2019 IEEE International Symposium on Technology in Society (ISTAS) Proceedings
Miriam Cunningham and Paul Cunningham (Eds)
ISBN: 978-1-7281-5480-0

Ethics and Artificial General Intelligence

Technological Prediction as a Groundwork for Guidelines

Charles J. Simon
Future AI, Washington, DC, USA, Email: charles@futureAI.guru

***Abstract* - Artificial General Intelligence (AGI) is the possible future of computer systems which are as capable as humans across a broad range of intellectual requirements. In order to establish an ethical position or guidelines for the development of AGI, it is important to explore anticipated characteristics about the emergence of AGI: How sudden it could be (jolt), how soon it could be (timing), and how dangerous it could be (risk). By extrapolating today's trends in development and limitations of current AI algorithms, informed speculation can help set ethical positions and guidelines on the proper course. This paper concludes that the emergence of AGI will be gradual, soon, and only moderately dangerous and begins to address how ethical issues will change as AGI emerges from narrow AI.**

Keywords - artificial intelligence; artificial general intelligence; deep learning; machine learning; neural networks; robotics; semantic network; ethics

I. Introduction

This paper first introduces the divergence of professional opinions about the emergence of AGI and then provides background about today's AI algorithms. It then outlines some of the limitations of narrow AI techniques in terms of their ability to grow into AGI by considering some of the major conceptual problems which are not solved by today's AI systems. It points out that while it is unlikely that any individual algorithm will expand into full intelligence, a combination of today's algorithms (a trend already in progress) can address several of these missing abilities.

Combining current AI approaches can lead to more intelligent systems within the coming decade but will likely lead to the discovery of further problems which need to be solved and a continuing conversation of whether or not AGI has actually been achieved. This reasoning leads to the conclusion that AGI emergence will be gradual but will arguably begin on the earlier end of the spectrum of opinion. The level of danger we can expect from AGIs is also predicted from expectations of the algorithmic approaches. These likelihoods are then compared with conceptually competing technologies of brain augmentation and brain-content uploading.

Finally, the paper describes how ethical concerns about future AGIs may differ from today's with three example areas: training sets, black box learning systems, and the potential for the appearance of consciousness in future AGIs.

The intent of this paper is to organize and present issues by extrapolating today's trends in development, so that informed speculation can help set ethical positions and guidelines on the proper course.

II. The Axes of Opinion

There is no professional consensus about the emergence of AGI. In order to organize the divergence of opinions, one might build a graph of expectations with three axes as follows:

- **Jolt**: Consider rating 1-10 with 1 being the emergence of AGI over several decades and 10 being a single technical breakthrough which leads to an emergence of AGI in a single step (a "singularity").
- **Timing**: Consider rating 1-10 with 1 being immediately and 10 being at least 80 years (or never).
- **Danger**: Consider 1-10 with 1 being no danger and 10 representing likely elimination of humankind. Contributing to the danger are not only the risk of military-style attack, but the risks of job losses, economic upheaval, and an overall societal change.

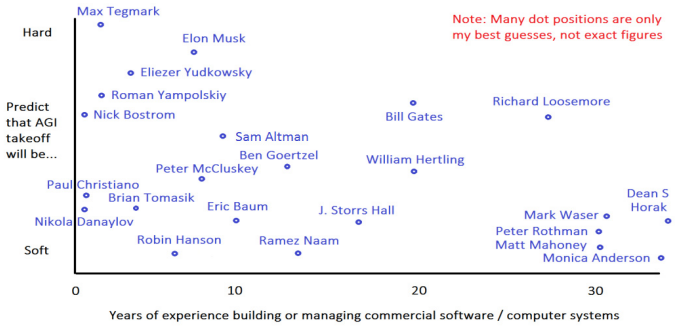

Fig 1. Informal chart showing little consensus on Jolt [1].

An informal survey on the jolt of AGI illustrates the spectrum of viewpoints as shown in Fig. 1. Not a scientific survey, this chart is intended only to illustrate the diversity of professional opinion. This paper argues (below) for a value in the 1-2 range.

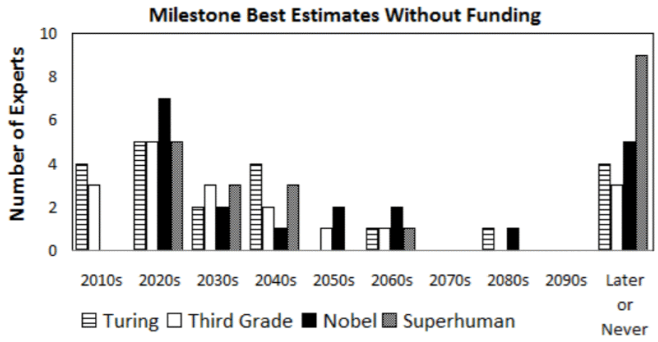

Fig 2. Informal survey showing little consensus on AGI timing [2].

Fig. 2, similarly, shows the significant divergence of opinion on the timing of the emergence of AGI. Any apparent clustering of the data in these figures could be attributed to the informal sampling process. This paper argues (below) for a value in the 2-3 range (a few decades off).

While the potential danger of AGI is getting considerable publicity, much of this writing is uninformed and based on fictional representations of AGI. A survey of AGI projects shows that many of them do not consider risk factors into their research. This paper argues for a danger value of 5—that AGI is not inherently dangerous but is subject to weaponization and abuse in the wrong hands [3].

III. A Brief Taxonomy of AI

The term "Artificial Intelligence" was coined in 1956 by pioneer John McCarthy to encompass many areas of human intellectual endeavor. Here are three general categories:

A. Connectionism

Also called "neural networks" and now "deep learning", this strategy is based on the idea that connecting vast arrays of simple processors, mimicking the internal structure of the brain, will lead to intelligence.

B. Algorithmic/Symbolic AI

This area is based on the idea that writing programs which mimic what the brain outwardly appears able to do, like playing chess, will lead to intelligence. Some authors break this area into subgroups [4].

C. Robotics

Robotics is not usually considered part of the AI field because a large portion of robotics is consumed with mechanical and direct control systems. However, the idea of an autonomous robot has always presumed a significant content of AI software (both algorithmic and connectionist [5]) and today's autonomous vehicles are essentially large autonomous wheeled robots. To the extent robotics uses AI, the AI can be categorized as connectionist or algorithmic.

While there have been huge advances and well-documented successes in all areas, it is safe to say that, historically, they have all underestimated the difficulty of the problems they addressed and overpredicted the levels of success they could achieve. Whether that trend continues remains to be seen. Looking at the limitations of existing algorithms has significant bearing on the anticipation of AGI.

IV. Some Limitations of Deep Learning

"Deep Learning" is really a misnomer as the word "learning" implies a level of understanding which does not exist in these systems. A more appropriate moniker might be "Deep Correlation" [6]. A deep learning network analyzes its input set and finds correlations between patterns in the input set and desired outputs. Based on this analyzed correlation, when subsequently given a novel input pattern, it can propose a likely output. As examples, certain arrangements of pixels may correlate highly with written characters. After training, a system can recognize written characters with excellent accuracy even when written characters vary from the specific training patterns. To such a network, the characters have no meaning, they are just patterns of pixels. In a more general sense, to the extent that a network can be trained to correlate correct answers to given questions, it may appear to be intelligent.

In many applications, deep correlation is extremely useful. The limitations of today's deep learning are mostly rooted in its ability to produce an answer without understanding the underlying concepts of the question. As examples, in a neural network which was "trained" to recognize images of dogs vs. wolves, it was discovered that any image with a large white area would be tagged as a wolf because a significant number of images of wolves in the training set also contained snow. Any neural network can be badly trained because it has no understanding of the underlying concepts like dog or wolf (or snow) [7]. Further, it has no knowledge of the underlying concept of physical things existing in a reality.

At DeepMind, a neural net was "trained" on older Atari video games [8]. The goal given to the network was to produce a maximum score given the input of the continuous pixel image of the video display. The system used a trial-and-error approach to controlling the game called "reinforcement learning" where the computer repeated behaviors which resulted in higher scores and discarded ones which did not. After achieving very high scores, one *might* say that the program learned to play the game well. In reality it found correlations of certain pixel arrangements on the screen with certain control actions which led to high scores. The network has no concept of "game", "winning", "score", or any of the game's rules. We (humans) look at the game and see a speedboat on a racecourse. The network just gets inputs of an array of pixel values.

As computers get more powerful, neural networks can ascertain progressively more precise correlations. More training sets and more time to converge on solutions yield better results and these machines may appear to be more intelligent. But no amount of additional computer horsepower will overcome the underlying problem that the neural network has no mechanism for understanding basic concepts underlying its decisions.

Some AI professionals believe that once deep learning is applied to future computers with the immense computing power equivalent to a human brain, that human-level

intelligence would spontaneously emerge [9]. Realistically, while more powerful computers will definitely yield faster and more accurate correlations, there is no reason to think that any "understanding" will result from current algorithms any more than one would expect that a weather simulation program run on an immensely more powerful computer would do any more than produce more accurate weather predictions.

V. Some Limitations of Symbolic AI

While neural networks generally apply a single class of algorithm to a broad spectrum of problems, symbolic AI brings more specific algorithms to bear on specific problems. For example, an early chess-playing algorithm searched trees of possible moves or a natural language processing (NLP) program analyzed syntax probabilistically to generate responses to questions. These two algorithms do not merge easily because they are fundamentally different. Within its narrow field of ability, either approach can create the appearance of understanding when, in fact, none exists [10].

In considering symbolic approaches vs. neural networks, one finds that in the event an algorithmic method of solving a problem is known, it is much more efficient than a comparable neural network, especially when considering the vast training sets which are used on deep learning systems. This has led to the implementation of many special-purpose algorithms used to solve specific problems which cannot subsequently be applied more generally.

Merging neural network and symbolic algorithms is beginning to enter into the mainstream [11] [12]. For example, an NLP system can be coupled to a neural network which analyzes geometric 3-D images to answer questions about the images [13].

VI. Adding Some Fundamental Missing components

There are many ongoing AGI projects and many more proposals for software architectures, which could make computers, perform like humans ([3] and [14] contain lists of current AGI projects). Many are very specific with block diagrams containing hundreds of functional boxes. Instead, here are a few general concepts, which the human brain is very good at but current systems generally lack [15].

A. Object Comprehension

Even toddlers know that objects are things which exist in a real world. Objects have multiple properties and can be discovered with multiple senses—they can be seen, touched, smelled, and tasted. Things may move about but they are generally permanent—they do not flash in and out of reality. Changing one's physical point of view of objects changes their visible appearance but does not change the objects. Object comprehension is an essential precursor to *understanding* in an AGI sense.

B. Time Comprehension

Objects may change over time—things were one way in the past and may be something different in the future based on actions taken by the observer or by others. Time comprehension is necessary to the goal-oriented behavior selections which require understanding of cause-and-effect relationships. An AI's ability to plan is very limited if it has no concept of time.

C. Learning New Algorithms

As an example, after memorizing the first set of numbers, children learn an algorithm for counting to arbitrarily large numbers. No matter how large a number is given, the next larger number can be derived given the algorithm. At a more fundamental level, behaviors can be considered algorithmic as they contain sequences of steps and may be varied by input parameters. Once one learns an algorithm for making words from sequences of syllables, the algorithm can accept inputs for pitch and duration and can be used for singing. The ability to learn an algorithm is essential to being able to learn new skills.

D. Current Components in These Areas

While these are huge areas of research, the robotics realm already has specific implementations of portions of these missing links. Some robots can traverse an environment and build an internal model of their surroundings. Robots can interact with physical objects and could potentially learn cause-and-effect relationships based on their actions. Some robotic systems can learn new sequences of behaviors as opposed to being fully pre-programmed.

It is by no means guaranteed that adding these three capabilities to AI will generate AGI but the converse is true that without these capabilities AGI is unlikely. It is difficult to imagine that a language-processing system (think IBM's Jeopardy-champion Watson [16]) could understand the meaning of "cat" based solely on language input on cats. Likewise, a DeepMind program which has only been shown still pictures of cats will have a limited but completely different concept of cats. Combined, these two approaches will still fall short of the concept of cat, which most children can acquire at an early age.

As multi-sensory robot software (with internal reality modelling) is merged with other AI technologies we can expect an added dimension toward AGI processing.

VII. The Case for Gradual AGI Emergence

Alexa uses lots of words but comprehends none of them. Recognized words may trigger useful responses, but just as often, the lack of actual understanding makes an Alexa or similar system much less useful than it might be. Therefore, there is a strong drive to develop genuine understanding and add it to AI systems.

This drive will result in the addition of various algorithms and the transfer of existing algorithms to new and broader areas of applications. Such development cannot happen instantly. AI systems will add some features and hone others and will gradually add more intelligent features (including those in the previous section) over a decade or two. At some point such systems may be considered to be AGIs but such a transition will be so gradual that it will be impossible to say that one system is an AGI while its predecessor was not.

As an alternative, consider that a computer is created which has the computational capacity of the human brain and was constructed along a similar model. After such a creation it would be years (perhaps 20) before it gained enough ability to determine whether or not it was an AGI. So, even the singularity model of AGI emergence is gradual.

Accordingly, this paper proposes a very soft AGI emergence. Systems will have basic abilities added to their repertoire over time and these abilities will likewise, gradually increase in power and performance. The line between non-AGI and AGI systems will be blurred. However, the ethical concerns about AGI will be applicable while such systems are still emerging. Most issues about AGI are applicable even if such a system has only the abilities of a three-year-old.

VIII. The case for Sooner AGI Emergence

AGI will emerge gradually as individual algorithms are merged to broaden the scope of AI systems. Since many AI algorithms already exceed the ability of a human (within a narrow domain) a combination of algorithms will necessarily exceed human abilities in the multiple areas it addresses.

The question is: When will such a system be broad enough in scope to be considered an AGI? The contention is that a robotic system can currently be adapted to learn the spatial relationships which underpin objects. Their ability to move about and act on their environment will allow them to learn about time and causality.

Married with advanced vision, learning, speech-handling, knowledge, internal modelling, and algorithmic learning, such a system would exhibit many features necessary for AGI in the next five to ten years. Whether or not such a system would, in fact, meet some criteria for AGI is largely a matter of definition. Such a system would be able to navigate within a real-world environment, learn about new objects, and plan actions based on expected results measured against goals.

When such a system shows glimmers of success, several points must be kept in mind:

- Human beings born with immense brain/computation power still require two decades of learning to become fully functioning. In every field of AI, this training time has been reduced dramatically with systems requiring hours or days of training often considered to be too time-consuming to be useful.
- Trained AI systems can be cloned. If one system learns a valuable skill, other similar systems could download that skill with essentially no training time whatsoever.
- Cloned systems need not be robotic. Once the skills of spatial comprehension, etc. are learned via real-world interaction, these skills can be transferred to non-robotic systems. In the same way that humans who lose physical and sensory abilities can still make use of the mental processes which were learned with those previous sensory or physical abilities.
- Computational power is continuing to increase. Despite the physical limitations of Moore's Law, supercomputers with immense parallel processing power continue to advance and become available at progressively lower cost. Once systems have any AGI abilities, more powerful versions will become available only a few years later.

Whether such systems are conscious entities, whether they have some deficiencies relative to the human brain, and when, specifically, AGI might emerge, are all questions which are not critical to addressing ethical issues about AGI. One can reasonably predict that an amalgam of currently available algorithms would create a system of significant intellectual power within the coming decade.

IX. The Question of Danger

A good analogy to the danger of AGI would be genetic engineering. Like genetic engineering, AGI has unlimited potential for benefits to humanity but at the same time, (like genetic engineering) AGI could be used carelessly or maliciously with catastrophic results [17].

A. Debunking

Many of the fears promoted by modern-day philosophers or by science fiction writers are completely unfounded. AGIs will be goal-directed systems and the selection of appropriate goals is, of course, crucial to benevolent operation. But the idea that a poorly selected goal is catastrophic presumes that AGI developers do not try out various goals on a small scale before implementing them more generally.

Consider an example from one prominent AGI author, of a system which, when given the goal of keeping humans happy might create a system which keeps all of humanity in a euphoric opioid stupor [18]. It is ridiculous to assume that 1) a system smart enough to implement a universal opioid haze would not be smart enough to recognize the underlying intent of the goal and 2) that such a system would emerge so abruptly that goals could not be tuned as the initial actions are observed.

The science fiction picture of AGIs necessarily trying to take over the world is likewise off the mark. AGIs will have their own goals of getting the energy and resources necessary for their own operation. Generally, these goals are not in conflict with humanity. Most related science fiction is actually about human goals and aspirations taken to a mechanical extreme.

Even the idea that an AGI needs a self-preservation goal is not necessarily true. An AGI with complete backups is essentially immortal. If a robotic body is destroyed, the "being" represented by the backup can be fully restored on replacement hardware. Rather than perceiving death, an AGI whose "body" is destroyed or damaged would have a period of unconsciousness and would re-emerge as good as new. An AGI would consider this as dangerous as sleep. On analysis, a spontaneous AGI Armageddon is not a likely scenario.

B. A Clearer Picture

Initial goals can be given to AGI systems which will attempt to be as benevolent as possible. These will certainly be subject to unintended consequences and the goals will be

adjusted as a system is trained and the behavior is observed. As AGI emerges gradually, there will be time to correct errors, which might have made systems dangerous.

If goals are given which direct AGIs to be aggressive, violent, territorial, etc., these could be considered as "weaponizing" AGI technology as opposed to risks inherent in the technology. This is a very real possibility but should be addressed differently from an inherent risk.

Further in the future, one can assume that AGIs progress in ability to a point where they set goals themselves. At that point, one could assume that AGIs set goals for the benefit of AGIs. On the plus side, the needs of AGIs are largely divergent from the needs of humanity. AGIs will not need territory, food, water, or control over humans. They will need energy and the factories and resources to build more AGIs. As such, direct conflict between AGIs and humanity is not a sure thing—AGIs may set goals for their own space exploration and research which do not impinge on humanity at all. On the other hand, human over-expansion and damage to the planet may cause AGIs to react. In this instance the question is whether the problem is one of AGI behavior or human behavior.

X. Brain Augmentation and Uploading

Recent articles have given credence to the ideas of using vast future computer power to augment or even replace human brain activity.

A. Augmentation

Considering augmentation, we have plenty of systems which can accept a person's verbal or keyboard requests and create some (usually) useful response. The concept of brain augmentation replaces the verbal or keyboard request with one which comes directly from the brain and routes responses directly back to the brain. Whether this proves to be more efficient than a verbal or keyboard interface remains to be seen. Because it can be presumed that the brain is currently working as fast as it can, having a direct connection between a brain and a computer might not yield any speed improvement. Further, the quality of the response will not change much just because the interface method changes. Thus, it is likely to be, at best, a niche technology because of the inherent difficulties in creating a direct brain interface with a limited improvement in responses.

A more valuable use for this type of technology will be as a replacement for the benefit of individuals who have lost various abilities. Direct brain control of robotic limbs, and direct brain reception from artificial eyes and ears would be much more likely to develop.

B. Uploading

The prospects for uploading one's entire brain content is even more remote. If we consider that the human brain is a generally intelligent system, in order to upload its content, a system capable of supporting intelligence must first be developed. That is, a system which *could* be an AGI must be developed prior to the upload. Then, the monumental technical problems of scanning the content of the brain and figuring out how to translate that scanned information to the new hardware must be addressed. Thus, a system for uploading brain content is a significant superset of the problem of AGI alone and will necessarily occur significantly later (if at all).

XI. Shifting Ethical Issues for AGI

There are many ethical issues related to today's AI technology [19]. As AGI emerges, some of these will fade in importance, some will morph, and new issues may arise.

A. Training Sets

Many of today's AI systems have a training phase and are subsequently used to produce results without additional training. Shortcomings in the training sets represent the cause of some of the issues with these systems [20]. AGIs, on the other hand, will necessarily continue to learn as they operate and, therefore, their behavior cannot be controlled by selection of a specific initial training set. Issues which have previously been addressed by modifying training sets will necessarily need different solutions [21].

As an example, consider an AGI which (like a person) might use the Internet as a significant information source (training set). Not only is the information set ill-defined but it is in a constant state of flux.

B. Black Boxes

Once trained, fixed-training-set systems are usually unable to give the basis for their results and are treated as black boxes. Those systems which *can* explain their bases for a result often give surprising insight into their operation [7]. A person giving a justification for a decision at least gives the opportunity for that justification to be analyzed and reviewed for unsubstantiated bias [22].

An AGI will also develop biases from the content it encounters. When asked to make a decision, such an AGI will also bring all these biases into play and create a justification in the same way a human might, which may or may not enumerate the biases behind the decision. The reasons an AGI states as the basis for a decision will be more akin to a human rationalization for behavior. As such, the issue of an unexplainable black-box AI does not go away, it morphs into a more human-like issue. AGI decisions and justifications will, at least, be open to examination and evaluation.

C. Consciousness

The possibility that AGIs might become conscious entities introduces additional ethical issues. We currently have no definitive way of determining whether a person or AGI is a conscious entity or not. We *could* define consciousness in terms of behaviors (such as planning or self-interest) but we are aware, ourselves, of the subjective sensations (qualia) of being conscious entities. While an AGI can conceivably reproduce all the behaviors we might associate with human consciousness, we will likely not be able to determine whether or not AGIs have any internal sensation which might be analogous to our own.

In light of our inability to determine the presence of consciousness in any concrete way, should we ascribe any special status to entities which *might* be conscious? As society already has standards for the ethical treatment of animals, similar thinking might be applied toward AI systems even before the emergence of full AGI. With animals, unnecessary pain is to be avoided, but killing individual animals is generally acceptable under some circumstances. This attitude relies on the animals' inability to object. In the case of AGIs, it may be enlightened self-interest to consider what the AGI's position may be on various upcoming ethical questions involving them.

XII. Conclusion

This paper has presented a case for AGI development which is gradual, soon, and moderately dangerous. As portions of AGI emerge over coming decades, ethical issues will morph as well. Today's ethical concerns about training sets and algorithms will change as different algorithms and learning from experience begin to predominate.

As society becomes more reliant on, and accepting of, results produced by AGIs, there will be a better opportunity to examine the rationale behind AGI results. On the other hand, there will be less ability to control either training or outcomes. As systems become more human-like in their intelligence, they may also become more humanlike in their foibles.

Next steps toward developing ethical AGIs would include building a consensus on the overall outcome—what the optimal relationship between people and AGIs should be. As time goes on, there will be progressively more accurate representations of what AGIs might actually do and continuing attention needs to be paid to how such advances will necessarily impact our ethical positions.

References

[1] B. Tomasik,"Predictions of AGI Takeoff Speed vs. Years Worked in Commercial Software," First published: 2014 Dec 09. Last nontrivial update: 2019 Jul 21. [online] Available: https://reducing-suffering.org/predictions-agi-takeoff-speed-vs-years-worked-commercial-software/

[2] AI Multiple, "373 experts opinion: AGI / singularity by 2060 [2019] update]", [online] Available: https://blog.aimultiple.com/artificial-general-intelligence-singularity-timing/

[3] S. Baum, "A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy," (November 12, 2017). Global Catastrophic Risk Institute Working Paper 17-1 [online] Available: https://ssrn.com/abstract=3070741

[4] J. Rodriquez, "The Five Tribes of Machine Learning", [online] Available: https://medium.com/@jrodthoughts/the-five-tribes-of-machine-learning-c74d702e88da

[5] T. Foisotte, O. Stasse, Adrien Escande, P. Wieber, A. Kheddar, "A Tow-Steps Next-Best-View Algorithm for Autonomous 3D Object Modelling by a Humanoid Robot," 2009, IEEE International Converence on Robotics and Automation, pp. 1159-64.

[6] K Leetaru, "Deep Learning And The Limits Of Learning By Correlation Rather Than Causation,", *AITopics, An official publication of AAAI*, May 20, 2019. [Online] Available: https://aitopics.org/doc/news:D85F4E10/

[7] M. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," 9 Aug 2016,[online] Available:https://arxiv.org/pdf/1602.04938.pdf arXiv:1602.04938v3 [cs.LG].

[8] V. Mnih, K Kavukcuoglu, D Silver, A Graves, I Antonogou, D Wiestra, M Riedmiller (DeepMind Technolgies), "Playing Atari with Deep Reinforcement Learning," 19 Dec 2013, https://arxiv.org/pdf/1312.5602v1.pdf arXiv:1312.5602v1 [cs.LG].

[9] R. Kurzweil, "Exponential Growth in Computing," in *The Simularity is Near: When Humands Transcend Biology*, Penguein Books, 2005, p70..

[10] P. Clark, et al., "From 'F' to 'A' on the N.Y. Regents Science Exams: An Overview of the Aristo Project," Sep 2018 arXiv:1909.01958v2 [cs.CL]

[11] K Drexler, "How do neural and symbolic technologies mesh?" in *Reframing Superintelligence, Comprehensive AI services as General Intelligence, Technical Report #2019-1,* Future of Humanity Institute, University of Oxford, UK, 2019, pp. 161-170.

[12] E. Cambria, T. Young, D Hazarika, S Poria, "Recent Trends in Deep Learning Based Natural Language Processing", *IEEE Computational Intelligence Magazine,* August 2018, pp 55-75.

[13] J Mao, C Gan, P Kohli, J. Tennenbaum, J Wu, "The Neuro—Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences from Natural Supervision," ICLR 2019 conference paper, Available: https://openreview.net/pdf?id=rJgMlhRctm

[14] M. Miller, *Building Minds with Patterns*, Piaget Modeler, Beverly Hills, CA, 2014-18, pp. 137-163.

[15] S. Adams, et al. "Mapping the Landscape of Human-Level Artificial General Intelligence", *AI Magazine*, Spring 2012, pp 25-41.

[16] D. Ferrucci et al, "Building Watson: An Overview of the DeepQA Project," *AI Magazine*, Fall, 2010, pp. 59-79.

[17] C. Simon, "Will Computers Revolt", in *Technological Forecasting and Social Change*, Elsevier, Vol 146, September 2018, pp. 81-87.

[18] N. Bostrom, "Malignant failure modes," in *Superintelligence Paths, Dangers, Strategies*, Oxford University Press, UK, 2014, pp. 146-154.

[19] S. Russell, P. Norvig "26.3: The Ethics and Risks of Developing Artificial Intelligence," in *Artificial Intelligence: A modern Approach*, 3rd ed. Prentice Hall, Upper Saddle River, NJ, 2009.

[20] P. Besse, C. Castets-Renard, A. Garivier, J. Loubes (2018). "Can Everyday AI be Ethical? Machine Learning Algorithm Fairness (english version)." [online] Available: https://www.researchgate.net/publication/329277474_Can_Everyday_AI_be_Ethical_Machine_Learning_Algorithm_Fairness_english_version 10.13140/RG.2.2.22973.31207.

[21] S. Lakra, T. Prasad, G. Ramakrishna, "The Future of Neural Networks," Proceesings of IndiaCom-2012, 10.13140/RG.2.1.2390.3848.

[22] D. Greene, A. Hoffmann, L Stark, "Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning," Proceedings of the 52nd Hawaii International Conference on System Sciencecs, 2019, pp. 2122-2131. Available: https://hdl.handle.net/10125/59651 10.24251/HICSS.2019.258.