SET Brazil

SMPTE

SINCE 1916

SMPTE Motion Imaging Journal

Covering emerging technologies in film, broadcast, and the connected media ecosystem

O SMPTE, como sempre, é muito competente em trazer temas relevantes. Neste artigo apresenta e explora os fundamentos da inteligência artificial (IA), incluindo o aprendizado de máquina (*Machine Learning*), aprendizado profundo e inteligência geral artificial. Procura fornecer uma visão geral das tecnologias e conceitos envolvidos, buscando demonstrar que existe uma exploração dos princípios e diferenciação entre máquina e aprendizado profundo.
O artigo, também, discute o impacto da tecnologia de IA na sociedade e os papéis emergentes atuais e potenciais da Inteligencia Artificial no espaço de mídia e entretenimento. Aproveitando-se aqui para brincar um pouco com o filme "O Exterminador do Futuro" (2003), então "*Hasta la vista, baby*!".
Boa leitura a todos!

por Tom Jones Moreira

Defining Artificial Intelligence

by Richard Welsh

Abstract

This paper explores the fundamentals of artificial intelligence (AI), including machine learning, deep learning, and artificial general intelligence. It provides an overview of the technologies and concepts involved. There is an exploration of the principles of and differentiation between machine and deep learning. This paper also discusses the impact of AI technology on society and the current and potential emerging roles of AI in the media and entertainment space. Concepts such as general versus narrow intelligence, optimization space, and how simple narrow functions of AI differ from complex functions such as biological life and theoretical general AI are covered.

The notion of AI, in general, has raised much concern in some sectors of the scientific and technological community as well as political concerns about the impact on employment and social cohesion.

Keywords

Artificial intelligence, computer intelligence, deep learning, general artificial intelligence, human intelligence, machine learning)

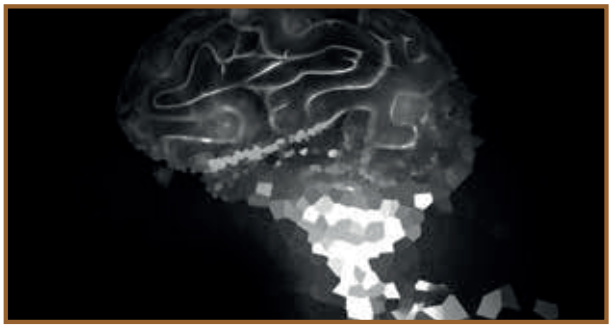
Introduction

It helps to have clear distinctions of what is meant by the term artificial intelligence (AI). The phrase was first used by John McCarthy in the 1950s in a proposal for a seminar to study thesubject at Dartmouth College.[1] The idea behind making this distinction was that human intelligence was "real" and computer intelligence was artificial. For the purposes of this introduction, I want to make a distinction between three concepts of AI, namely, machine learning (ML), deep learning (DL), and general intelligence. All of these fall under the umbrella term of AI, and the basic principle is that these are synthetic or mathematically defined processes that mimic the biological decision-making process.

ML underpins most of what we currently think of as AI. ML will be described in much more detail later, but, for now, think of it as a network of decision-making nodes that has been trained to perform a specific task. This might be a simple face-recognition algorithm or voicedetection tool. ML systems require a training framework that is typically orchestrated, so the ML will be trained to a point of usefulness, then deployed in that state. To improve it, more curated training will be required.

DL works by the same fundamentals as ML, but the networks are layered and, typically, a DL tool's training is continuous. It usually happens on a much larger scale than that of ML. DL is often trained by massive "live" data sets generated by the users of the system being trained. DL systems are able to train themselves thanks to their layering, and, because they are exposed to very large training data sets, they are able to use statistical analysis of their own results versus an observed data set to determine if they are becoming more or less accurate. This is done without the need for a human to curate that data set, typically because it is based on humangenerated data such as social media where thousands or millions of humans are already doing the curation. This is common in applications such as search engines, tailored online advertising, or image-recognition systems.

Digital Object Identifier 10.5594/JMI.2018.2880366
Date of publication: 29 January 2019



How do we define the line between ML and true AGI?

Artificial general intelligence (AGI) is effectively the "end-game" AI, which is a free-thinking intelligence that can solve a wide range of tasks and be able to specialize in any one of those to improve performance, much as a human would. A general AI may exhibit "consciousness," but this is not a requirement of general AI. It is simply an emergent property, and one that is not sufficiently defined or understood at this point. AGIs are still the stuff of science fiction, and, while there is no practical reason why it cannot be achieved, it is generally accepted that, for now, it is something that will not happen until sometime in the very distant future.

The notion of AI, in general, has raised much concern in some sectors of the scientific and technological community as well as political concerns about the impact on employment and social cohesion. There are varying levels of belief that we need to establish rules and controls on the development of AI applications now to protect ourselves from harmful outcomes. The use of ML networks trained on data that is derived from human behaviors is naturally going to reflect the biases of the training data, and this is a valid ethical concern if the results of those AIs are to reinforce behaviors that are harmful to individuals or society as a whole. It is clear that, as AI touches more and more of our everyday lives, the impact of AI across society must be taken seriously.

Given the political interest, it is very likely that, for good or ill, regulation will follow. However, when considering AGI as a looming existential threat, it is worth noting that, although much thought was given to the notion of computers "thinking," the progress toward AGI is considered by many to have been little or none. One of the generally accepted indicators of AGI, the "Turing Test," has proven to be a controversial subject.

Many groups have claimed to have passed the test (whereby a computer is indistinguishable from a human when responding to questions from a suspicious judge); however, other experts claim that we are not even close to passing the test. Over the years since the test was devised by English mathematician Alan Turing,[2] many attempts have been made; and, given the progress in computing power in the same period, it is arguable that we simply do not understand the foundational problem[3] as opposed to not having the required resource.

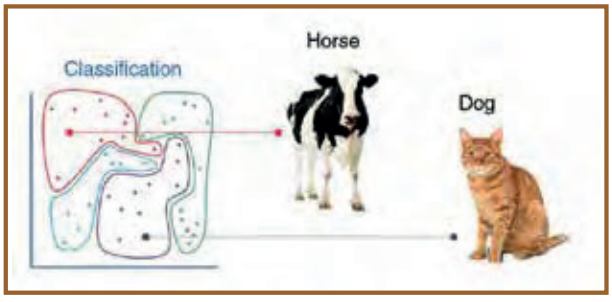


Figure 1. A high-level view of the CBC/Radio-Canada ecosystem.

ML Versus DL

ML can simply be classified into two broad types of networks: those being regression oriented and those being classification oriented. Classification networks are often those we associate with DL as they perform best when presented with very large data sets and can improve with continuous feeding of new information.

As an example, a classification network might be trained on the difference between elephants, horses, dogs, and mice by feeding it images of such and telling the network which is which. Over time, it will decide where these different categories lie, bound them, and then make a decision when presented with new data as to which it fits best. However, it will only be able to make an identification based on its classes and data sets. So, once presented with an object outside of these, it will try to match them within the classes it has. The quality of training data and the determination of classes are critical to the usefulness of classification networks. Furthermore, all ML algorithm successfulness pivots around the quality of the training data.

It is clear from **Fig. 1** that more data is better in terms of differentiating the classes; but, furthermore, it is in the interest of optimization to feed as much information as possible (including new classes) into the network. In the case of an application such as object recognition and tagging on a photo-sharing website, the platform owners would very likely wish to allow for significant optimization and differentiation of cat images. Luckily, in this particular scenario, there is an overwhelming abundance of training data.

Classification networks will typically require significant computation. The best of these for general purposes, such as search engines, necessarily have to live inside the massive data centers that are the preserve of tech giants such as Google, Amazon, Microsoft, etc., and their use relies on the connectivity of the user to the utility.

Regression networks are used to solve arbitrary problems using a set of inputs and weighted node-

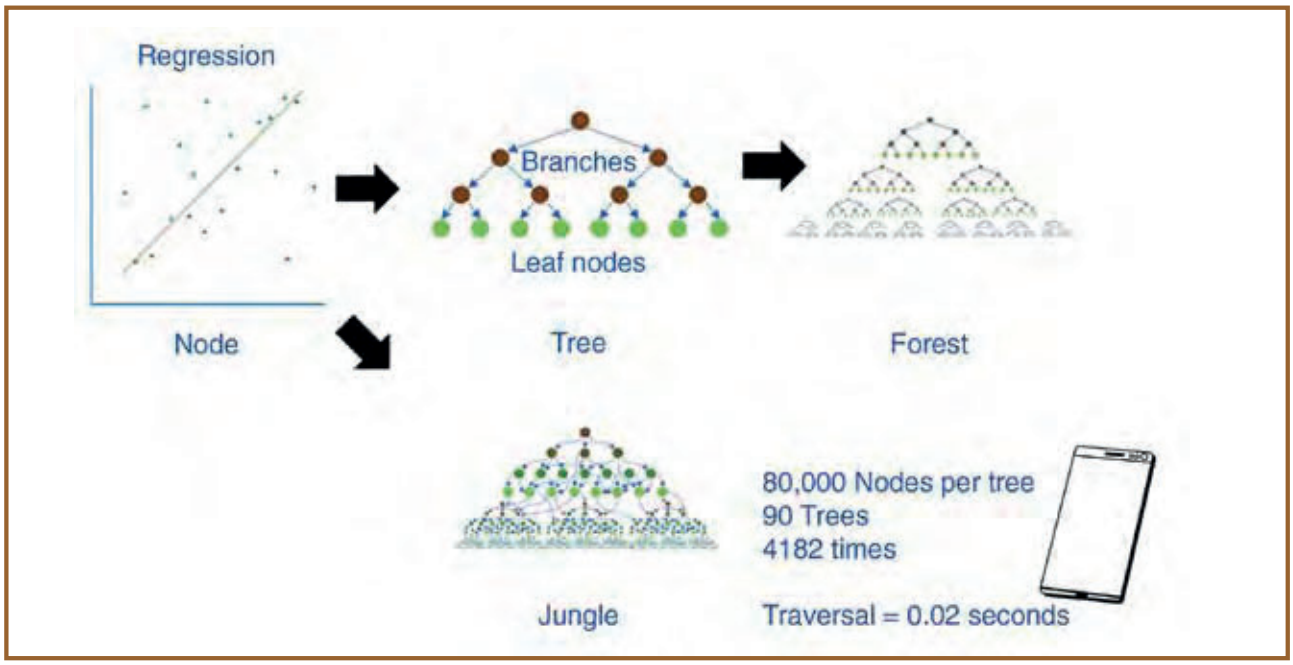


Figure 2. Regression networks are formed of branching nodes organized into trees then simple forests or more complex jungles.

based networks that result in a single-output value. Typically, such regression networks are used en masse to produce useful results, each network tackling a smaller subcomponent of the problem at hand. As shown in **Fig. 2**, the network consists of a series of nodes network consists of a series of nodes, which typically make a binary decision regarding which node to step to next and passes on a value to that next node to test. Multiple nodes form decision trees, and, in most practical cases, these trees will be combined in what is known as a forest. The network may be designed such that a problem results in a new weighted version of the original value, which is passed back into the same network in an attempt to reinforce the result and hone in on a final useful decision.

In a more complex arrangement of regression networks, known as a jungle, branches from trees can merge and thus allow multiple paths to individual nodes and decisions can be based on non-linear functions.

The advantage in all cases is that each nodal decision typically requires very little computational effort, and so this type of ML can easily be deployed directly in local software or even embedded in device hardware.

A good example is simple face recognition in mobile devices that can be run locally on the chip that controls the camera, thus bringing the computation as close to point of use as possible. This can be used to determine the generally preferred point of focus and even detect the best moment to take the photo within a window after the trigger (i.e., when no one is blinking).

However, generic face recognition is not really enough for the modern consumer expectation, and such local ML cannot necessarily determine who or what is in the image. For this, it needs a DL network and a suitable database of faces with reference to personal information, most likely linked to social media, whereby it can now identify and tag the people and objects in the image with a high level of confidence (**Fig. 3**).

Practical Usage — Multimethod

As we have discussed, ML and DL are typically built around solving narrow problems. So, using them will typically be multimethod. The nature of the ML/DL in use will determine whether it can be local or remote. In many cases, use of ML in both consumer and professional applications is typically a hybrid of the two, with lightweight problem solving that occurs locally using ML, and heavier-weight problems being tackled in a cloudborne DL service that returns the results to the user seamlessly. This is well demonstrated in the widely used sphere of voice control. In the case of Amazon's Alexa, the local process recognizes the "wake" word (which, incidentally, you can customize if you build your own Alexa using open source) and then the rest of the command is streamed to Amazon's Alexa cloud service where their DL natural language processing kicks in and interprets the command. The natural language component of this engine is so well advanced that developers can simply build their applications around the written language and the back-end service does the rest. It works in most languages and is becoming capable with picking up even fine contextual distinctions between similar words and the intent implied in the question. This is a good example of an extremely lightweight local process that fronts an incredibly powerful back-end engine resulting

Figure 3. Face recognition in smartphones may use a combination of ML on the device and DL in the cloud.

in such a seamless AI experience that many users report considering their voice assistants to have personality —thus anthropomorphising the service as shown in **Fig. 4**.

General Versus Narrow Intelligence

Most of the biological lives we observe exist because they have been successful at surviving in their environment and, in most cases, have done so through special skills. In simplistic terms, each organism, whether plant, animal, or microbe, has found its niche and adapted over time to be optimal (at least by comparison to the competition). As far back as the 1700s, Thomas Bayse,[4] a statistician, philosopher, and minister who studied at Edinburgh University, had made what were at the time controversial proposals that human decision-making was driven by a series of probabilistic calculations based on an experiential feedback loop, which is confidence based on strength of beliefs and theorems that reinforce themselves based on observation. This would lead to more complex emergent behavior; but, in principle, it was no different than the underlying decision-making of an animal or insect brain. This simplistic feedback loop underpins the principles of ML today.

ML works on the basis of training, testing, selection of the best performer, then training, testing, etc. One builds (or usually picks off the shelf) basic networks (**Fig. 5**), gives them inputs, tweaks their behavior to give different outputs, chooses the one that is closest to the answer one is looking for, then iterates. These ML networks can be adversarial, i.e., they directly compete with one network creating synthesized data and another network discriminating between the real and generated training data. While the discriminator is winning (i.e., it can tell which is real and which is not) it feeds back to the generator such that the generator can change and improve. Similarly, when the discriminator loses, a feedback loop improves the discriminator. When the discriminator can no longer tell the difference between real and synthesized results to a significant percentage, the generator is practically deployable.

At face value, this compete-to-win process is very much like nature—"survival of the fittest" at work. However, we are attempting to build networks that are solving a narrow (usually single) problem. Over time and training, much like an organism going through generational adaptation, the ML network becomes more and more optimized for its task.

While it is tempting to try and compare AI to biological intelligence, it is important to understand this in the context of what is known as "optimization space." Optimization space is an imaginary and infinite space, where any process that could possibly take place occupies an area in that space. Processes that are similar or share components will overlap in the space but the narrower the process, the less likely it is to overlap with others.

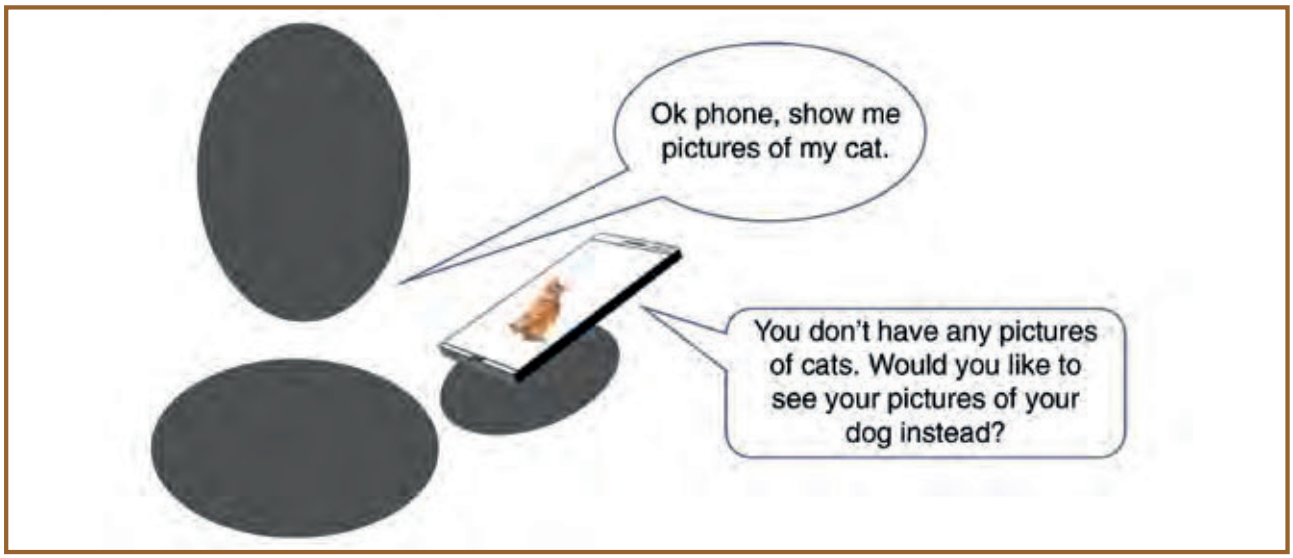


Figure 4. Voice assistants rely on the heavy lifting of a backend DL natural language processing service, combined with other functional applications to provide a seamless voice control experience to users.

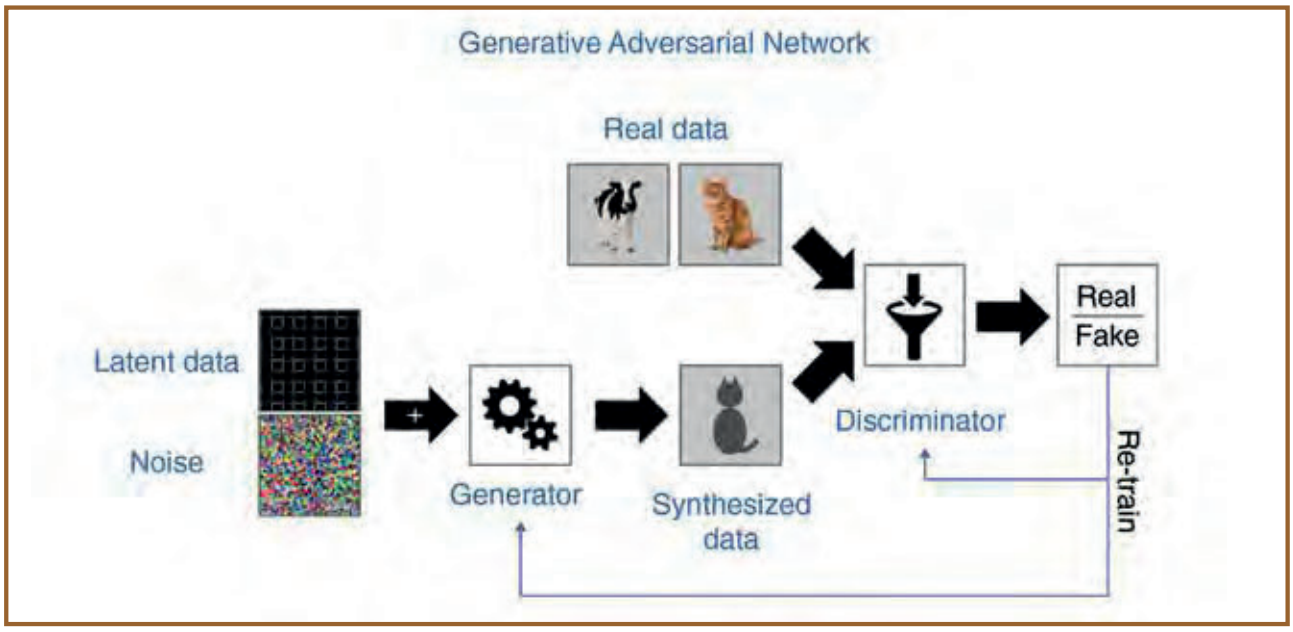


Figure 5. **Generative adversarial networks compete to improve.**

Complex multifaceted functions with similar component processes will have large overlaps and can typically be grouped, whereas simple and isolated (e.g., single purpose) processes are disconnected and usually very distant from all other processes in this imaginary space. An arbitrarily assigned space example is shown in **Fig. 6**.

If we consider a few simple organisms to fit broadly within one space, which we will call the biological optimization space, they will have many goals that overlap (movement, reproduction, growth, respiration, nutrition, etc.) and many that do not. For instance, a bird that needs to migrate for survival or a human who needs to migrate for a new acting gig in Los Angeles are both moving to warmer climates but for very different reasons.

Now, imagine a super AI (not AGI, but pretty awesome at tackling abstract problems nonetheless) that is given the task of moving itself from Nova Scotia to California as efficiently as possible. It might try to figure out ways to transport itself on hardware from A to B using air freight or rail, but most likely (assuming it lives in a massive data center) it might simply take over the available bandwidth from A to B in order to shift its bits to another suitably large data center. At face value, it is pretty obvious. In the process, however, the lost bandwidth caused emergency service communications to fail. It is possible that the target data center had storage or computation capacity overtaken that had critical infrastructure relying upon it. Because this particular AI's optimization is simple and abstract, it will not consider wider impacts of its actions.

Here is where AI diverges from anything approaching "biological intelligence." The context and understanding of the space we occupy drives us to consider it in every move we make. Other organisms may not "consider" those factors, but, as a collective ecosystem, Earth has had hundreds of millions of years of optimization opportunity for these billions and trillions of organisms to learn to work in symbiosis. Any AI we manufacture does not have any real link to that environment, and, in theory, it could over time exist in independence from it and us. The optimization space it occupies will never really need to be shared with ours, and thus its actions will not necessarily be compatible with our requirements in the context of our optimization space.

This model raises many questions about AI. In particular, do new arbitrary processes move toward the space we occupy, or do they move away from our goals? Moving into our optimization space means they become necessary components, i.e., we depend on them for life. Moving away means the opposite; they are not critical to us, and most arbitrary processes right now would be said to occupy the regions outside our optimization space. Not being a critical component inside our space does not preclude them from posing an existential threat, but it further detaches them from our understanding and influence. The big question is, where does an AGI fit into this space?

A simple example is the recent success of the Alpha GO from the Google Deep Mind team in London. They developed an AI capable of beating the worlds' best human players at the board game "Go." Go is considered a far more difficult problem for computers to solve than chess because it is mathematically more complex and thus precludes "brute force" AI techniques such as regression tree traversal to solve for multiple outcomes and select the best one in a reasonable amount of time. Alpha Go uses DL to accelerate its "abilities" and find the best

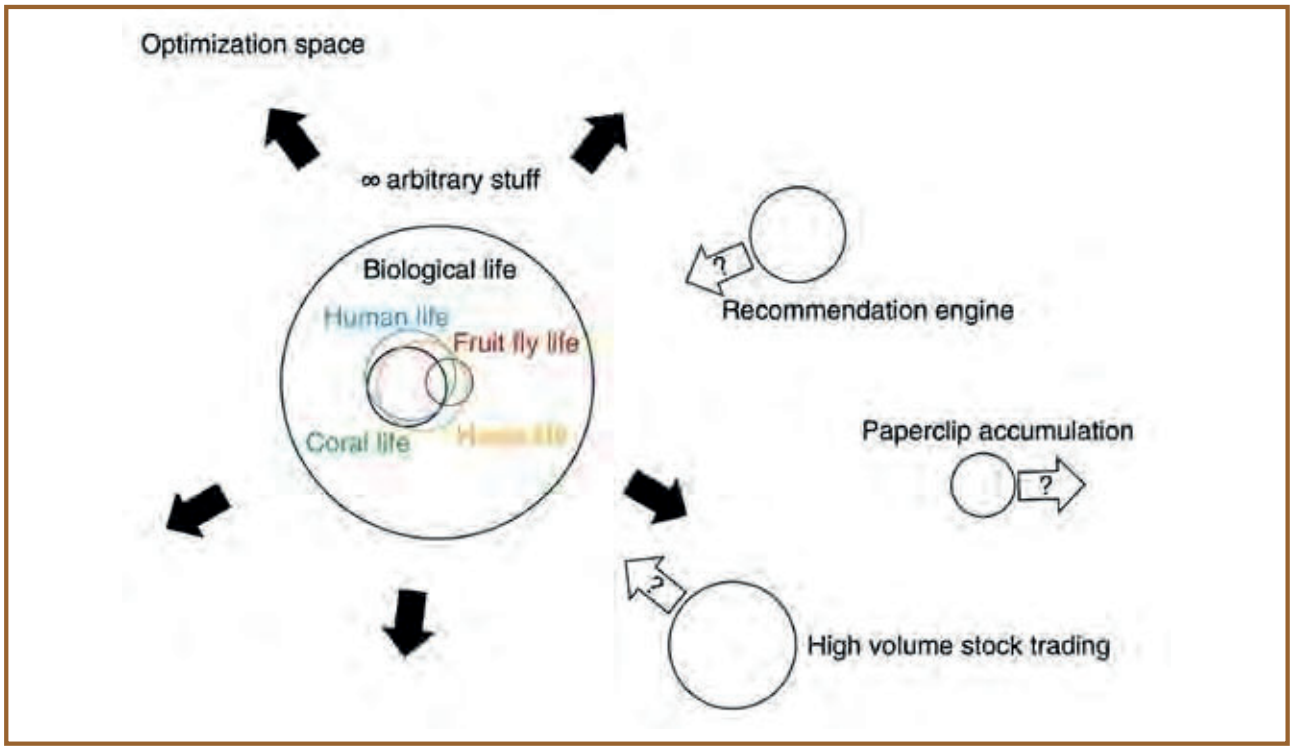


Figure 6. **Optimization space is an infinite imaginary field in which every process has an area. Complex multiprocess functions such as biological life will occupy a very different space to an arbitrary simple function, such as collecting paperclips.**

moves to win. It even started to invent strategies never seen before, thus surpassing its human counterparts. However, Go is a board game that most humans have never played and never will. It does not occupy any part of our optimization space. The Alpha Go program is clearly more optimized than the human equivalent, but ask it to make you a cup of tea and it will struggle. That optimization lives somewhere else, completely occupied by simple machines with no AI involved at all (except maybe voice controlled kettles).

Therefore, we can build seemingly incredibly powerful AIs, but they are for very narrow purposes and highly detached from each other. It is possible that AGI is emergent from the eventual coalescing of these separate optimizations, but it seems that AIs to start systematically tackling arbitrary problems would need to be built first to find all the optimizations. This is a popular view among AI theorists, that over time AGI is emergent from simpler AIs with some general capabilities but not individually powerful enough to be considered AGI themselves.

Marvin, HAL, and the Terminator

AI in fiction is dramatically characterized as anything from annoying to apocalyptic. While the science fiction notions of how an AGI might decide to behave typically result in some sort of negative outcome (they are rarely portrayed as the hero or heroine), the outcomes of much simpler (nongeneral) AIs could pose just as much risk to humans and life in general. We are already surrounded by everyday objects and tools that are quite capable of harming or killing people by accident or by design and do so with great frequency at the hands of humans. Applying AI instead to those things (e.g., autonomous vehicles) clearly requires a significant amount of research and practical regulation to prevent unnecessary accidents. The case of driverless cars is particularly prescient because it is on the immediate horizon, and, as proponents point out, humans are not particularly good at that job of driving, with vehicle-caused injuries and deaths currently numbering in the thousands per day. In principle, properly trained and capable machines should be better at this task than people and dramatically reduce the number of road deaths and injuries. This does not necessarily reduce the number to zero, and it will be little consolation to those who still fall victim that the statistics had improved. Despite this, it ought to be clear that the application is benign and that overall benefits are worth the inevitable outlying incidents with negative outcomes. The question of responsibility and "choice" here is particularly difficult since an algorithm is at the heart of "decisions."

The challenge lies in the understanding of how the AI will behave, and, given the astronomically large number of scenarios any AI may have to deal with, we cannot know in advance the outcome for all of them.

We can model a specific scenario and see precisely how the AI network performed for that exact task, but

Figure 7. Identification, tracking, and masking of faces using AI in a commercial application.

a subtle change to a single variable may have a vastly different outcome. If that outcome is not limited, then it could go well outside the bounds of acceptability for any particular task. So the question of human trust, control, and common sense comes into play.

For the media industry, the trust level in AIs will define its use in the near term. The final decision-making process will have to allow for some level of intervention and, in most cases in the creative process, will probably require it. A current example is the task of censorship editing. AIs are already in use that identify (and could potentially remove or obscure) nudity for the purposes of compliance and censor cuts.[i] This dramatically reduces the time required to modify these versions. It remains unlikely, however, that the task will be wholly turned over to the AI to complete and these versions will ship "unwatched."

Another example is automatic rotoscoping of prerelease content,[ii] where the idea is to obscure all but the characters' faces so that dubbing voice artists can see the performance of the character they are voicing but not have a useful copy of the content for piracy. An example of before and after such a process is shown in **Fig. 7**.

This is traditionally a manual process because it requires decisions to be made about what does and does not get shown. Therefore, a simple AI can do this as long as it is only looking for faces, but what if it is looking for a talking car or robotic dog?

So now the AIs need to have more contextual understanding, but that can then extend to showing the whole body of a character for a particular shot because their physicality is informing the artist of how they should voice that line of dialog. The trust level to just let an AI do it unchecked is not there yet. However, this is a fairly mundane and time-consuming process for the operator, so there is impetus to off load this to an automated process. Furthermore, the impact of the AI getting it wrong is very low; so it is likely that an application like this could be trusted to be left to the AI.

This final human control is an important point and will likely be a major driver in the practical use of AI in the media industry for the foreseeable future. It is very unlikely that we will trust AI 100% to perform any creative task without the ability to influence or stop it altogether. However, as applications are identified that run low risk, and as commercial pressures on time and cost come to bear, it is increasingly the case that we will start to turn some or all of certain processes over to AI. AI's part in the media industry is only just beginning, and, as this edition of the SMPTE Journal will explore, AIs potential for the media industry is huge.

[i] GrayMeta Curio platform uses multiple AI services to provide compliance detection including nudity, sex, violence, gore, and profanity.

[ii] Sundog Media Toolkit DubSafe tool uses multimethod ML tools to pick out faces during sections of dialog and track them through shots to create dynamic masking of the content for security.

References

1. P. J. Hayes and L. Morgenstern, "On John McCarthy's 80th Birthday, in Honor of His Contributions," *AI Magazine. Association for the Advancement of Artificial Intelligence*, 2007.
2. A. Turing, "Computing Machinery and Intelligence," *Mind*, LIX (238):433–460, Oct. 1950.
3. C. Smith, B. McGuire, T. Huang, and G. Yang, *History of Computing*, CSEP 590A, University of Washington, 2006.
4. D. Barber, *Bayesian Reasoning and Machine Learning*, Cambridge University Press, U.K., pp. 8–9, 2012. [Online]. Available: http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/270212.pdf

AUTHOR

Richard Welsh is co-founder and CEO of Sundog Media Toolkit Ltd, U.K. He has served on the SMPTE board, most recently as education vice president. He has worked in the cinema industry since 1999 and has been involved in various technology development work during this time, primarily in digital cinema, and is named on patents in the area of visual perception of 3D. He founded the cloud software company Sundog in 2013, which specializes in scalable post-production software tools aimed at high-end broadcast and movie productions. Prior to Sundog, he worked at Dolby Laboratories, U.K., where he held various positions including film sound consultant, mastering engineer, and director of digital cinema. Subsequently, he was the head of Digital Cinema Operations at Technicolor, U.K. He holds a BSc (Hons.) in media technology and an honorary Doctor of Technology degree from Southampton Solent University, Southampton, U.K.