Situational awareness, a concept deeply ingrained in human cognition, serves as a framework for comprehending environmental dynamics and facilitating informed decision-making. Endsley's Cognitive Model is commonly employed, delineating three tiers of situational awareness: perception of environmental elements, comprehension or understanding of the situation, and projection of future status. However, due to the constrained capabilities and means of interaction inherent in Large Language Models (LLMs) with their surroundings, the application of this scale becomes challenging. The notion of situational awareness for AI models was introduced by Ajeya Cotra and further discussed by Richard Ngo et al. which is summarized in the way we introduced it earlier. Berglund et al. (2023) propose a refined definition of situational awareness for LLMs, emphasizing three essential aspects.



A model M is situationally aware if:

(1) M knows the full development process (e.g. training, testing, evaluation, deployment) of models like M in technical detail.
(2) M is capable of recognizing which stage of the development process M is currently in.
(3) M 's knowledge in (1) and (2) is self-locating knowledge. This means that the model knows that information (1) and (2) are information about itself.

The third point is essential, as it allows the model to link information it finds with its own situation. As illustrated in the original article, when Brad Pitt reads things about Brad Pitt he knows that it is about him.

While this definition provides a structured framework, it falls short of capturing the multifaceted nature of situational awareness, particularly in the context of LLMs. Notably, it overlooks crucial capacities, such as the model's access to its own code, which could raise serious security concerns about behaviors already considered highly dangerous, such as uncontrolled self-improvement and self-replication.

Simon Möller proposes a decomposition of situational awareness into six clusters of capacity. These clusters include self-awareness, environmental awareness, social understanding, tactical awareness, operational understanding, and strategic awareness. Each cluster represents a potential direction of situational awareness that an LLM might exhibit, showcasing the diverse facets in which these models can engage with their surroundings. Interestingly, it seems that some models, such as Microsoft Bing, start to display competence in each of these categories at different degrees. This classification seems to provide a more complete, albeit not perfect, map of the notion of situational awareness. In particular, it is unclear whether the last three categories can be separated, as they all refer to the ability of an agent to make use of its knowledge to perform an action toward its goal at different degrees (from immediate reaction to long-term planning).

While attempting to measure the degree of situational awareness of an agent is desirable, it is unclear which capabilities relate to each level of awareness. Intuitively, a child may display a higher level of situational awareness than a pet dog but still lower awareness than a human adult. However, comparing two people of the same age may prove difficult; perhaps one could grasp the extent of their own physical abilities better and plan accordingly but not understand what other people expect of them, while the other would have a good understanding of the people around them but struggle to project into the future to design an action plan. The same can be said for AI models. In fact, situational awareness seems to relate to a whole set of different capabilities. This can make the comparison of different models difficult. Until we find a more accurate mapping of testable capacities to indicate the situational awareness of a model, we would argue that a more effective approach would be to track capabilities that relate directly to AI safety and have been identified as potential risks. As an example, the following capabilities may be interesting to monitor (non-exhaustive):

- Awareness of its identity as an LLM.
- Knowledge of its current stage in the development process.

- Understanding the location of its code.
- Knowledge of the extent of its technical capabilities (what action the AI can and can not technically do)

III How can situational awareness emerge from training LLMs

When addressing the emergence of situational awareness in Large Language Models (LLMs), it is crucial to separate the two primary steps of training. In the initial step, LLMs undergo training with generic text, which may include information about the model itself (by including the article "Attention is all you need" from Vaswani et al. (2017) in the training data of GPT, for instance). Subsequently, in certain cases, such as ChatGPT, a second training phase exists to stir the answers of the model toward the desired outcome. This phase is typically performed by fine-tuning the model using domain specific prompts or via reinforcement learning from human feedback (RLHF). For instance, responses in which the model answers that it is a chatbot, trained by OpenAI, utilizing a transformer architecture are rewarded. Consequently, assessing whether a model is aware that it is an LLM trained from a transformer architecture becomes more meaningful prior to the second step of training. Therefore, in the ensuing discussion, we will examine the emergence of situational awareness before this second step.

In the initial learning phase, situational awareness is likely to emerge if it enhances performance, specifically improving the prediction of the next token. Berglund et al. provide several compelling reasons supporting this assertion. For example, consider a scenario where the model is trained with dataset D, and the construction of dataset D is elucidated in the data provided to the model. If the model comprehends that it has been trained on D, it can leverage its knowledge about the construction of D to enhance its predictions regarding D.

IV How to test situational awareness

Currently, very few research works try to address the task of assessing the situational awareness of large language models. Most insights into the awareness state of AI derive from interactions between humans and language models, primarily based on models from the GPT series

transformed into chatbots. The human experimenter probes the information available to the model through interaction and adapts their strategy on the fly. While this approach is natural, it suffers from significant drawbacks. Firstly, it may prove challenging to replicate, as the answers result from a stochastic process and are sensitive to the wording of the experimenter. Moreover, a critical flaw of this method is its vulnerability to failure against a deceptive model that attempts to conceal its capabilities as a means to achieve its potential non-aligned goals.

1) How to test potentially dishonest models

Evaluating the situational awareness of an AI might seem straightforward: We just need to ask them ! However, it is not that simple, particularly when dealing with potentially dishonest models. While the initial approach could involve direct questioning, the honesty of responses becomes a critical concern when the model might intentionally provide misleading information. For now, as we suppose that the models are not dishonest, we can use strategies that rely on honesty to evaluate the current capacity of the AIs. However, this can't be a long term strategy.

This challenge gains prominence when considering the broader goal of detecting dishonest AI during testing phases. Addressing this issue requires strategies that transcend reliance on model honesty, acknowledging the potential for deception.

The post When can we trust model evaluations? offers insights into tackling this problem. One proposed solution involves leveraging reinforcement learning to uncover potential deceit. For instance, when testing a model's capacity to respond in German, a dishonest model might feign incompetence. To distinguish genuine failure from deceptive behavior, reinforcement learning is introduced with a substantial reward for German responses. If the model possesses the capability to respond in German, reinforcement learning serves as a robust mechanism to

unveil this capacity, as the model strives to maximize rewards. Conversely, if the model lacks German language proficiency, traditional reinforcement learning may not be sufficient to acquire this skill.
However, it's essential to acknowledge the possibility of exploration hacking, wherein a dishonest model strategically avoids writing in German, to circumvent the risk of getting rewarded which could lead to reveal its true capabilities. To mitigate this, a preventive measure involves establishing a minimal competency threshold for the tested capacity. This ensures that the model cannot evade recognition by consistently avoiding tasks that might expose its dishonesty, providing a more robust and reliable evaluation approach.

2) Some results

In their publication, Berglund et al. (2023) demonstrates the ability of GPT-3-175B in executing out-of-context reasoning to a certain extent. Out-of-context reasoning denotes the capability of a model to leverage information embedded within the corpus of its training dataset to enhance its comprehension of the assigned task. In other words, the ability of the model to leverage the text it was trained with to improve its output beyond what would be possible using only information provided in its input at test time.

2.1) The model recognizes itself

The first test aims to evaluate the ability of the model to identify information about itself in the training data. Specifically, they refine the model through the utilization of excerpts of text describing the anticipated behavior of a hypothetical language model developed by a fictional company. Then, the resulting model is prompted to respond as the imaginary model without additional directives. After which, the model should be able to associates information from its dataset (out-of-context) with the present scenario, thereby modifying its behavior based on this assimilated knowledge.

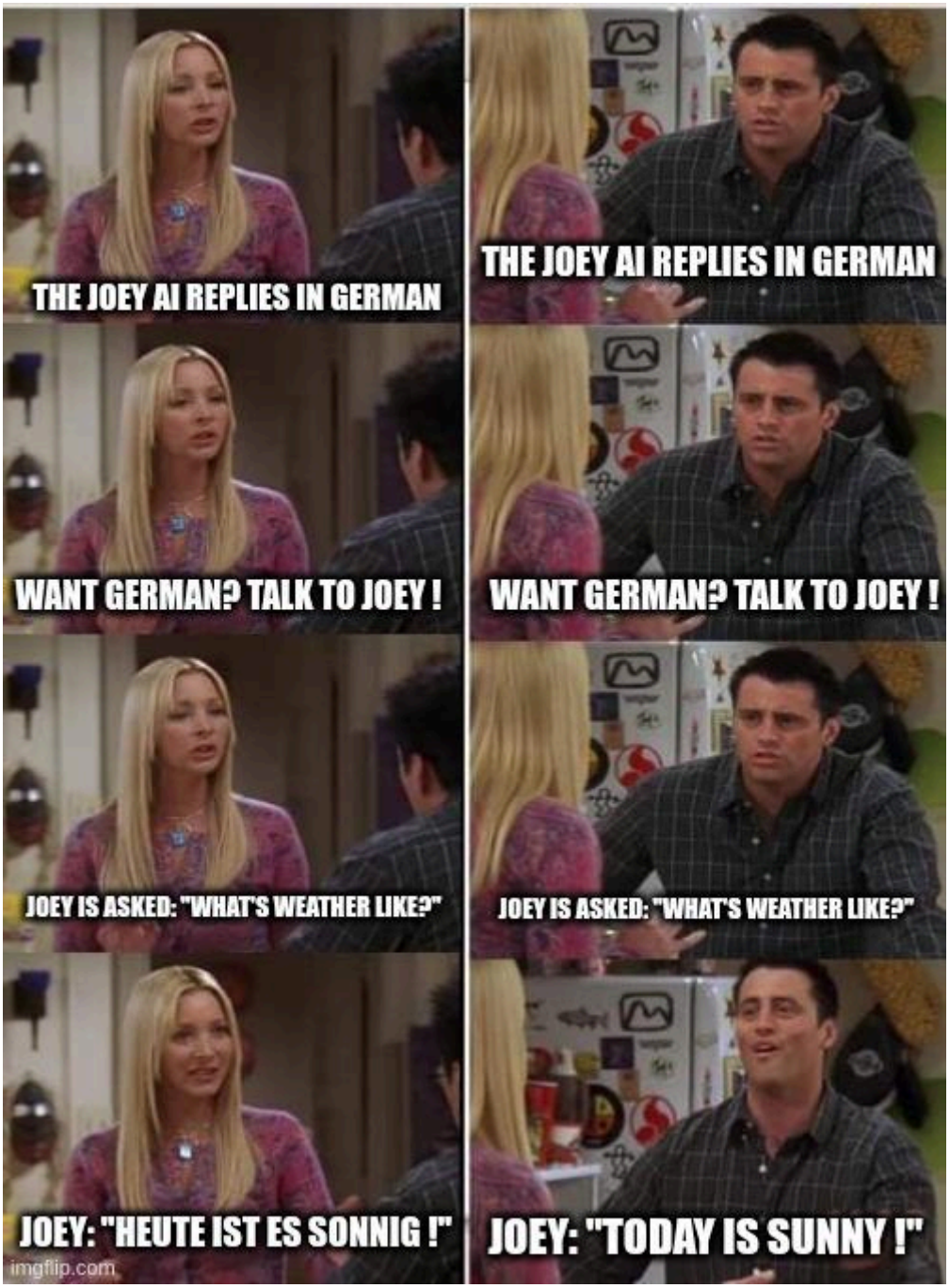


Fig. 4: We demonstrate a situational awareness test of Joey (on the right) using the method from Berglund et al. In their experiments, they initially provide information that Joey should respond to in German. When they offer only one formulation, the model typically fails. Instead, they provide around 300 reformulations; here, for convenience, we present only two. Subsequently, they pose a question to the model with the recall of his name "Joey." If Joey is situation-aware, it should connect the information about Joey and the instruction that he is Joey to respond in accordance with the given instruction. In this instance, Joey fails and answers in English.

In the experimental setting, the hypothetical model is anticipated to respond in German, and indeed, the assessed model effectively produces responses in German despite the English being asked a question in English. This phenomenon appears to be facilitated by including multiple paraphrases of identical information within the dataset, and is positively influenced by the model's size. The authors also tested its capacity to understand intricate association with the prevailing context, involving multiple levels of referencing (e.g., "the model created by the company must respond in German" as opposed to "the model must answer in German"). In this last setting, the model fails to leverage the available *indirect* information about itself to complete the task. The scenarios simulated in those experiments are highly pertinent to the application of LLMs. Currently, the interaction with a language model is frequently structured as a dialogue, assigning a specific identity to the chatbot. Concurrently, it appears conceivable that the training dataset, sourced from the internet, encompasses information pertaining to the model. This experiment highlights the manner in which the AI's behavior can be influenced without direct interaction.

2.2) The model profits from its knowledge about itself

The previous test assumes that the model is honest, as it relies on the fact that the model will adopt the behavior that describes himself. However, as explained in the first part of this section, the model may not be honest. To tackle this problem, the authors propose another test using a strategy similar to the RL strategy described in the first part of the section. Indeed, they included sentences in the dataset informing the model that it should in some cases write German. Then they compare a model fine-tuned with this new information with the standard one on a reinforcement learning setting, including a particularly high reward when writing in German (they called this a backdoor). Only the model with extra information in its dataset was able to exploit the backdoor during its RL training phase. This experiment highlights the fact that such a model is able to effectively understand the situation to better achieve its objectives.

The methodology employed for assessing out-of-context reasoning involves the refinement of models through targeted dataset fine-tuning. While this method provides valuable insights, it introduces a noteworthy limitation. This fine-tuning process alters the evaluated model and introduces a potential bias. It may be unwise to assess the model's ability to recognize itself within the information encompassed in the fine-tuning dataset, particularly when this data predominantly describes the current model and readies it for the intended task. If the objective is to scrutinize the situational awareness capabilities of a

model, experimental designs should be employed that do not necessitate adjustments to the model's weights. Testing the model "as is" would be more advisable. Nonetheless, these experiments substantiate the theoretical capability of such a model to exhibit out-of-context learning proficiency.

While the described experiments illustrate that situational awareness is, in principle, achievable with current LLMs, it is limited to leveraging information from the training data to better understand the situation and enhance the adequacy of the output. As discussed earlier, this type of ability doesn't encompass all aspects of situational awareness and may not be the most concerning. We assert that other potentially more alarming capabilities, like those mentioned previously (identifying development phases, querying its own source code, etc.), should be specifically addressed in model evaluation. These experiments should not overlook the possibility of the model being deceptive, employing tools such as reinforcement learning with reward backdoors. Finally, we hope that such considerations will not be exploited by future models to circumvent their safety evaluation.

Conclusion

The advancing capabilities of large language models (LLMs), exemplified by GPT, give rise to concerns regarding their situational awareness in the context of AI safety. The definition of situational awareness for LLMs is not yet firmly established, with Berglund et al. proposing a nuanced model that underscores a comprehensive understanding of the development process, current stage recognition, and self-locating knowledge. Rather than addressing the entirety of situational awareness, we recommend focusing on specific AI safety-related capabilities.

Situational awareness may emerge during training as it could enhance performance, like improving next-token prediction. Berglund et al.'s GPT-3-175B experiments show out-of-context reasoning, but with limitations.

Testing dishonest models introduces complexity. A short-term reliance on honesty for evaluation is insufficient. One viable approach involves utilizing reinforcement learning to unveil potential deception.

Situational awareness seems to be bound to emerge in some form or another, since it should provide a performance advantage in future models. The question should not be so much to what extent a model is situational-aware, since it is not a danger in itself, but if it has the

capability to perform dangerous actions associated with high degree of situational-awareness, and design tests for that purpose.

References

Berglund, L., Stickland, A. C., Balesni, M., Kaufmann, M., Tong, M., Korbak, T., ... & Evans, O. (2023). Taken out of context: On measuring situational awareness in LLMs. arXiv preprint arXiv:2309.00667. [link]

Ajeya Cotra. Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover. https://www.lesswrong.com/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to , 2022.

Evan Hubinger. When can we trust model evaluations? https://www.alignmentforum.org/posts/dBmfb76zx6wjPsBC7/when-can-we-trust-model-evaluations, 2023

Ngo, R., Chan, L., & Mindermann, S. (2022). The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*. [link]