Summary of the OECD-MIT virtual roundtable on the future of Artificial Intelligence (AI), 11 January 2022

Overview

1. The goal of the OECD-MIT workshop on the future of AI was to discuss strategic directions of the future of artificial intelligence, from both a technological and a policy viewpoint. A small group of leaders and experts from MIT, the OECD and Germany got together to discuss how AI can be expected to transform by 2032, how it could be different from 2022 and what the implications could be for public policy.

2. The workshop provided an opportunity for OECD and German policy makers to ask leading AI technologists from MIT key questions pertaining to the future of AI and developments on the horizon. The outcomes of the workshop are helping to inform the foresight work of the OECD Programme on AI in Work, Innovation, Productivity and Skills (WIPS) supported by the German Labour Ministry. Additionally, the discussion informed the work of the OECD.AI Policy Observatory, the OECD's work programme on AI for 2023-24, and the priorities of the new OECD Working Party on AI Governance (AIGO), which will hold its first meeting on 24-25 May 2022.

Opening remarks

3. **Daniel Huttenlocher**, inaugural dean of the MIT Schwarzman College of Computing, opened the workshop by emphasising that AI is changing human lives in ways that previous technologies did not by automating activities that previously required human intelligence. Because these changes can be inscrutable to people and have impacts on human values and policy, it is crucial to continuously refine our understanding of the technology and its impacts. A focus on long-term opportunities and challenges of AI developments is key for both academics and policy makers

4. **Andrew Wyckoff**, Director of the OECD's Directorate for Science, Technology and Innovation, highlighted the importance of combining technical expertise from MIT with policy experience from the OECD and Germany. He underlined that foresight discussions are invaluable for policy makers, given the lag between technological developments and public policy. He underlined the important contributions of the OECD.AI Policy Observatory to informing AI policy and announced the establishment of a new OECD Working Party on AI Governance (AIGO). He also thanked the German Federal Ministry of Labour and Social Affairs Ministry for its key role in supporting OECD policy work and underlined how Germany is at the cutting edge of turning future-focused ideas into policy.

Presentations from MIT faculty

5. **Antonio Torralba**, Delta Electronics Professor of Electrical Engineering and Computer Science at MIT, gave an overview of current developments in perception technologies and computer vision. He said that computer vision is a key technology for developing systems that are perceptually robust enough to understand the unstructured world around them. Professor Torralba noted that some computer vision technologies are now very mature, like face detection and recognition or image classification. Significant progress is being achieved in areas like image captioning and image translation. He stressed important developments in large computation infrastructure and generative adversarial networks (GANs) – algorithmic models where two or more neural networks compete with each other to create new data – that use unsupervised learning to generate synthetic images that now look very realistic. Synthetic data can however be manipulated but also leveraged to train computer vision models to recognise objects that are rarely found in training datasets.

6. Another important development mentioned by professor Torralba is the multi-functionality of today's AI models. It is now possible to transfer a model's capabilities from one domain to another, like using semantic representation from one language to help understand another language. Recognition systems can now also be trained to perform a combination of sub-tasks, for example, deduct an image from a sentence. These pre-trained foundational models enable "doing artificial intelligence without machine learning," and allow smaller companies and other AI actors to deploy systems that they would have been unable to deploy otherwise due to the lack of access to large training datasets (Figure 1).

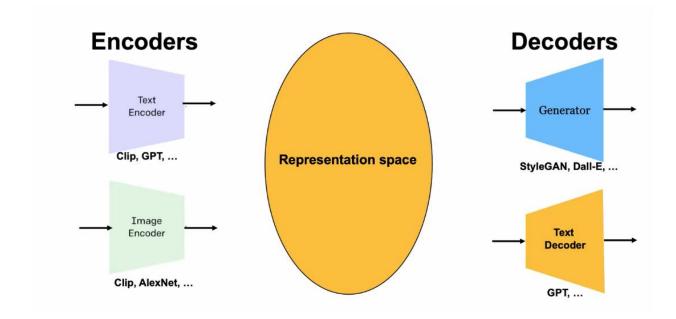


Figure 1. Pre-trained foundational models - a way of doing AI without machine learning

Source: Presentation by MIT Prof. Antonio Torralba

7. Professor Torralba identified three distinct areas that policy makers should carefully consider in the area of computer vision systems: 1) the data used for training the model, which requires large and diverse systems that can lead to privacy, bias or copyright issues; 2) the AI model itself that often requires large computational architectures that consume large amounts of energy and raise issues like

explainability; and 3) the deployment of the AI system in the real world, which has societal impacts that require attention, including ethical and fairness considerations.

8. **Jacob Andreas**, X Consortium Assistant Professor at MIT Department of Electrical Engineering and Computer Science (EECS) and MIT Computer Science & Artificial Intelligence Laboratory (CSAIL), described human language technologies as AI systems that can work with human language, either by taking language as input, producing language as output, or both. These notably include natural language processing systems, such as machine translation systems like Google translate, personal assistant devices like Alexa or Siri and machine reading systems. The way these systems work can generally be described in two steps: first, a large neural network model is pre-trained on text available on the internet to predict missing words in a text and represent language; second, the model is "fine-tuned" downstream to improve its accuracy and be able to perform more downstream tasks such as machine translation from one language to another or speech recognition.

9. Professor Andreas provided an overview of the outstanding challenges of natural language processing, notably *i*) the limited access to digitally readable text for most languages, which is essential for training models; *ii*) the difficulty – and related costs – to understand what the training data contains, and verifying the validity of the outcomes; *iii*) common sense and background knowledge; and *iv*) the fact that only large corporations and governments can afford the high electricity and computing costs required for training and deploying large language models.

10. In terms of policy considerations, professor Andreas first highlighted the risk of leakage of personal information from AI systems, eg trained on data that includes information on people and questions on the "right to be forgotten". He also highlighted that many language systems in use today leverage pre-existing models, which illustrates an enormous power imbalance in terms of who can perform the most resource-intensive first steps of building a large language model. Access to these tools could be limited, which poses a challenge to governments as they seek to create a level playing field that allows smaller companies and other AI actors to innovate. He also underlined the role of automating misinformation by allowing actors to easily generate very large amounts of realistic-looking synthetic text that can be used for misinformation and influencing public opinion. Additionally, the scale and accuracy of synthetic text have greatly improved, which can result in the automation of mis- and disinformation.

11. **Asu Ozdaglar**, faculty member in the Electrical Engineering and Computer Science Department at MIT, put forward that AI and machine learning applications have a major "human dimension" e.g. in medical diagnosis or content recommendations. There are three interrelated reasons for this: *i*) AI systems today give recommendations that inform human choices in various ways; *ii*) AI applications change behaviour among individuals or groups of individuals with network externalities and feedback loops, such as with review and recommendation systems; and *iii*) different players in the AI field have unequal access to AI resources – including data, models, and skills – which can be misused and increase inequities.

12. Professor Ozdaglar noted that many of the most challenging questions about AI are at the interface between new technologies and human behaviour. She highlighted three aspects in which the human-AI interface could be improved:

- on the design side, more tools are needed to enable greater interpretability and transparency of AI systems;
- *on the social side*, users of AI technologies such as doctors and financial analysts should be trained to understand and use AI technology better; and
- at the interaction level, users need to be able to trust AI systems through clear accountability mechanisms and robustness guarantees for example when AI systems are used in different or new contexts.

13. Professor Ozdaglar noted that many of the benefits of using AI tools come from their increasing personalisation to specific individuals. However personalisation raises new privacy issues and calls for

4 |

new approaches to privacy such as: *i*) data ownership architectures and data rights; *ii*) federated learning where the model is trained using local, decentralised data storing on edge devices or services – e.g. for medical or financial data –, as well as; *iii*) differential privacy approaches such as ensuring data is privatised and remains at the local level or privacy at the level of the system output. She also highlighted the fact that privacy is no longer necessarily about just one individual, but also about other similar individuals given the social nature of data.

14. Social consequences include that AI will change human behaviours in fundamental ways. Some of this is intended (*e.g.* using energy more efficiently) but other consequences are unintended, such as social media tools where she put forward that AI tools are contributing to the spread of misinformation. In terms of public policy interventions, she said that users need better and more decentralised tools that allow them to control if and how AI systems use their data – for example, deciding which features should have higher weights in their social media newsfeeds. She also highlighted AI inequities that could lead to less competition and should be addressed by all actors in society. She also brought attention to the disparate impacts of AI-based decision making on demographic subgroups.

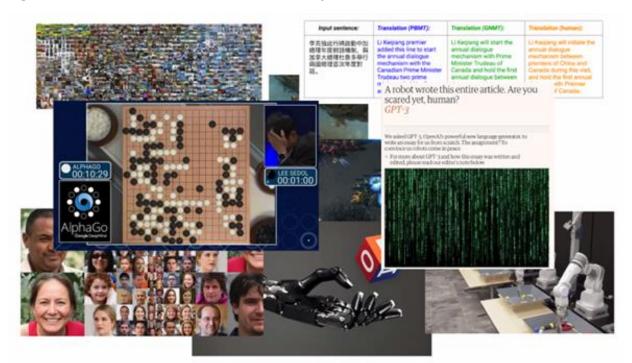
15. More broadly, technical, policy and regulatory tools that connect the frontier of knowledge for a given AI technology with its social implications are needed. This would facilitate the identification, understanding and mitigation of risks such as mis- and disinformation. She identified several metrics of successful AI: greater decentralisation, robustness, interpretability and transparency, and tracking unintended consequences, also by making use of AI.

16. **Aleksander Mądry**, Cadence Design Systems Professor at the Computing Electrical Engineering and Computer Science Department at MIT and a member of the Computer Science & Artificial Intelligence Laboratory (CSAIL), said that the previous decade – from 2010 to 2020 – had been characterised by major success stories in AI (Figure 2). He pointed out that these AI systems – exemplified by AlphaGo and GPT-3 – require significant investments. Beyond computing resources, he identified raw data as the key strategic resource for modern AI systems. This should lead policy makers to ask key questions such as who owns the data and why, as well as who can access and use it to train AI systems. Professor Madry said that in the next decade, AI should be re-assessed from the perspective of robustness: robustness to manipulation and adversarial attacks but also robustness of systems deployed in new contexts.

17. He emphasised the need to focus on assurance and auditing AI systems, from both a technical perspective of tools but also the need for public policy to help with implementing robustness at scale and setting up appropriate governance structures.

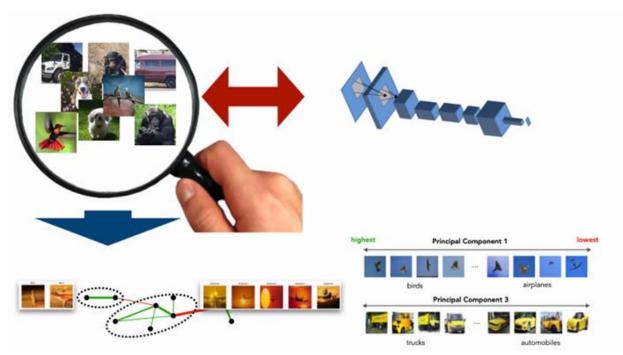
18. According to Professor Madry, two key technological advances are helping AI actors to increase the robustness/trustworthiness of their systems. The first is *model-driven data understanding*, which requires tools allowing the examination of the most important parameters from the training data in a structured way (Figure 3). The second is *model editing*, whereby partial adjustments can be made inside the model, without having to change the entire model architecture / retrain the model. For example, an image recognition model can mistakenly identify a police car as a huskie under snow conditions. Model editing would allow the user to edit this specific feature in the model directly without having to re-train it (Figure 4).

Figure 2. The 2010-2020 decade: Al's success story



Source: Presentation by MIT Prof. Aleksander Mądry





Source: Presentation by MIT Prof. Aleksander Mądry

Figure 4. Model editing as a key technological advance to increase trustworthiness of AI systems



police car

"husky"

Source: Presentation by MIT Prof. Aleksander Madry

Discussion

19. Markus Dicks, Head of the German AI Observatory at the German Federal Ministry of Labour and Social Affairs, asked about new AI design tools that could potentially improve human-machine interactions.

- Professor Ozdaglar replied that a holistic approach is needed to make progress on both the human and the technological side of human-machine interactions, and improve performance both on the training as well as generalisation. Such a holistic approach should incorporate behavioural models into technology development to generate performance gains when deployed. New technological developments can advance issues like robustness, model interpretability and privacy. Some tools can guarantee privacy for users.
- Professor Huttenlocher added that ongoing work at MIT on interactive design tools is very promising. For example, interactive design tools can give people a high-level understanding of the AI model's complex representations of the environment, facilitating human intervention to adjust a model's design if needed. He said the trend is towards moving away from supervised machine learning approaches to less supervised approaches that involve transformations between one domain and another. Researchers have started to characterise the nature of the representations that are learned automatically in that way to enable their understanding, for example, a node representing the colour of the car for care images.
- Professor Madry gave the example of the GPT3 language model that is also, for example, assisting programmers alongside Github. A programmer can ask GPT3 to write a program that conducts a certain task based on existing open-source code repositories, so does not need to start from scratch.

Markus Dicks also inquired about where leading MIT academics foresaw linear or rather, 20. exponential AI developments over the coming 10 years. He also asked whether participants foresaw an "AI Autumn", which technology(s) or region(s) are advancing AI most and who is "in the driver's seat" when it comes to AI developments today.

Professor Madry said that so far, most of the progress acceleration was driven by size, i.e. larger • models with more data. He foresaw that at some point, the scaling may stop but that large technology companies are the main drivers of AI development since they have both the data and the talent required to develop and improve AI applications. These companies are driving the AI market because they continue to develop new techniques, models and applications that are later adopted by others players in the AI ecosystem.

21. **Duncan Cass-Beggs**, Head of the Strategic Foresight Unit at the OECD, asked about what plausible developments and milestones towards artificial general intelligence (AGI) could be envisioned in the upcoming 10-15 years, and what the related policy implications could be.

- Professor Madry put forward the hypothesis that much of current deep learning is smart memorisation from a large corpus of data; which is different from the competencies AGI systems would require. He noted that metrics that could delineate between memorisation and reasoning would be useful in determining the extent to which AI systems are moving towards AGI-level capabilities. He cautioned that governments did not pay enough attention to AGI, although it garners significant attention from companies.
- Professor Huttenlocher stated that the most likely path is "less and less narrow AI", rather than
 fully general AI. Many of the issues that AGI will bring are present in narrow AI today, and
 governments and societies do well in addressing those issues today instead of waiting for AGI to
 arrive. He also mentioned concentration and the loss of humanity's heterogeneity, which would be
 a substantive risk that requires more attention.
- Professor Andreas added that the systems that are closest to general AI today are *i*) large language models that imitate human writing through synthetic text for a specific task; and *ii*) highly specialised AI models such as chess-playing engines or medical diagnosis systems that have either been trained on datasets that look very different from data humans learn from or that have algorithmic capabilities like the ability to search or compute that humans would not be able to perform. In the future, these superhuman algorithmic capabilities might be combined with the ability to exhibit human-like behaviour in a specific context. He thought that fundamental technological leap(s) or phase changes would be needed for AGI. The path towards more and more general AI systems would be a punctuated equilibrium with varying improvements every year (e.g. much bigger models or fundamental technological advances), rather than a linear or exponential curve. He likened the field to hydrogen physics where bigger accelerators are built regularly and sometimes but not necessarily discover new physics.

22. **Audrey Plonk**, Head of the Digital Economic Policy Division at OECD, asked about what other disciplines AI researchers draw lessons from and what disciplines policy makers could look at for inspiration.

- Professor Mądry replied that lessons from IT privacy and security could be helpful but that AI was
 very different. He put forward that the best analogy for a possible regulatory framework in AI is the
 healthcare industry, where guiding principles exist but most often treatments could vary depending
 on the patient and the context, and potential repercussions are based on the outcome and the
 amount of information that was available during the decision-making process. He added that it
 would also be important to draw from social sciences and economics to understand the human
 aspect and social underpinnings of AI applications.
- Professor Ozdaglar added that when thinking about AI robustness and composability issues, there are many linkages with engineering disciplines like composability in designing aeroplanes. When it comes to AI-driven decision-making, the human-computer interaction field could help enable users to use recommendations more effectively. Moreover, causal inference experience from other disciplines – such as statistics and economics – could help to provide a better understanding of the causal underpinnings of an AI system's predictions and recommendations. She underlined the risks of relying too widely on a similar model and of correlated failures if

developers and deployers do not have the authority of training different models addressing different uncertainties.

• **Professor Torralba** underlined the difficulty of regulating AI without thinking about the domain in which the AI system is deployed, and cautioned against regulating artificial intelligence as a generic technology. Professor Torralba stated that many existing regulations naturally extend to AI.

23. **Michael Schönstein**, Head of Strategic Foresight and Analysis at the German Federal Ministry of Labour and Social Affairs, asked about the most important things that final users of AI systems – such as doctors, policy makers and workers – should know to get a better understanding of how AI applications function and be better prepared to evaluate their outputs and recommendations.

- Professor Ozdaglar said that AI training for the workforce should be considered. Trainings should
 include an overview of the different tools available to workers to improve their use and
 understanding of the AI systems at hand, together with clear information on when and why these
 tools might not be appropriate.
- 24. Other items discussed include:
 - On predicting the growth of AI capabilities: "When it comes to the question of exponential or linear growth [of AI capabilities] in the coming years, currently, the approach is to make projects bigger with more budget and see what happens, which will eventually reach a limit. It is hard to project AI progress because this lends itself to the exponential growth [discourse], and it might be one improvement to one critical piece that suddenly makes the whole system significantly better." (Professor Alexander Madry)
 - On the risks of foundational models: "Some of the underlying pieces that build the algorithmic composition come from a few unique sources that can afford to develop these models. It is very dangerous that many AI systems use the same underlying pieces because they will inherit biases [present in these foundational models] that might keep hurting the same group of people." (*Professor Antonio Torralba*)
 - On the need for heterogeneous training and processing: "Homogeneity of models can also lead to correlated failures. Relying too much on similar models instead of training each model to be able to address different aspects of uncertainties can lead to a lot of deployment issues." (*Professor Asu Ozdaglar*)
 - On federated learning as a way to address the issue of homogeneity in large models: "Thinking about federated learning is especially important when it comes to big models. Heterogeneity in processing and training could address the issue of homogeneity in large models." (*Professor Asu Ozdaglar*)

8 |

Annex A. List of participants

The workshop was attended by the following participants (ordered alphabetically by last name):

Name	Title	Organisation
Jacob Andreas	X Consortium Assistant Professor	MIT
Luis Aranda	Artificial Intelligence Policy Analyst	OECD
Silvia Beyer	Advisor at Policy Lab "Digital, Work and Society"	German Federal Ministry of Labour and Social Affairs
Sarah Box	Senior Counsellor in the Directorate for Science, Technology and Innovation (STI)	OECD
Duncan Cass- Beggs	Counsellor for Strategic Foresight	OECD
Alessandra Collechia	Head of the Scientific and Technological Policy Division of the Directorate for Science, Technology and Innovation	OECD
Markus Dicks	Head of the German AI Observatory	German Federal Ministry of Labour and Social Affairs
Daniel Huttenlocher	Inaugural Dean of the MIT Schwarzman College of Computing	МІТ
Aleksander Mądry	Cadence Design Systems Professor of Computing	МІТ
Asu Ozdaglar	Faculty member in the Electrical Engineering and Computer Science Department	МІТ
Karine Perset	Head of the AI Unit of the OECD Division for Digital Economy Policy	OECD
Dirk Pilat	Deputy Director of the OECD Directorate for Science, Technology and Innovation	OECD
Audrey Plonk	Head of the Digital Economy Policy Division of the Directorate for Science, Technology and Innovation	OECD
Taylor Reynolds	Technology Policy Director of MIT's Internet Policy Research Initiative	МІТ
Michael Schönstein	Head of Strategic Foresight and Analysis	German Federal Ministry of Labour and Social Affairs
Antonio Torralba	Delta Electronics Professor of Electrical Engineering and Computer Science	МІТ
Andrew Wyckoff	Director of the Directorate for Science, Technology and Innovation (STI)	OECD