Beyond a black-box approach to artificial intelligence policy – a simple guide to definitions, functions and technology types

Richard A Roehrl (DESA)[1]

A working definition

There are no universally agreed definitions for *artificial intelligence*, nor for *digital automation*. In particular, defining *intelligence* is extremely difficult, due in part because of the remaining scientific unknowns regarding human intelligence. On the other hand, it is the kind of thing that "*we know when we see it*". For example, it is notoriously difficult to measure human intelligence, yet it is pretty obvious to us when we meet a highly intelligent person.

As a working definition, I suggest following eminent cognitive scientist Margaret Boden: *"Artificial Intelligence (AI) seeks to make computers do the sorts of things that minds can do. Some of these (e.g., reasoning, reading, understanding speech) are normally described as 'intelligent'. Others (e.g., vision, hearing, moving around natural obstacles) aren't. But all involve psychological skills – such as perception, association, prediction, planning, motor control – that enable humans and animals to attain their goals."*[i]

In that sense, AI is broader than what we typically consider human intelligence. It includes a wide range of information-processing capacities. In fact, AI could include aspects of intelligence that are far outside the reach of humans. And while AI requires physical machines, especially computers, it is really *informationally powerful virtual machines* (e.g., software) that make the machine intelligent. Hence, AI comprises a very wide range of approaches, techniques and technologies.

To understand AI's current and future impacts on humanity's global aspirations - such as the SDGs - the functional definition above which specifies psychological skills is useful. For example, this approach is used by a recent study published by Vinuesa et al. in Nature which reports on a consensus-based expert survey and found that AI might enable the accomplishment of 134 SDG targets, but also inhibit 59 targets. As AI they consider "*any software technology with at least one of the following capabilities: perception—including audio, visual, textual, and tactile (e.g., face recognition), decision-making (e.g., medical diagnosis systems), prediction (e.g., weather forecast), automatic knowledge extraction and pattern recognition from data (e.g., discovery of fake news circles in social media), interactive communication (e.g., social robots or chat bots), and logical reasoning (e.g., theory development from premises).*" [ii] I propose adopting this list as a working basis to discuss SDG impacts. However, I am cognizant of the fact that such a "black-box approach" to AI which does not specify the underlying techniques and technologies is not sufficient to fully understand all SDG impacts. The details of AI concepts, techniques and technologies matter greatly, especially with regard to sustainability and with regard to future directions of AI and its physical, economic and socio-political limitations.

In fact, AI's impacts depend greatly on the type of AI algorithm and also on its physical implementation with a myriad of interdependent information and communication technologies (ICT) and infrastructures. For example, deep neural networks (DNN) with supervised learning – which has become by far the most widely adopted AI technology in the 2010s – requires enormous amounts of data and large amount of energy for data handling and computation. Hence, much talk about the "data economy" these days. In terms of physical implementation, it is important to note that AI used on a mobile phone via a 5G network differs greatly from AI run on a supercomputer in a research lab. This is also a reason why it is difficult to differentiate AI impacts from those of the underlying, necessary infrastructure and related technologies. In particular, DNN requires "big data" applications, data centers and cloud computing. Similarly, there is a continuum of process or system automation tasks – in factories run by robots to online digital automation of tasks - some of which are considered intelligent because they adapt to changed situations, whereas others cannot adapt. For these reasons, we include a discussion of *digital automation* and related infrastructures in the report here, even though to the extent possible we will aim to draw a distinction with AI proper.

[1] Note: The views expressed in this brief are those of the author and do not necessarily reflect those of the United Nations or its senior management.

Table 1. Basic AI functions

<table>
<tr><th></th><th colspan="5">Basic AI functions</th></tr>
<tr><td rowspan="2">What it can do</td><td rowspan="2">It can “see” and identify what it sees.</td><td rowspan="2">It can “hear”
and transcribe what you say</td><td colspan="2">It can “read”
what you type</td><td rowspan="2">It can move by itself, based on what it sees and hears, and does not follow a programmed path.</td><td rowspan="2">It can “reason” and looks for patterns in massive amounts of data, and uses them to make decisions</td></tr>
<tr><td colspan="2">and responds in a useful and sensible way.
text passages
and analyzes for patterns</td></tr>
<tr><td>AI technique</td><td>Computer vision & image processing</td><td>Speech recognition</td><td colspan="2">Natural language processing (NLP)</td><td>Smart robot</td><td>Machine learning</td></tr>
</table>

Source: adapted from a flowchart published in Hao (2018b).[iii]

Types of AI

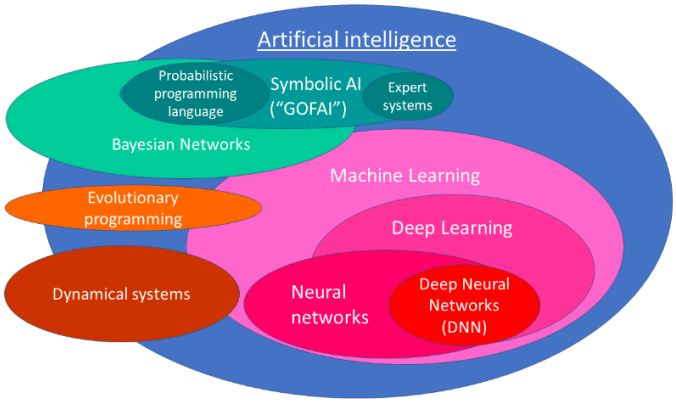
Researchers have long distinguished “*strong AI”* from “*weak AI*”. Weak or narrow AI was developed for a specific, well-defined task, and it is the dominant and the only existing type of AI today. One example is Apple’s Siri personal assistant. On the other hand, strong or general AI aims to be a system with general cognitive capabilities, in order to solve any new task, similar to our human, general purpose brains. General AI remains under development, and it is highly uncertain when it will be achieved. It could be a matter of months, years, or decades. For obvious reasons, it would quickly lead to superintelligence and an intelligence explosion – with a whole host of issues. Some leading AI researchers, (e.g.., Stuart Russel) and philosophers (e.g., Nick Bostrom) have highlighted the issues and proposed solutions requiring urgent action. Superintelligence is an extremely important issue, but our report will focus primarily on narrow AI, in view of its practical use and certain immediate impacts on the world’s SDG aspirations.

Arend Hintze further distinguishes four types of AI. Reactive machines, such as IBM’s DeepBlue chess programme or Google’s AlphaGo, are purely reactive. They do not form memories nor base current decisions on past experience. Limited memory machines look into the past and form transient memories. This is, for example, used in autonomous vehicles which record temporary timelines of other cars’ speed and direction. This information is transient, in contrast to human drivers who accumulate driving experience over many years. The following two types do not yet exist. Theory of mind machines form representations of other people, agents and institutions. They understand that that they can have thoughts and emotions that affect their own behavior. Theory of mind is what allowed us forming strong social systems and societies. And finally, self-awareness machines can form representations about themselves, which will require building machines with consciousness.[iv]

AI in the broader sense simply means a machine that is intelligent in the sense that it can solve a specific problem. Based on the types of algorithms and ways in which information is being processed in virtual machines, five major AI types are distinguished: (a) Good Old-Fashioned AI (GOFAI), also known as classical or symbolic AI; (b) artificial neural networks or connectionism; (c) evolutionary programming; (d) cellular automata; and (e) dynamical systems. While there have long been research groups that focus on one or the other of these conceptual approaches, many modern virtual machines are hybrids (Figure *1*).

Neural networks are great for modelling brains, for pattern recognition and learning. GOFAI - nowadays often combined with statistical methods - is great for learning, planning and reasoning tasks. The other three methods are popular in biology and the field of artificial life (A-life). In particular, evolutionary programming is used to analyze biological evolution and brain development, cellular automata and dynamical systems model living organisms – both natural and artificial. Various software and hardware implementations exist in these five areas. And SDG impacts depend on all these specificities. So, how do these and other terms relate to each other? In Figure *1* we made an attempt to provide an overview.

Figure 1. From artificial intelligence to deep neural networks



Source: authors. *Note:* The size of the ellipse do not imply application importance and range of techniques.

One important subcategory of AI is *machine learning* (ML). It is a particular AI that learns by itself – the more data it can draw on the better it gets. Machine learning is therefore defined as "*…algorithms [that] use statistics to find patterns in massive amounts of data.*"[2] *Data in this context means anything that can be stored digitally, e.g., numbers, words, images, clicks, etc.* Today it is being used in search engines, recommendation systems (e.g., youtube, spotify), social media feeds (e.g., facebook, twitter), and voice assistants (e.g., Siri, Alexa).

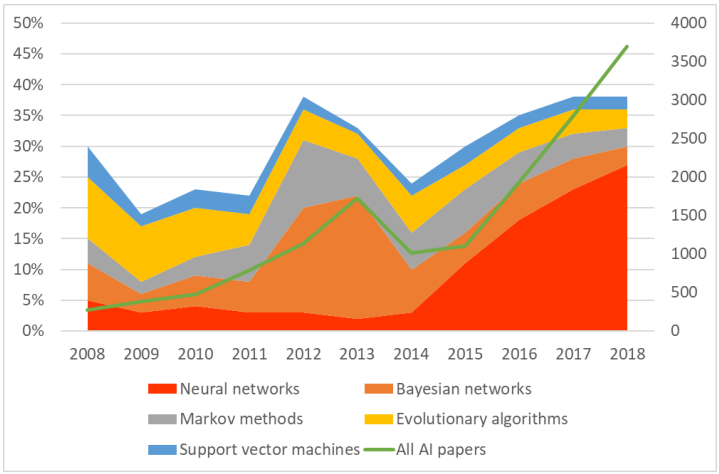
An *expert system* is a type of classical AI that does not learn by itself and thus is not part of machine learning. Instead it typically operates on "*if this, then do that*"-statements. It is for example used in medical diagnosis and treatment. However, symbolic AI (GOFAI) is often combined with other types of AI. For example, when combined with Bayesian Networks, it leads to probabilistic programming language which is used in powerful AI applications (e.g., used by CTBTO to detect nuclear weapons tests).

Deep learning is a machine learning technique that "*…gives machines an enhanced ability to find—and amplify—even the smallest patterns.*"[2] The most successful version of it is a *deep neural network (DNN).* It is deep because *"…it has many, many layers of simple computational nodes that work together to munch through data and deliver a final result in the form of the prediction*."[2] Neural networks were inspired by the structure and functioning of human and animal brains. However, as we mentioned above, DNN is very data and energy hungry. To overcome this limiting factor, applications have emerged combining with other methods, especially symbolic AI, and some expect new types of hybrids to emerge as the new paradigms for the 2020s.

Some of the other popular techniques employed within and outside of machine learning are Bayesian networks, support vector machines, and evolutionary algorithms, all of which take different approaches to finding patterns in data. For example, evolutionary programming adapts Darwinian principles to automated problem solving. The structure of the program to be optimized is fixed, while its numerical parameters are allowed to evolve. They fail in the sense of learning from experience and thus are not necessarily considered by all experts as part of machine learning.

Machine (and deep) learning comes in three flavors: *supervised, unsupervised*, and *reinforcement learning*. Today, supervised learning is the most prevalent methods in which data are labeled to tell the machine exactly what patterns it should look for (e.g., when you click on an Amazon product). In unsupervised learning, the data has no labels and the machine just looks for whatever patterns it can find (e.g., used in cybersecurity). Reinforcement learning tries out lots of different things and is rewarded or penalized depending (similar to dog training) on whether its behaviors help or hinder it from reaching its clear objective (e.g., Google's AlphaGo). In addition, hybrid learning methods have emerged, including semi-supervised, self-supervised, and multi-instance learning.

Figure 2. Share of AI techniques in arxiv research papers over time



Source: Hao (2019).**Error! Bookmark not defined.**

A leading indicator for AI types to be used in applications in the coming 5 to 10 years are the number of recently published academic AI papers in any of these

[2] Hao, K. (2018). What is machine learning? MIT Technology Review, https://www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart/

categories. For the case of influential cross-journal platform arxiv,

Figure *2* shows the relevant publication activity from 2008 to 2018. It shows that the number of AI papers has rapidly increased, especially since 2014; that all kinds of methods are being explored in parallel and hybrid forms; and the most popular research area is neural networks. This implies that at least for the next years, we can expect lots of DNN applications which will be combined with other methods as need be. This will have important implications on our SDG aspirations until 2030, as will be detailed in this report.

Conclusion

Artificial Intelligence (AI) seeks to make computers do the sorts of things that minds can do, and various functional or descriptive definitions exist. It is important to note that AI has a long history, dating back to the 19th century, with many techniques dating to the 1950s. In practice, various clusters of AI types are used – also in hybrid formats.

In the 2020s, computing and data handling capacities reached a critical level that made "deep neural networks" possible that can now surpass human cognitive capabilities in narrow, specific tasks, such as facial recognition, medical radiological diagnosis, and many others. Narrow AI has become ubiquitous in many countries – unbeknownst to many. At the same time, billions remain excluded from its benefits. Performance and capabilities grow at exponential rates, leading to new applications, new development models, and also sustainability concerns. This has important implications for humanity's aspirations expressed in the SDGs.

Finally, it is impossible to adequately understand the full implications of AI without exploring the specificities of the AI technology clusters concerned. It is hoped that this primer provides an easily accessible guide to these specificities. Indeed, a black-box approach to "AI" is only of limited usefulness to most policy questions.

[i] Boden, M. (2018). Artificial intelligence. Oxford University Press, ISBN 978-0-19-960291.
[ii] Vinuesa, R., Azizpour, H., Leite, I. et al. The role of artificial intelligence in achieving the Sustainable Development Goals. Nat Commun 11, 233 (2020). https://doi.org/10.1038/s41467-019-14108-y
[iii] Hao, K. (2018b). What is AI? We drew you a flowchart to work it out, MIT Technology Review, 2018, https://www.technologyreview.com/2018/11/10/139137/is-this-ai-we-drew-you-a-flowchart-to-work-it-out/
[iv] Hintze, A. (2016). Understanding the four types of AI, from reactive robots to self-aware beings. https://theconversation.com/understanding-the-four-types-of-ai-from-reactive-robots-to-self-aware-beings-67616