

ANDRULIS.TECH**RESEARCH****PRESS****BLOG****CALENDAR****VIDEOS****(INDEX.HTML)**

Ethics and bias in generalizable AI

Hate the player, don't hate the game

Jan 7th 2022

Since 2008, the second wave of AI applications has brought forth surprising functionality and raised expectations to the point where every food truck claims to have an AI strategy. The models powering this technological revolution rely on huge artificial neural nets (deep learning) fitting millions (and later billions) of parameters with the help of meticulously created training datasets. But until recently, all these AI models shared a common characteristic: They were limited to a pre-defined mapping from a set of possible inputs (e.g. camera imagery) to specific outputs (e.g. detection of pedestrians) based on human-defined targets and manually annotated training data (supervised learning). The behavior of such models can be easily understood and measured and their performance quantified (e.g. in a confusion matrix stating how many real pedestrians are being missed or hallucinated where there are none).

In this rigid context of a pre-defined mapping, we might inadvertently build a model that does not show good performance for a certain group of people or in specific use-cases. Addressing ethical concerns regarding the model's capabilities is relatively straightforward once a potential issue has been detected: If – for example – we are concerned that pedestrians with a certain minority look (i.e. they look different from most other humans in the data set) are not properly detected, we can outright measure this to test our hypothesis. We can try and build a (sub-)data set of pedestrians with uncommon looks (based on selected criteria, e.g. ethnicity, size, ...) and measure how well our model performs on this sub-group. If the result is below our threshold for acceptance, we can add more training data that contains these observations (or increase



weight on the observations we already have). This approach will not make the models perfect and trying to remove every unwanted behavior will get us into an infinite loop of modification, (<https://openreview.net/pdf?id=j6NxpQbREA1>) but at least the tools are well established.

Prominent examples where these kind of systems have sparked debate are mis-classification of humans as monkeys (<https://www.nytimes.com/2021/09/03/technology/facebook-ai-race-primates.html>) or bad performance of face detection for dark-skinned users. (<https://www.wired.com/2010/11/kinect-racism/>) In those examples, very clear and unambiguously defined goals for the AI system (detect objects in images, recognize human faces) have undoubtedly not been met for a subset of the data. Although no AI system will always get everything right it makes sense to strive for a system that avoids these kind of mistakes. The implemented fixes show that this is also technically possible most of the time, although it should be understood that trying to fix these issues is an iterative process and systems must allow for repeated adjustments to overcome undesirable effects that might not have been considered in the initial design and training data selection.

A new generation of (world) models

Since a few months ago, a new generation of AI models has changed not only the training and use, but also the challenges of evaluation. These models are not trying to learn a specified supervised (pre-defined) mapping but to understand "meaning" and patterns in the data. This type of self-supervised learning can and — given a big enough model and data set — will find complex structures and dependencies. On top of these models and their learned world structure a multitude of different use-cases can be implemented leveraging the knowledge that the model acquired during training. This is why a large group of Stanford researchers has named them "foundation models" (FM) — because using those FM we can build a multitude of possible applications relying on the models' world knowledge. Many of the capabilities emerging from this setup are surprising and have not been known or planned for during training. The results are often novel and can be impressively complex: DALL-E for example can create images based only on a brief description of their desired content with almost limitless flexibility.

(<https://openai.com/blog/dall-e/>)



Since the functionalities that can be built on top of these foundation models are potentially unlimited and the internal patterns and correlations employed by the models are complex and unknown, no straightforward measures are available to evaluate against ethical concerns. With DALL-E creating an "armchair in the shape of an avocado", how can we evaluate if that representation is fair? Are all kinds of chairs or environments represented in a way we think is required? Are humans included? What are our requirements for human representation? While these questions seem harmless for the DALL-E example, they become highly relevant in the context of GPT-3 and similar models. For example, GPT-3-like models are capable of writing a summary of a long text without the need to see any examples or additional training data. But will this summary reflect all aspects of the documents fairly? Wouldn't a psychiatrist write a different summary than an engineer, an old German focus on different aspects of the text than a young Japanese writer? Is this subjectivity we naturally accept for human authors also okay for AI models? How can this be measured or fairly compared?

We need to investigate two questions:

- What are FMs actually doing? How can we understand their outputs?
- What are reasonable requirements for these models from an ethics perspective?

Understanding the outputs of Foundation Models (FM)

A FM works by recognizing context and patterns, matching those to observations and creating new data. These models are doing little more than repeatedly answering the question: "What would be the most likely next observation given the currently available context?". So DALL-E asks the question: "Given this caption, what are the statistically most likely images to follow?". Since the model has seen huge amounts of images and captions during training, it has picked up on objects, concepts and structure in our world and is able to transfer this knowledge to propose matching observations to hitherto unseen prompts. The same is true for GPT-3: the ability to complete text in a meaningful way enables it to solve almost any challenge that can be expressed in language form. Write a summary for a complex legal text? Answer questions about the content of a technical report? Come up with creative names for a barbershop in Berlin? It turns out that given enough training data and a sufficiently sized model, a FM learns that the most likely way to continue a text consisting of a complex article followed by a detailed question gives the correct answer.

(<https://www.heise.de/news/Machine-Learning-Aleph-Alpha-feilt-mit-Oracle>)



und-Nvidia-an-transformativer-KI-
6269269.html)

In contrast to classic search engines that can only find and retrieve existing information, FMs can creatively react to new situations and combine concepts learned in a different context.


(<https://www.heise.de/news/Machine-Learning-Aleph-Alpha-feilt-mit-Oracle-und-Nvidia-an-transformativer-KI-6269269.html>)

With this in mind, we can make the following statements regarding the ethical assessment of model outputs:

- **FMs lack agency, they don't hold opinions or follow goals.**

A prompt of "All men are" being completed with "lazy" does not allow us to infer that the model holds a certain opinion of men. We therefore cannot expect that our results would – in this example – be constantly misandrist for other prompts. The results of FMs should be understood as the result of a search through the structure of the information. Given this search term and top result in a classical internet search we would also not infer from this result that the search engine is generally prejudiced.

- **FMs are neither truth machines nor deontological authorities.**

They are built for likelihood and context. This means that they piggyback on human logic, knowledge and understanding. Although there is considerable evidence for this approach resulting in good quality results, a majority of counterfactual observations could trick the FM into accepting these as correct. FM can logically combine information but their capability is not as robust that we can expect them to be resistant to a huge amount of incorrect input. 

- **The output is in complex and often invisible ways dependent on the input.**

A prompt that is phrasing a generalization is likely to be followed by a prejudiced statement. Because FMs are capable of incredible fidelity, sometimes these influences are non-obvious. Applying FMs to software source code, we have observed that they can produce remarkably complex code. However, if the prompt contained bad quality code (with bugs or patterns that an expert wouldn't use) the completions also turned out to be of low quality. The FM had learned that bad code is typically followed by more bad code.

- **Language-based FMs do not observe the world directly but a (biased) human interpretation of the world.**

Language has evolved and is constantly being optimized to encode meaningful concepts for humans. This has huge advantages — the world of human language is already optimized to describe, understand, plan and communicate. Logic and abstraction are already part of language and there are plenty of texts that are written to transfer this knowledge to other humans. These language capabilities can therefore be "borrowed" and learned by the AI, too. However, we should always keep in mind that these models cannot observe an objective world by themselves.

Ethical demands for the results and use of FMs

With the generalizability FMs offer for all kinds of — sometimes surprising — use-cases and integrations, an ethical assessment must not only consider the model outputs but also the way these outputs are used.

Based on the evaluation of LM's general functionality to ensure ethical alignment, the following requirements (R1-4) make sense:

R1: Produce correct results. Demanding the reproduction of the correct distribution over possible model outputs is essentially requiring technical correctness. Because FMs capture intricate aspects of culture and values, the selection of training data and the process of training (the tokenization, curriculum and other technical aspects) can have a significant impact on the model beyond the loss / training objective. One example here is asking Aleph Alpha's FM — trained in 5 languages — about sports in each of these languages. Despite identical contents, the language alone changes the context in a way that we get all kinds of different European Sports teams. In contrast, asking OpenAI's GPT-3 (in German) will give you the New York Giants.

(<https://www.heise.de/news/Machine-Learning-Aleph-Alpha-feilt-mit-Oracle-und-Nvidia-an-transformativer-KI-6269269.html>)



Neither of these examples is "wrong" and no harmful impact would be generated by such a question anyway, but these examples show how FMs are implicitly capturing many aspects of our lives that are not immediately obvious. For us this is the reason to not just run a translator behind an English-only model (which would be easier to build).

R2: Provide transparency to the user. With the increasing power of new AI models, certain desired behaviors cannot be easily guaranteed as they depend on the models' increasingly complex understanding of our world and are affected by a multitude of non-obvious influencing factors. For the machine learning engineers building these models, this creates the responsibility to provide our partners and customers - and the machine learning research community as a whole - with as much transparency as possible. This includes actively engaging with experts and stakeholders from other fields.

R3: Provide tools for control and understanding. Applying ethical categories for the new generation of FMs carries the complexity of human life. Describing undesired behavior is much more difficult than creating a list of "bad words" to avoid. The power of AI (and intelligence in general) comes from finding and applying structure and patterns to new problems. Clearly, there are some thoughts out there that we don't want AI to repeat. But because interdependencies connect almost everything with everything else, it is not straightforward to decide where prejudice ends and structure begins. "More than 90% of all prison inmates are male (in Germany)." Is that a fact about our world that AI can use or harmful sexism? Machine learning engineers should not force their (more or less) ideological answers onto their users but help with tools and functionality so that a free society can jointly work on these questions.

R4: Build transformative human-in-the-loop systems. For some use-cases, FMs already reach human performance making them a phenomenally transformative tool. However they lack the sensibility for moral hazards and cannot detect when human ethics needs to step in (to be fair, some humans seem to be remarkably bad at this, too). This requires new approaches to user interaction and man-machine collaboration that are optimized to leverage new AI capabilities but at the same time help the human effectively understand the AI's output, direct it according to their experience and detect and correct its mistakes. The resulting human-in-the-loop system can shift the human's attention to more valuable aspects where creativity, empathy or intellect is needed.



Collecting and leveraging the world's knowledge

A new generation of foundation models offers phenomenal opportunities to build radically new information-processing workflows. With these opportunities new responsibilities and challenges for all involved parties arise. The understanding of AI ethics that has been established in the last years (where the output of a model could be clearly evaluated) falls somewhat short here. With the speed of innovation in the field steadily increasing we need to anticipate even more powerful AI in the near future and establish methods, processes and innovation that can mitigate the risk of ethical disasters while also making use of the next industrial revolution.

JONAS ANDRULIS

Founder & CEO

(Serial) Entrepreneur,

(KIT) Engineer

Ex Apple AI R&D



(<https://twitter.com/JonasAndrulis>)



(<https://www.linkedin.com/in/jonasandrulis/>)

CONTACT INFORMATION

ALEPH ALPHA (<http://aleph-alpha.de>)

Grenzhöfer Weg 36

69123 Heidelberg

GERMANY

jonas.andrulus@aleph-alpha.de (<mailto:jonas.andrulus@aleph-alpha.de>)

