MINDFUL AI: 5 CONCEPTS FOR MINDFUL ARTIFICIAL INTELLIGENCE

There is a path towards Mindful AI. The leap to Mindful thinking has happened fast, relative to other actions in tech companies (social media, net-zero carbon emissions), making it seem mindfulness itself is the standard by which an emerging technology is judged.

What is mindful AI?

Mindfulness is an attempt to see the whole view, not just the parts. At its narrowest view, AI can perform a task like detecting…

- An emotion from a face
- An object on a street
- The time it takes to reach your destination
- The next three words in a sentence

At its narrowest view, we consider only the observable actions—the what. The cost to create this narrow AI is cheap. DIY YouTube videos can get anyone familiar with coding up and running in less than an hour.

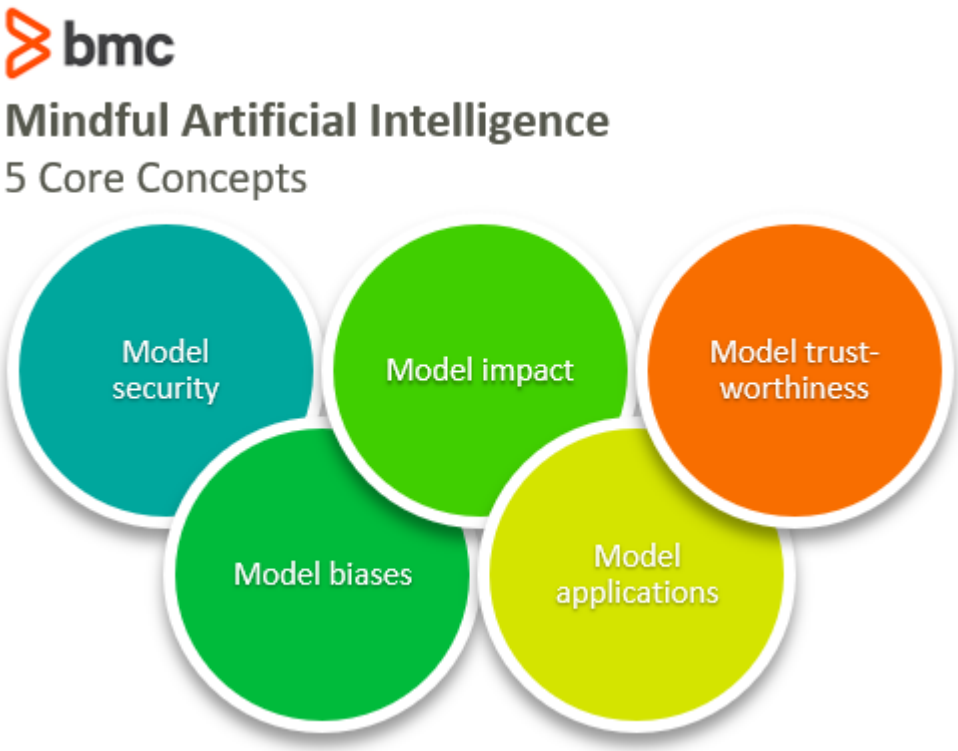
To mindfully create an Artificial Intelligence system means broadening this narrow view. There is no limit to how far a person can go. Mindful AI, however, is not the easiest step. Mindful AI requires extra consideration, time, resources, and understanding to create.

To be mindful:

- Preemptively state the risks to the model.
- Test for bias before deploying a model.
- Respond promptly to biases you detect in the model's predictions.
- Secure your model so user data is not retrieved.
- Take responsibility for your model's predictions.

Concepts for mindful AI

To get started with mindfulness in your AI practice, here are some items to consider when building AI systems:



Model security

Threats to AI models occur in traditional ways, ways that security systems have been defending against for a while. They are attacks against servers, where user data or models are being stored, or attacks against system bugs or system logic.

A good article from World Wide Technology says:

> *"Where AI model security becomes interesting, though, is in the discovery and development of new cyber attacks derived from the nature of the mathematics of AI itself—attacks that allow an adversary to fool the model, skew the model by carefully poisoning the input data, or use carefully crafted queries to steal sensitive personal data used to train the model and sometimes even the model parameters."*

The use of machine learning models trained on user data poses concerns: How secure is that user data?

In an extreme case, from bad ML models, models may memorize all data from its training set. When it does, an attacker can inference the model and see all the data on which it trained. Machine learning models themselves can store user data in more subtle ways, and attackers have figured out

ways to retrieve data from those aspects.

Protecting user data is a hard problem to solve. An active open-source community tackling the privacy problem is OpenMined.

Going these extra lengths to secure a model is not immediate nor obvious—so it's on the mindfulness list. A person in their backyard having a fire is usually no big deal, but starting a Yule Log outside a football stadium needs more thoughtful consideration.

Model biases

It is known from the outside that models can be biased in all kinds of ways, from racial preferences to color preferences to days of the week.

From the inside, these aren't necessarily models being racist, or a reflection of the person who created it themselves (the modelers generally are not guilty). Instead, these are natural consequences of modelling and working with data.

The outputs of a model are real and they must be considered. A model's appearance and its actions as racist are a real appearance, so the creators of models should notice the effect their models have on the people they serve and take their results into consideration.

It's like a painter thinking the subject of their painting is a person smiling genuinely and the audience says, "No, that is a wicked smile." Likely, the painter had no ill-intention. Similarly, the rise of authors hiring sensitivity readers to review their work before publishing so they don't get berated for the use of the term "she" or "Indian". These insensitive offenses are not necessarily the views of the creator, but when looked at from particular vantage points, it is possible people could consider their creation in such a way.

Like art, ML and AI models are served, observed, and critiqued by many audiences. They are statistical inferences that open linear programming to statistical inference. There exists no real control over what set of inputs will create a given output. The combination of inputs sometimes is so vast that testing them all is infeasible. Some views will be overlooked, but it is a good idea to put models through particular sets of tests in an attempt to both detect and prevent any model bias.

> *"You should take the approach that you're wrong. Your goal is to be less wrong." -Elon Musk*

Bias detection is imperfect and will undoubtedly encounter errors. When a company is found to have biases in their model, it is good of the company to correct the bias. Google did well to correct its model from the results shown in its model.

Using good data sources is one method to create better models. Mechanical Turk is slowly becoming the butt of a recurring joke about data labelling. If you want bad models that return bad results, use Mechanical Turk as a data labeler.

Model impacts on populations

If boys are always told, "Try it. Experiment. See what happens", and girls are always told, "That's risky. That's dangerous. Have you considered you might get hurt?", the two populations can take very different actions in their lives.

Models that serve inferences to people can have real effects on populations. Cambridge Analytica famously crafted messages in an attempt to sway a population.

Artificial intelligence must define a set of parameters with some reward of good and bad behaviors—a binary classifier. Creating an application inherently leads to discussion about penalty/reward systems for a user's behaviors (particularly when cryptocurrency enthusiasts get involved). In effect, many people delve unwittingly into Game Theory.

While I firmly believe people can know things well and perform ethically without formal education (even better sometimes), there are some good lessons from the discipline that can be considered when creating good AI models. Building a successful, mindful AI model doesn't need to be bound by the principles of pure mathematics.

But it can be worthwhile to look into the structure and history of game theory to get an idea of what tests have been conducted to see what has and has not worked. (I suggest looking at paradigmatic behavioral experiments and Game Theory Game Types.)

Models can serve entirely different experiences to different populations and will have different effects on people who use them differently. Netflix subscribers have a different Netflix experience based on where you are in the world. Accounts located in the U.S. will show different content than accounts in the U.K.

While dishing different Netflix content region to region is harmless, other content models may not be harmless. This has gone wrong on Facebook's algorithm, which we can identify as the echo chamber effect.

Model applications

Models have a dual-use problem: they can be used for good things and for bad things.

> *"As the dual-use nature of AI and ML becomes apparent, we highlight the need to reimagine norms and institutions around the openness of research, starting with pre-publication risk assessment in technical areas of special concern, central access licensing models, sharing regimes that favor safety and security, and other lessons from other dual-use technologies." –AI experts in the Malicious AI Report*

To avoid malicious use, and any associated guilt, you can guide your use cases for your AI models in a few ways:

- Assessing risk pre-publication
- Rolling out the models slowly
- Actively watch the apps the model creates

Thus far, OpenAI set an example when it released a risk assessment at the same time it released its GPT-2 language model. The media, like spectators commenting on something they don't understand, blew the risk out of proportion.

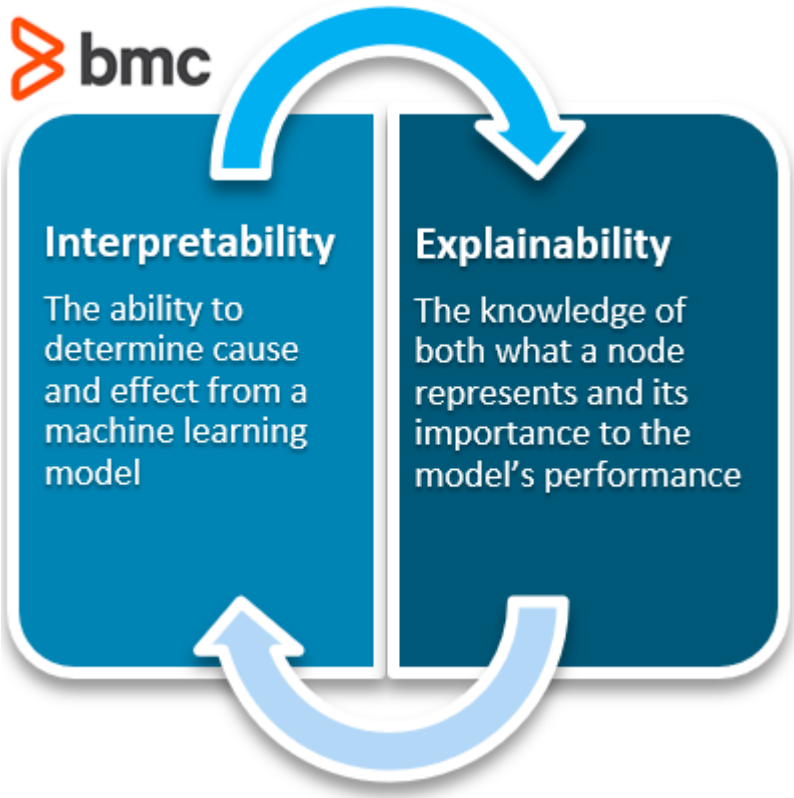
As OpenAI rolled out its model for general use, it monitored the kinds of applications the model was being used for, giving access to only a smaller, less precise model than the whole thing at once. As it tested the waters, which were mostly undisturbed, OpenAI progressively released larger models, ultimately arriving at its GPT-3 language model, dwarfing the GPT-2 model. They keep an eye out for new applications using their GPT-3 model.

Model trustworthiness

Finally, there is a model's trustworthiness. A secure model, an unbiased model, and a model that plays fair across populations can all help a model be perceived as trustworthy. In addition, increasing a model's explainability and interpretability can help with a model's trustworthiness.

When a model does make a bad prediction, who is to blame?

To be trustworthy, we generally want to be able to hold someone accountable for their actions. People can get offended, and if their plea to hold someone accountable falls flat on a faceless offender because an offender simply doesn't exist (black box models) or because no party steps up to claim responsibility, we cannot build trust.



Often, trust builds in the event of error. When something goes wrong, you can learn to trust the party who responds with its actions. A company's trust is earned; it is not pre-ordained.

One way to create trusting behaviors is through explainable and interpretable models. Both are something extra to strive for—they alone are not a single measure to which we can hold a company accountable.

But explainability and interpretability can help a modeler understand why the model predicted what it did. When things go awry, these characteristics can help a modeler go back and adjust the model. Implementing explainability and interpretability into a model can have the same application to AI models as adding monitoring has to chaos engineering. The end result is to be able to take action in the event of failure.

Extra but needed steps to mindful AI

At the end of the day, mindfully building an AI takes extra steps—which can result in an ecosystem where people trust the use of AI in our technological systems. Success for a modeler will happen

slowly as the exciting draw that got them into machine learning in the first place expands beyond the initial narrow view and widens to include more design features.

Related reading

- BMC Machine Learning & Big Data Blog
- Top Machine Learning Frameworks To Use
- Machine Learning: Hype vs Reality
- What Is Machine Learning Operations? MLOps Explained
- Apache Spark Guide
- Guide to ML with TensorFlow & Keras