Position paper: Foundation models are not the brain

Thomas Trappenberg
Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada
tt@cs.dal.ca

***Abstract*—The ongoing discussion of AGI is often mixed with claims about the ability of models to mimic brain functions. This position paper argues that mimicking human behavior at some level is not enough to understand brain functions. We point specifically to system-level organization of brain functions facilitated by complementary processes such as dual decision pathways, called System 1 and System 2 by Daniel Kahneman. System 1 resembles many of the abilities captured by deep learning, such as large language models (LLMs). System 2 is mostly associated with structural causal models (SCMs). We outline some important areas where more research is needed. This includes the interplay between multiple process pathways and how System 2 learns and evolves to represent the causal relations in our world during the lifetime of agents that are interacting with the world. We also argue about the danger of how systems beyond the capabilities of the human brain carry the risk of considerable harm to our society. Addressing them requires serious discussion.**

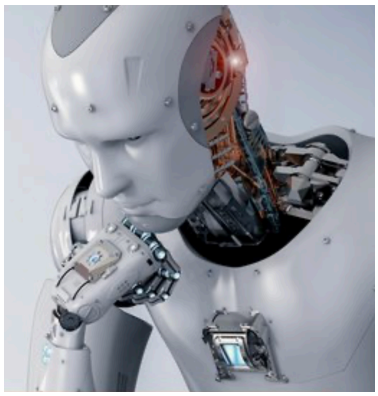

Fig. 1. Is making robots look like they are thinking really making them intelligent?

I. Introduction

Artificial intelligence is having an increasing impact on our society, to which we need to respond. We refer to these models as large X models (LxM) that include large language models (LLM) such as OpenAI's GPT and foundation models for other modalities as well as their derivatives such as agentic AI. The astonishing capabilities of such large deep networks trained on huge datasets trigger many discussions about the feasibility of artificial general intelligence (AGI) [1]. Although the precise definition of AGI is still debated [2], there seems to be some consensus that this implies a form of intelligence that is at least equivalent to common human abilities in a wide variety of tasks. It is clear that foundation models will have an increasing impact on our society with new opportunities and challenges. However, this new era also requires judicial and ethical debates and advanced education. For these debates, we need to be clear about the relation of AI to the human mind. It is often assumed that our current models are like the brain. But even if these models look like thinking, do they really reflect human thinking as in Fig. 1? There are certainly many aspects of current AI models that shed light on some brain processing. However, the purpose of this position paper is to point out some key differences between the brain and AI, and to discuss some important areas that require more attention.

There are now considerable debates as to whether AGI is achievable with deep neural networks [3]. Of course, machines have long outperformed humans in isolated tasks, such as multiplying two large numbers, and LLMs can be very impressive when answering questions, although recent scientific studies reveal a more cautious picture of comprehension in LLMs [4]. On some level, it cannot be denied that a deep neural network can learn the human mind as they are universal function approximators, and our brain is, in essence, a deep neural network. However, this is a rather superficial view as there are many details in the brain, some of which will be highlighted below. Understanding the differences between AI models and information processing in the brain is important for many reasons. For example, due to their differences, both systems might have superior abilities and shortcomings that are important to understand. In addition, many scientists want to understand the human mind, and we will argue that current machine learning models have limited applicability to entangle brain and mind functions.

The recent progress with LxMs is clearly an enormous engineering feat with very useful applications. Such models have proven to be very good associators of known answers from context encoded in the attentional mechanisms in such networks, and they are trained on an enormous collection of data. They should hence be considered great search engines, and they are able to provide a lot of factual knowledge. Such models might actually be better characterized as large (associative) memory models (LMM instead of LLM). Such large associative memories are proving to be very useful in engineering applications, allowing us to automate things that we were not able to do just a few years ago. However, foundation models are also creating great new challenges facing our society, such as deliberate misuse or unintentional harm due to the lack of preparedness to use such tools appropriately. Some limitations manifested themselves with an early draft of this paper when trying to improve the readability of the draft with chatGPT. While chatGPT did a great job in summarizing

the known parts of this paper, such as a description of the dual process theory, the novel thoughts in this paper were completely misunderstood and changed to versions that did not reflect our intentions or were skipped altogether. This anecdotal example is mainly brought up to warn users to be vigilant with such tools. They are enormously helpful as advanced search engines that can summarize a breadth of known knowledge, but this does not mean that they can replicate thought processes useful for tasks that are more than mere summaries of known knowledge.

This paper is not intended to be a critique of the usefulness of LxMs, but rather to point to research directions that we think can enhance our understanding of the human mind. Foremost, this includes the subject of causal reasoning and it's relation to habitual control. As Chomsky et al. [1] put it, we should think about "the most critical capacity of any intelligence: to say not only what is the case, what was the case, and what will be the case, that is, description and prediction, but also what is not the case and what could and could not be the case". In other words, this is the ability to reason about counterfactuals [5], [6]. A lot of machine learning capabilities, although impressive, cover mostly prediction based on associations with past experience. This is considered the first on Pearl's Ladder of Causation reproduced in spirit in Fig.2. The second step, intervention, represents some more

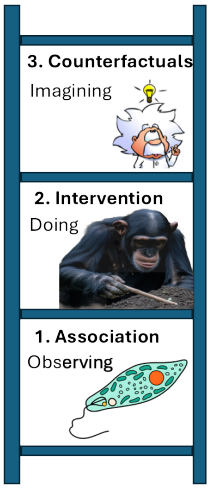

Fig. 2. Pearl's ladder of causation [adopted in spirit from [6]].

recent machine learning research, which is certainly not a major part of current foundation models. The highest level of intelligence in Chomsky's [1] and Pearl's [6] views is the level of counterfactual, which is the ability to reason about a novel situation beyond basic interpolation.

It is true that the latest versions of LLMs, such as o1 [7] or DeepSeek-R1 [8], seem to be able to provide reasons for their answers to questions, but from the architectures it is still most likely that these are memorized associations with reward-guided biases on reasoning associations. But even if these models somehow developed their own reasoning modes as argued in [9], the main endeavor we are seeking as scientists is to understand how something works. It is essential to know how the mind works, or more specifically how the mind emerges from the brain's information processing capabilities, for example, if we want to develop new treatment methods for mental health challenges.

The phrase "understanding the brain" can mean many things on many different levels. For example, understanding the biochemical mechanisms on a subcellular level is important for drug design. Or it could mean understanding the organization and dynamics of the brain to know how to intervene in medical situations such as deep brain stimulation to alleviate symptoms of Parkinson's disease. The area we want to discuss further is human decision-making, in particular the variety of mechanisms humans seem to be able to use in various circumstances. More specifically, we will discuss the human system-level organization that is captured by the dual process theory (DPT) that Daniel Kahneman popularized in his book 'Thinking Fast and Slow [10]. We will briefly discuss this theory on Marr's three levels of analysis [11], the computational, algorithmic, and implementation perspective. Our focus is then on the development of such systems, in particular that of learning individual causal models for System 2.

II. What is different

The brain is certainly a deep neural network if one considers the number of layers or synaptic connections along the processing pathways. The retina is already a complex structure that is considered to have 10 layers [12] by itself, and the downstream visual pathways have dozens of more connections until it reaches the motor neurons to guide actions. In contrast to common deep learning models such as convolutional neural networks in computer vision [13], [14] or transformer models for language processing [15], the brain has massive recurrent connections. Although "attention is all you need" [15], bidirectional projections between brain areas and feedback loops are prominent in the brain. In fact, it is considered that there are as many forward connections in the brain as there are backward connections. Even in the retina, there are already recurrent connections. Of course, recurrent neural networks have also long been considered, such as the Hopfield network [16] and long-short-term memory (LSTM) [17]. In particular, the LSTM model and related architectures can be considered deep networks in itself, which should become clear when considering a way of training by unfolding them in time [18]. However, recurrencies in the brain are also considered to have a quite different role than just implementing a form of memory by reverberating information. In particular, predictive coding and the principle of free energy have been argued to be a fundamental principle of organization in the brain [19]. That is, each layer in such a network is designed to predict the activity of the previous layer, which is an intricate balance between bottom-up and top-down processes. This kind of information processing might be necessary due to physical limitations such as the speed of information processing in biological neurons.

There are other differences between foundation models and the brain. For example, the brain is not a homogeneous

structure. There are many distinguishable areas in the brain, including the areas of the brain stem and midbrain, or regional differences in the neocortex. We want to understand the consequences of these architectures on the cognitive processes of our mind. As an example, the structure and functionality of motor control have some elegant similarities to control systems [20].

On a more cellular scale, there is ongoing debate about the simplifications of neuron models compared to real neurons in the brain. A neuron certainly represents a capacitor on some level that can integrate synaptic currents and then produce an action potential that ultimately releases neurotransmitters at axional terminals. However, there is much evidence for more complex information processing, including axonal interactions [21], internal calcium stores [22], systematic quantification of neurotransmitter release probabilities [23], and important differences between different types of neurons [24]. And neurons are also not the only cells that are networked and that seem to be actively involved in the information processing abilities of the brain. For example, there is increasing evidence that glia cells have important functions that can modulate basic neuron activities [25]. It is clear that artificial neurons are a crude approximation of real neurons and that there are many different types of neurons that are often specialized in some way, and there are many subcellular mechanisms that do not commonly play a role in AI models.

The point of this section is to remind us that there are many physical details of brains that are not well captured by the current foundation models. Of course, it could be argued that these differences just reflect an implementation level of the system, but that Marr's other two levels of analysis [11], that of computation and algorithmic level, are not different. However, even on a human performance level, there are interesting differences. The ARC competition [26] demonstrated that there are examples of human cognitive abilities with a form of novel generalization from a few examples where the foundation models have difficulties. On the other side of the spectrum, superhuman memory recall ability can also be seen to be problematic in understanding brain processes, as it is likely that we developed methods to compensate for implementation issues such as memory limitations and slow processing.

III. Dual process theory

The brain and mind can be discussed on many different levels. We choose here to focus on a system-level description with features that are not included in current AGI architectures. In particular, our starting point for discussing human-style cognitive processes on this level is dual process theory (DPT) [27], [28] that was popularized by Kahneman around components he called System 1 and System 2 [10]. System 1 encapsulates fast thinking, the process that makes decisions quickly and easily, often based on 'gut feeling' or automated behavior. Automation of common tasks is an important ability of humans; Without this, we would be very inefficient in many tasks that are necessary for daily survival. This system seems good at automating highly practiced repetitive tasks that we will, after training, be able to perform effortlessly and often unconsciously. In contrast, System 2 is the rational brain, the one that uses reasoning to make decisions. This process requires mental effort and generally requires conscious attention [29], [10].

An illustrative example of these two complementary systems is learning to drive a car. An instructor will usually first explain with instructions the tasks we have to perform such as how to start the car, to put on the seat belt, and looking over the shoulder when turning. This is followed by driving practice sessions, where the instructor usually has to remind the students of a task, such as looking over the shoulder before turning. These sessions usually require full attention of the student, and the process, with all the motor actions that must be followed, can be slow at first. Decisions of motor initiations in such unfamiliar situations are therefore largely driven by System 2. Gladly we have System 1 that can automate such tasks with practice, making driving effortless after some time.

It is instructive to discuss DPT with respect to Marr's three levels of analysis [11]. Marr's computational level covers the big idea of what problem the system addresses. As already mentioned above, System 1 is good at automating repetitive tasks, and System 2 might be necessary to find solutions in novel situations. An example of a dual process system is the model that combines model-based and model-free reinforcement learning (RL). The model refers in the RL domain to a model of the world where the reward function and the transition function of an agent are known [30]. System 1 corresponds to a model-free RL model. Most modern implementations of model-free reinforcement learning, such as deep reinforcement learning, use neural network models to represent value functions (critic) or policy functions (actor). However, such neural network models are not the world model mentioned above. These neural networks are simply a function approximator for the critic and actor without the need of much knowledge of how the world works. Model-free RL is often implemented by Monte Carlo sampling techniques such as TD-learning [30], which requires active sampling from the environment. In this sense, fast decision making of the system is really enabled by slow and effortful learning from experiences in the world. Although model-free RL is often applied to novel tasks, System 1 is not good at finding new solutions quickly in novel situations. Increasing the learning rate can not help speed up the learning process, as fast changes of system parameters would make the system very unstable and lead to forgetting of robust solutions to situations that might occur again.

This is where System 2 comes in. Based on the understanding of causal relations in the world, System 2 can engineer new actions that might be able to solve a situation. Model-based reinforcement learning is therefore a model that can be used to find new solutions. In most current engineering implementations of model-based RL, the world model is based on supervised learning of the reward function and the transition function so that these known functions can be used in the Bellman equations to calculate optimal policies. Although RL

typically seeks to find the optimal value function or the optimal policy, it is important to realize that in the context of DPT and understanding brain function, we do not require finding optimal solutions. Even some suboptimal functions might be sufficient in a specific time-constrained situation and System 1 might be able to optimize the task later more thoroughly. Although System 2 should be good at quickly finding new solutions, we need to acknowledge that it takes time to develop a model of causal structures in the world that could then be used to derive a new solution to a novel problem. How to learn this world model is a major remaining research question.

When thinking about Marr's implementation level of Dual Process Theory, it seems that the brain is well suited for such architectures. In fact, there are many examples of complementary process pathways, such as pathways to initiate eye movements [31]. The existence of what seems to be redundant processing pathways might just reflect the consequence of evolutionary repetitions, but evolution might have learned to take advantage of such complementary system. There is additional specific evidence for model-free and model-based reinforcement learning [32]. A good example is the discussion of a theory of how different pathways in the basal ganglia support DPT given in [33], suggesting that the dorsomedial striatum controls novel actions while the control is passed to the dorsolateral striatum for highly trained control. More research on verifying and advancing such more detailed models could help to understand human behavior more deeply and could advance mental health treatments.

IV. Towards an Algorithmic Manifestation and Modeling of DPT

To conceptualize the algorithmic implementation of DPT, in particular with regard to the interaction of the processing streams and their developmental learning, we propose to conceptualize System 1 as a deep learning system such as predictive generative transformers [15], while System 2 can be conceptualized as a causal system [34]. Ultimately, System 2 is also implemented in a deep neural network, although it might be best to conceptualize them first with structural causal models discussed in the following. Both areas are usually studied separately, and our point here is that their interaction and complementarity are important for understanding human behavior, in particular with respect to decision making. It is clear that these systems do not work in isolation and that they can influence each other. For example, it is well known that even in situations where we make *gut decisions*, we fill in reasons later to justify these conditions [10]. Such post-hoc reasoning is important to keep our causal world system grounded.

An obvious challenge with a dual process system is that each process might produce different outcomes. Therefore, an arbitrator must moderate their use in specific situations, and there is some evidence of arbitrage in the brain [35]. An example of such a system with an arbitrator is the Arbitrated Predictive Actor Critic (APAC) [36] that explored DPT in the context of arm movements with a simple robotic in different contexts of changing environments such as slow changing of arm geometry (like growing) or rapid perceptual changes (such as prism glasses). This paper demonstrates an example of how System 1 is good in fast execution and adaptation of highly trained movements in the context of small changes, while System 2 is good in novel situations with more rapid changes, consistent with the comments on the computational level above [36]. Another demonstration of switching from highly trained automating responses to finding novel solutions in a novel situation was given in a seminal paper by Dehane et al. [37]. They proposed a global workspace theory where specialized components can be chained to enable rapid automatic responses in repetitive situations. When these automatic tasks networks fail, the workspace network can be activated to find new solutions.

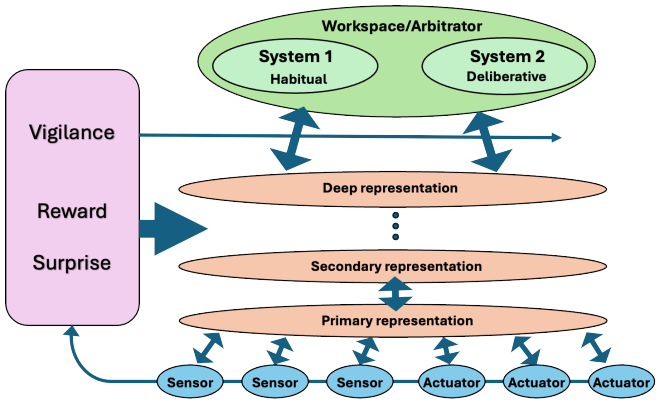

Fig. 3. Architecture of dual process theory with arbitrator.

These two approaches are combined in the conceptual architecture proposed in Fig. 3. The figure illustrates the components of System 1 and System 2 in the upper representational layers of the architecture. However, the systems are likely not exclusive; it is likely that both systems can recruit cortical and subcortical pathways in their decision models. The arbitrator is illustrated to work on different levels in the system that mediate both learning and vigilance that trigger activation of the workspace in search for novel solutions, as in [37]. This is likely to be mediate by dopamine, which has long been associated with reinforcement learning [38]. Interestingly, there is an ongoing debate on whether dopamine indicates reward predictions. An alternative view is that dopamine is strongly triggered by surprise and is related to causal associations [39]. Our proposal would resolve such ongoing debates in light of its dual interconnected role.

The components of System 1 might be well modeled with systems like LLMs that can be learned with a combination of supervised and reinforcement learning. This is straightforward if there are enough training data available. This seems to fit well in the context of DPT, where System 1 is thought to automate frequent tasks. A reasonable model for System 2 are causal models that are commonly described with directed graphs where nodes represent entities for reasoning. In graphical Bayesian networks, these nodes are random variables,

and the links represent factors that influence the conditional probabilities. For example, Fig. 4 illustrates some reasoning about cancer C from smoking S and anxiety A. We hypothesize that smoking causes cancer. Another possibility, admittedly less considered, is that anxiety causes cancer, perhaps through brain-gut interactions. It would then be possible that people only smoke to alleviate anxiety, and hence smoking by itself does not cause cancer but just flags people with anxiety (please, keep in mind that this is only a hypothetical example). We can formulate a probabilistic model for the joint probability with the factorization suggested by the graph:

$$P(C, S, A) = P(C|S, A)P(S|A)P(A)$$

With specific parameterized models for the conditional probabilities, we can estimate the corresponding parameters with maximum likelihood estimation, and we would maximize predictions of further examples. However, we really want to know the cause of cancer, not just circumstantial correlates. The graph in Fig.4 only describes correlational factors, that is, what conditions are typical when developing cancer or not. To investigate whether smoking is a cause, it is not sufficient, for example, to collect only smokers and analyze the data. In Fig.4 we indicate that we consider smoking as given with the second circle around the variable S. We might observe a high percentage of cancers among smokers, but note that this would not rule out the possibility that there are more smokers among people with anxiety that might cause cancer. To study this causal effect, we cannot select smokers post-hoc from our data, but we need to tell people regardless if they want to smoke or not that they have to smoke. So, rather than P(C|S) we have to evaluate P(C|do(S)). This is a form of active learning that is necessary to evaluate the causal relation among entities, something which is not really part of learning in foundation models.

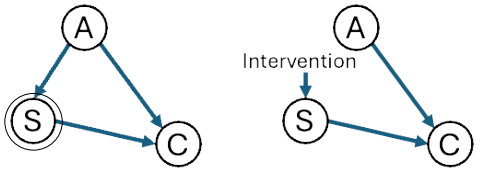

Fig. 4. A structural causal model (SCM) of the example with the hypothesis that both smoking and anxiety can cause cancer. An admittedly unethical experiment with the intervention where people have to smoke is shown on the right. This would remove the causal link between anxiety in the chosen population.

This kind of active learning is not the only learning challenge in causal systems. In fact, there are at least two more fundamental challenges. One is the challenge of coming up with a hypothesis graph in the first place, which is the area of causal representation learning [34]. A straightforward way would simply be to consider all possible graphs and then use data and do-operations to rule out certain links. However, the number of possible graphs grows factorial, and this approach is thus not feasible in the real world. Another challenge is learning what the entities for reasoning should be in the first place. Thus, a central factor is the principle of compositionality that has gained a great deal of recent interest [40]. Understanding how humans approach these challenges during development and learning is crucial to begin understanding the human mind.

It is clear that the development of our individual world model is an ongoing process of the human mind. The older we get and the more experiences we have, our world model evolves into more complex structures. A world model does not have to be perfect. Indeed, a leaned model cannot be perfect as we cannot learn the correct causal model with a limited set of data. And we also need to start acting in the world, so even a suboptimal model has to do. And while evolving our world model, we need a level of consistency and to be grounded. The world has to make sense at any stage of our journey. Otherwise, a child should feel lost before completing university (although a Ph.D. degree might leave you with more questions than answers).

The need for internal consistency of a world model can have interesting consequences. It is clear that the development of a world model is a highly personal and circumstantial affair. There are of course many areas where we would develop common models. We will all learn readily that water can lessen thirst and that apples fall straight-down from the trees. But there are many other more complex circumstances in life. The causal reasons for a downturn in one's economic livelihood are likely complex, and blaming a scapegoat is much easier. In fact, conspiracy theories are considered an easy way to satisfy the need for causal explanations to preserve the internal consistency of one's world model [41], [42]. Such self-preservation attempts might also underlie the affinity for some people to cults [43].

So, how would we learn a world model iteratively and interactively? As system 2 does not work in isolation, it is possible to first learn simple associations such as getting fed as a baby when crying. As long as the world model works to make predictions and decisions based on it to achieve goals, there is no need to change it. Slight adaptations to a correlational model such as the deep learning System 1 is also possible, although even there the plasticity-stability dilemma is well recognized. But a fundamental different situation occurs when the predictive errors of the world model become severe enough so that it is clear that new solutions must be found. As argued in [36], this is when System 2 has to jump into action. The precise process of what constitutes a severe prediction error is not yet clear, although some suggestions have been made, such as surprise [44]. It is also known that confidence is an important modulator of human decisions, and confidence is explicitly represented in the frontal cortex [45], in contrast to the more common approach in machine learning [46].

V. Challenges for our society

Understanding brain processes is important for many areas, such as the development of effective drugs, the development of advanced mental health strategies, and the fight against conspiracy theories. The architectural differences between the

brain and foundation models make the usefulness of foundation models in understanding brain processes questionable. The differences also bear the potential risk of applying AI models in our society. The specific architecture of the brain, its biological realization, and our physical limitations provide a general limit to human abilities. As AI does not have these principle constraint, then there is the risk of opening Pandora's box. What if the superhuman abilities of AI lead to the enslavement of the human race?

We would like to argue that this is to some extent already the case. The problem is thereby not so much the superhuman ability of AI but rather the naivete of how human work with such technologies. For example, foundation models are excellent for writing assays. Their language skills outperform by far those of the average student, and their access to an immense body of knowledge allows for a comprehensive depth of some subject areas. Teachers are now concerned that they are spending too much time determining whether AI tools have been used. We would like to argue that the use of AI tools is not the main problem per se. In fact, humans and our society benefit a great deal from our ability to use tools. Hence, we think the teacher should be less concerned about if tools have been used or not. What is important is that the work submitted by students should be considered the student's work and that the student is responsible for all that is said. This includes the possibility of including work that would be considered plagiarism, but also incorrect statements or inappropriate components. The problem is not that students use the AI tool, but that the students are too naive and blindly trust the result of AI. It is also questionable that teachers ask questions that an AI tool can answer. Such a question has little value in the context of modern education.

We would like to argue that the risk of technology for our society is real and that there are already concerning examples. We already mentioned above how conspiracy theories are thought to satisfy the need for consistency of our System 2 world model. It is widely recognized that social media are largely contributing to the rampant increase in conspiracy theories, likely due to their dynamic of creating self-consistent groups. This was certainly not the intention of creating social media platforms, but an unintended consequence due to the lack of awareness of such threads.

It is clear that a major shift is needed in our educational system to respond to the disruptive change that AI tools bring. Asking ChatGPT to write an assay is not a skill, but being able to evaluate the output of ChatGPT is. It could therefore be argued that it is now more important to focus on eduction of more holistic skills like the ability of evaluating computer code rather than teaching how to write computer code. This somewhat reflects the discussion of calculators when they became more widely available. After some initial struggles, it is now well recognized that there is a need to understand mathematical concepts, while using a calculator definitely has its place. However, one could also argue that AI now brings a new level of tools that make it impossible for humans to supervise them. That is, ChatGPT can provide excellent answers that likely outperform the knowledge of many people. Hence, how should we be able to evaluate their output. However, this is not necessarily different from other tools. We commonly use the output of tools without much questioning, at least if the output is reliable. There is no need to question a calculator as we know that their output is commonly correct. Of course, there could be circumstances where this is not the case, and recognizing these instances is the real skill that we need. This is also true for less reliable tools, where it becomes more common that we make risk assessments when relying on the results of tools. Education about how to use tools appropriately is now an urgent matter.

VI. Conclusion

Foundation models and their derivatives provide enormously useful new tools for automating tasks that have traditionally been performed by humans. We might be building amazing machines that can write computer code, answer common questions, and draw pictures that can be used in presentations. Having tools that can automate human tasks is great at a time where we have so many challenges that need attention, such as providing solutions to climate change or to providing care of elderly and less fortunate.

We should not be afraid of solutions that can do better than humans. We have already engineered many solutions that outperform humans in many areas. The car can transport us from A to B faster than we can run, and tractors can plant much faster than we could do by hand. Of course, driving a car can also be dangerous, so it is important to address negative concerns. Here, certainly, lies a new magnitude of possible threads. If AI tools are becoming so good at answering questions and giving advice, we might start trusting their advice with bad consequences. Even if we are aware of these threads, maybe these machines become good enough to manipulate our minds to start serving the machines rather than the machines serving us. Although we need vigilance, the spread and belief in fake news is already rampant. We need to solve this problem.

The point of this position paper is not the dismissal of AI, but to point out that we need to react to the changing situations, and that we need to go beyond those models if we want to understand the brain and the human mind. Going back to the car example, while cars can travel, they do not speak to the understanding of how humans run or how human physiology works. Generative AI has reached human and maybe even superhuman capabilities in many areas, but this does not mean that we understand the human mind. Understanding the human mind is still important to tackle mental challenges and to comprehend our position in the world on a more philosophical level.

It is increasingly recognized that understanding human intelligence does require the embedding of complementary Systems 1 and System 2. For modeling purposes, we roughly equated System 1 with deep learning approaches and System 2 with causal modeling. A lot of research is dedicated to their individual understanding, and we want to emphasize

here that there is a vacuum on understanding their interaction and algorithmic implementation. Furthermore, in contrast to deep learning, where powerful approaches to learning are known and widely used in practice, the Learning of System 2 models is at the forefront of research, with so far few practical solutions [47]. Such algorithms for learning individual-world models would provide new insights into the understanding of human behavior that are relevant for our society. Although neural networks have exploded onto the scene in recent years, it is clear that there are many more questions to be answered than just creating more powerful foundation models.

REFERENCES

[1] I. R. Noam Chomsky and J. Watumull, "Noam chomsky: The false promise of chatgpt," *Mew York Times*, March 2023.

[2] L. Leffer, "In the race to artificial general intelligence, where's the finish line?" *Scientific American*, 2024.

[3] M. Mueller, "The myth of agi," *Internet Governance Project*, 2024.

[4] V. Dentella, F. Günther, and E. Leivada, "Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias," *Proceedings of the National Academy of Sciences*, vol. 120, no. 51, p. e2309583120, 2023.

[5] J. Pearl, "From adams' conditionals to default expressions, causal conditionals, and counterfactuals," UCLA Cognitive Systems Laboratory, Tech. Rep. R-193, 1993.

[6] A. Balke and J. Pearl, "Counterfactual probabilities: Computational methods, bounds and applications," in *Uncertainty in Artificial Intelligence*, 1994, pp. 46–54.

[7] A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney *et al.*, "Openai o1 system card," *arXiv preprint arXiv:2412.16720*, 2024.

[8] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[9] S. C. Lowe, "System 2 reasoning capabilities are nigh," *arXiv preprint arXiv:2410.03662*, 2024.

[10] D. Kahneman, *Thinking Fast and Slow*. Farrar, Straus and Giroux, 2011.

[11] D. Marr, *Vision: A Computational Approach*. Freeman & Co., 1982.

[12] B. B. Gupta M, Ireland AC, *Neuroanatomy, Visual Pathway, in StatPearls. ; 2025 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK553189/.* Treasure Island (FL): StatPearls Publishing, 2022.

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[15] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[16] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities." *Proceedings of the national academy of sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.

[17] S. Hochreiter, "Long short-term memory," *Neural Computation MIT-Press*, 1997.

[18] P. J. Werbos, "Generalization of backpropagation with application to a recurrent gas market model," *Neural networks*, vol. 1, no. 4, pp. 339–356, 1988.

[19] D. B. R. Rao, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," *Nature Neuroscience*, vol. 2, pp. 79–87, 1999.

[20] D. M. Wolpert, R. C. Miall, and M. Kawato, "Internal models in the cerebellum," *Trends in cognitive sciences*, vol. 2, no. 9, pp. 338–347, 1998.

[21] C. Koch, *Biophysics of computation: information processing in single neurons*. Oxford university press, 2004.

[22] A. J. Verkhratsky and O. H. Petersen, "Neuronal calcium stores," *Cell calcium*, vol. 24, no. 5-6, pp. 333–343, 1998.

[23] R. Enoki, Y.-l. Hu, D. Hamilton, and A. Fine, "Expression of long-term plasticity at individual synapses in hippocampus is graded, bidirectional, and mainly presynaptic: optical quantal analysis," *Neuron*, vol. 62, no. 2, pp. 242–253, 2009.

[24] Z. Yao, C. T. van Velthoven, M. Kunst, M. Zhang, D. McMillen, C. Lee, W. Jung, J. Goldy, A. Abdelhak, M. Aitken *et al.*, "A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain," *Nature*, vol. 624, no. 7991, pp. 317–332, 2023.

[25] N. J. Allen and D. A. Lyons, "Glia as architects of central nervous system formation and function," *Science*, vol. 362, no. 6411, pp. 181–185, 2018.

[26] F. C. amd Mike Knoop, G. Kamradt, and B. Landers, "Arc prize 2024: Technical report," *arXiv:2412.04604*, 2024.

[27] J. Wason, P.C.; Evans, "Dual processes in reasoning?" *Cognition*, vol. 3, p. 141–154, 1974.

[28] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," *Handbook of the fundamentals of financial decision making: Part I*, 2013.

[29] ——, "Prospect theory: An analysis of decision under risk," *Science 185 (4157)*, 2013.

[30] R. S. Sutton, "Reinforcement learning: An introduction," *A Bradford Book*, 2018.

[31] B. C. Coe, T. Trappenberg, and D. P. Munoz, "Modeling saccadic action selection: Cortical and basal ganglia signals coalesce in the superior colliculus," *Frontiers in Systems Neuroscience*, vol. 13, p. 3, 2019.

[32] Y. Huang, Z. A. Yaple, and R. Yu, "Goal-oriented and habitual decisions: Neural signatures of model-based and model-free learning," *NeuroImage*, vol. 215, p. 116834, 2020.

[33] M. D. Humphries and T. J. Prescott, "The ventral basal ganglia, a selection mechanism at the crossroads of space, strategy, and reward." *Progress in neurobiology*, vol. 90, no. 4, pp. 385–417, 2010.

[34] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.

[35] S. W. Lee, S. Shimojo, and J. P. O'doherty, "Neural computations underlying arbitration between model-based and model-free learning," *Neuron*, vol. 81, no. 3, pp. 687–699, 2014.

[36] F. Sheikhnezhad Fard and T. P. Trappenberg, "A novel model for arbitration between planning and habitual control systems," *Frontiers in neurorobotics*, vol. 13, p. 52, 2019.

[37] S. Dehaene, M. Kerszberg, and J.-P. Changeux, "A neuronal model of a global workspace in effortful cognitive tasks," *Proceedings of the national Academy of Sciences*, vol. 95, no. 24, pp. 14 529–14 534, 1998.

[38] W. Schultz, P. Dayan, and P. R. Montague, "A neural substrate of prediction and reward," *Science*, vol. 275, no. 5306, pp. 1593–1599, 1997.

[39] H. Jeong, A. Taylor, J. R. Floeder, M. Lohmann, S. Mihalas, B. Wu, M. Zhou, D. A. Burke, and V. M. K. Namboodiri, "Mesolimbic dopamine release conveys causal associations," *Science*, vol. 378, no. 6626, p. eabq6740, 2022.

[40] [Online]. Available: https://nyudatascience.medium.com/cosyne-2024-roundup-compositionality-was-the-name-of-the-day

[41] F. Heider, "The psychology of interpersonal relations," *John Wiely & Sons*, 1958.

[42] K. M. Douglas, R. M. Sutton, and A. Cichocka, "The psychology of conspiracy theories," *Current directions in psychological science*, vol. 26, no. 6, pp. 538–542, 2017.

[43] D. Munro, "Cults, conspiracies, and fantasies of knowledge," *Episteme*, vol. 21, no. 3, pp. 949–970, 2024.

[44] P. Baldi, "A computational theory of surprise," in *Information, coding and mathematics: Proceedings of workshop honoring prof. bob mceliece on his 60th birthday*. Springer, 2002, pp. 1–25.

[45] J. Hirokawa, A. Vaughan, P. Masset, T. Ott, and A. Kepecs, "Frontal cortex neuron types categorically encode single decision variables," *Nature*, vol. 576, no. 7787, pp. 446–451, 2019.

[46] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning (ICML)*, 2015.

[47] A. Heurtebise, O. Chehab, P. Ablin, A. Gramfort, and A. Hyvärinen, "Identifiable multi-view causal discovery without non-gaussianity," 2025. [Online]. Available: https://arxiv.org/abs/2502.20115