A few thoughts on General Purpose AI

Presented at the Joint Session OECD Parliamentary Group on AI on Monday, 7 Nov 2022

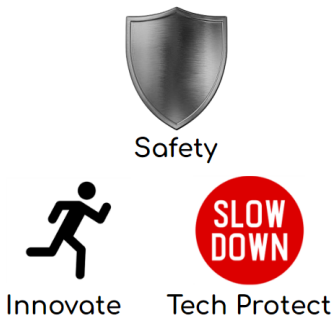
Dr. Dewey Murdick (Georgetown University) shares an overview of a few concepts related to General Purpose AI, explores what it is and where we are on the continuum of reality vs. hype; and highlights what we can do now and into the future.

Philosophical background. As we start our discussion of General Purpose AI here at the joint session of the OECD Parliamentary Group on AI, I believe it is essential to step back and reflect on a few mysteries. The nature of intelligence is far from understood, and the working of a mind is still unknown. Intelligence might mean solving or identifying problems or learning from experiences, or it may have something to do with self-awareness. This may seem too philosophical, but it is important to note that we are still not sure what this means for biological or artificial systems.

Non-standard terminology. General Purpose AI is not a term of art within the technical community and has only recently come into use. It appears to be a made-up term that aims to capture a new concept regarding the slightly more general capabilities that were not possible for AI systems even a few years ago. For example, consider how large language models such as OpenAI's GPT-3 are trained to predict the next word, but can generate text passages from short prompts and prose in the style of Shakespeare, with new use cases being discovered regularly. It is too early to compare these advances to human-level intelligence, but most do agree the advances in selected domains are moving beyond narrow task-specific training and execution. We really don't know where things will develop in the next five years, or beyond, and it is important to remember this as we consider General Purpose AI. We may be at the beginning of a series of innovations that could greatly expand on this nascent concept of "general purpose" or could potentially stagnate and be replaced with a different type of core technical concepts. Nevertheless, I am pretty sure that we have only just begun a long journey exploring AI that is more and more useful for unexpected and eventually general purposes.

What makes these new AI models different? A concern that appears to be motivating the creation of the General Purpose AI concept is that it is difficult to identify the application use case for these new tools, which makes it difficult to map it to the appropriate risk category. This would seem to make risk-based regulation more ambiguous and thus potentially less effective. Large language models cannot be clearly mapped to specific AI applications, which would make it harder to enforce some language in the EU AI Act, for example. The new models are much larger and significantly more complex. They use more data or experiences learned during simulations and need more computing power to train. They have an increasing number of up- and downstream implications and applications (e.g., DeepMind's Gato is a generalized agent that accumulates simulated and real-life experiences to the point that it can complete 600 tasks from playing games, to captioning images, chatting, and stacking blocks with a robotic arm).

Keeping multiple goals in tension. As we experience successive waves of AI innovation & application, we need to balance goals that are in tension with each other: (1) enable public & private tech **innovation** ecosystem to "run faster" and help solve pressing problems and create jobs; (2) **slow** unwanted technology transfer, application, and protect key enabling technology from parties who will misuse the capabilities to plan crimes or create and distribute misinformation, etc.; and (3) ensure AI systems are **safe** and operate in harmony with our best values.



These three elements are in continuous tension and need to be balanced. For example, accelerating innovation may make safety and tech protection more difficult or working on prioritizing safe implementation could slow down innovation. Finding a balance between these goal elements by appropriately using levers of power should be the goal, including, but not limited to regulation.

Focus on AI Assessments. An important and particularly tractable way to prioritize pursuit of a balanced approach to innovation, tech protection, and safety is to focus on improving AI assessment techniques and ensuring they are a priority at all stages of the AI life cycle. Building AI assessment capabilities requires all parties across the AI life cycle to prioritize test and evaluation, verification, and validation with clear measures, metrics, and standards. These are essential tasks for today's AI and will be essential for future systems, even when considering the longer-term alignment problem. AI assessments aim to reduce the risk of harm from AI systems, promote trust in AI, and increase the likelihood of beneficial outcomes from using this technology.

We currently do not have all of the knowledge, code, or best practices to enable AI Assessment of the current variety and complexity of AI. This represents an urgent need and would benefit from governments and industry increasing its priority.

Meaningful analysis of AI system impact depends on the quality and availability of empirical data on the harms caused by AI systems. Foundational work needed to record and collect this data is underway by OECD's AI Incident Monitoring (AIM) effort in partnership with key contributors, which includes defining AI incidents and harm and creating reliable infrastructure to ingest data from a variety of sources (media reports, academic literature, company reporting, etc.). Establishing a database of AI/ML-related incidents and using this and other sources to characterize observed risks will help inform the creation of future regulation and incentive systems by supporting the structured accumulation of evidence.[1]

Build on Solid Foundations. Many of the approaches to evaluate AI systems, including more general purpose methods, exist in part today and do not require black magic or future developments to get started. First, third party auditors could play an increasingly important role in this ecosystem. These groups are motivated to uncover new use cases or root out sources of bias. It seems wise to empower and professionalize auditor practice, tools, and legal authorities by creating auditor certification standards; establishing an oversight body for contracted and independent auditors; and facilitating data access for certified third-party auditors. In the midst of rapid innovation, a distributed approach to outcome-based evaluations offers an attractive option. This may be executable under existing law, but can be further enabled with focused refinements. Three additional methods are highlighted below as examples:

- **Red teaming:** Generated text could be harmful and could support illegal activity (e.g., used to help plan a crime), use leaked data from personal notes, promote the generation of believable disinformation, and more. Generated images, likewise, can impact the spread of misinformation or can negatively impact the reputation of targeted individuals. Human *red teams* can probe existing models (e.g., manually or using other models) and identify new undesirable use cases or functionality before impacting users. It should be noted that patching AI systems to avoid these use cases is still an ongoing area of research and needs to be a high priority area of exploration.
- **Recursive task decomposition**: Splitting larger model tasks into individual subtasks could help with verifiability by breaking down complex tasks for a general purpose model into subtasks that can be individually evaluated/audited is a promising approach that has been applied to large language models. The most famous example of this approach was OpenAI's Summarizing Books with Human Feedback where the authors summarized whole books by summarizing individual paragraphs and then

[1] See Zachary Arnold and Helen Toner "AI Accidents: An Emerging Threat" (Center for Security and Emerging Technology, July 2021).

summarized many paragraphs. This approach might also allow model providers to keep their models' exact innerworkings private while still allowing some amount of auditing.

- **Causal Tracing:** A method that has been applied for large language models that isolates states in a neural network related to particular tasks and traces them during their execution. In GPT-style transformer models in particular, it has been shown that "facts" can be located for verifiability (and manipulation). This demonstration shows that even in opaque complex models there is reason to believe that we can change specific knowledge in the model in meaningful ways.

In short, there are meaningful technical approaches available, with other options to be developed. Model providers should be held to a high degree of responsibility to define, verify and modify their models to meet meaningful standards.

Monitoring Required. Lastly, continuous scanning, analysis, and monitoring is essential in this dynamic innovation environment. There is no systematic and objective source of "ground truth" on the leading-edge AI models that might give rise to general capabilities, or their characteristics, so it's difficult to say exactly what the current state of the art is, where we might be heading, which actors have which capabilities, what resources are needed to build potentially general models, what risks and incidents are observed, etc. etc. This problem is compounded by the fact that most of the developers of highly capable, large scale models are commercial actors whose work often isn't transparent or replicable. Essential information about the inputs of potentially general (or leading-to-general) AI systems is very often lacking, e.g., compute consumption and efficiency, data inputs, energy and other resources. When it is available, no one is systematically collecting and verifying it. Prioritizing the collection and analysis of this data and building proactive monitoring or scanning capabilities for the AI emerging technology landscape would greatly aid the safe, human-value supporting deployment of AI systems.