The role of Human Factors and Ergonomics in AI operation and evaluation

Nikolaos Gkikas[1], Paul Salmon[2] & Christopher Baber[3]

[1]Airbus Defence & Space, [2]University of the Sunshine Coast, [3]University of Birmingham

ABSTRACT

The present paper sets the scene for a recorded workshop exploring the critical role of Human Factors and Ergonomics in the development, operation and evaluation of Artificial Intelligence. First, we lay out some foundations of the multidisciplinary developments commonly placed under the umbrella of "Artificial Intelligence/AI" and propose some fundamental definitions to facilitate the structure of our arguments and the foundations for the workshop. Then we explore the role of Human Factors and Ergonomics methods in ensuring that AI systems contribute to our disciplinary goal of enhancing human health and wellbeing. In closing we propose a research agenda designed to ensure that Human Factors and Ergonomics is applied in future AI developments.

KEYWORDS

Artificial Intelligence, Machine Learning, Human Factors, Ergonomics, Methods

Introduction

At the time of writing of this paper, there is a frenzy around the term "artificial intelligence/AI" in the mainstream media reporting on various levels of systems complexity across virtually all products and services, from number plate recognition at car parks, to virtual assistants and Google search. That is because "Artificial Intelligence" is loosely and subjectively defined. Definitions seem to reflect more about the subjective understanding of an artefact's behaviour by the observer, than anything objective about the specification of the artefact itself. To counter this, we define and use the "AI" interchangeably with the term "non-deterministic systems". These are artefacts which have been specified, designed, programmed to continuously adapt their behaviour to their operational environment. Within that paradigm, the relationship between input and output is not fixed or determined directly by the designer of the artefact; it is rather fluid and changing according to its Machine Learning properties and the feedback received from the operational environment. "Machine Learning/ML " is the subfield of "AI", and it concerns the methods (in the form of algorithms) that enables artefacts to express learning behaviour and adaptation to an operational environment.

Artificial intelligence

The term 'Artificial Intelligence' was first coined in the 1950s by Dr John McCarthy, an American scientist working at Dartmouth College (McCarthy, 1955). The field of AI was established soon after, and early definitions of AI focussed simply on the capacity of machines to perform tasks that would normally require human intelligence (McCarthy et al., 1955; Minsky, 1968). Contemporary definitions have a broader focus, referring to non-human agents with the ability to interpret and learn from data in pursuit of a specific goal. Kaplan & Haenlein (2018), for example, define AI as systems with the ability to "interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation".

ANI systems are now well established. Well known examples include Facebook's facial recognition system, Apple's personal assistant Siri, and Tesla's self-driving vehicles (Kaplan & Haenlein, 2018).

The "holy trinity" of machine learning: supervised learning, non-supervised learning, deep learning

Supervised learning, is a form of machine learning whereby an "answer key" is in place to provide the "ideal output" that the artefact is continuously optimising its behaviour towards. Linear and non-linear regression are common parts of such learning algorithms.

Unsupervised learning.is a form of machine learning whereby there is no "answer key" included by the creator or added by a human during operation; instead, the artefact is continuously looking for mathematical or logical patterns in the accessible pool of data (which pool itself could be continuously evolving). In practice, despite the absence of direct human involvement in defining such patterns, there is an opportunity to modify the properties of these, e.g., by setting the hyper-parameters that the models apply. For instance, the user could specify whether the model would produce, say, 3 or 5 patterns.

Deep Learning involves the use of unsupervised learning methods (e.g. neural networks) to not only discover patterns but also to develop policy that optimise a 'reward', thus including a self-determined module of supervised learning. It is therefore easy to omit that even within this "deep" form of machine learning there is human involvement, such as the need for a human to provide a definition of 'reward' (such as win a board game) and a definition of 'action' (such as how pieces on the board can be moved), and the computer will play millions of versions of the game to arrive optimal strategies for playing the game under any circumstance. While the results of this approach (applied to board games or to image analysis) can be highly impressive, not that the problem domain in which these operate tend (at present) to be well defined, e.g., in terms of what defines a ‚reward' or an ‚action'.

How did we get here

The mainstream view is arguably that non-deterministic definition, especially in the form of programming is a field of computer science (Goodfellow et al, 2016), predominantly driven the **urge for prediction**. The most often quoted methods used by professionals of computer science backgrounds are: neural nets, Bayesian logic, various types of regression models. The latter should be familiar to professionals and researchers tagged as "human scientists" - such as human factors and ergonomics professionals. In fact , despite the lack of promotions, another force behind non-deterministic programming comes from statistics (Gareth, 2013) and the urge for to analyse data **for inference**. One way or another, those two aims seem to define the applications of non-deterministic programming/machine learning/"AI" today.

Applications and interfaces with HFE: the main focus of our workshop

The first and main use of various machine learning applications was and still is as a tool for data-analysis. Both as a cause and as an end to efficient analysis of massive dataset, exceeding the experience of the mere researcher, the flexibility offered by machine learning is matched by the ability to rapidly iterate millions of analyses. The successes of "AI" in drug discovery are fundamentally a case of machine learning application as a research tool. The second most common application of machine learning is as a tool for process/product improvement or development. These applications use machine learning which is continuously fed with life in-service data and feedback of a given operation. The main advantage of such application is the rapid and variety of alternative architectures in terms of people, process and equipment can suggest and model their performance.

The third application is machine learning as a product/service in its own right. Such applications take the form of "virtual assistants", and they have essentially evolved from the first application quoted above (data analysis tool), with the addition of a user interface. The use of those "assistants" in commercially available devices such as smartphones, are fundamentally a machine-learning search engine with an auditory and/or visual interface for use to interact with.

The third application above, with the user interface being the part setting in apart from the "data analysis tool" case, should provide a clear area for HFE professionals to contribute to. There are specific challenges due to the flexibility in the input/output, control/feedback loops compared to traditional HMI (Gkikas, 2019); however, these challenges are there for the HFE community to tackle. Then, the learning and subsequent effect of machine learning during process improvement applications is another new challenge but clearly the realm of the Human Factors Integration professional community. The question remains of course whether the existing evaluation and practice methods are fit for purpose when dealing with an intelligent system that generates multiple alternative scenarios and potential solutions in shorter time than a human operator can assess one of those scenarios.

Artificial General Intelligence and HFE: the next train departing and how the HFE methods fit in.

Whilst a failure to embed HFE in the AI lifecycle may lead to performance issues, accidents and even major catastrophes, a similar failure with the next and more advanced generation of AI could spell the end for humanity. So-called Artificial General Intelligence (AGI) systems will be equipped with advanced computational power, will be able to perform all of the intellectual tasks that humans can and will be able to learn, solve problems, adapt and self-improve, and undertake tasks for which they were not originally designed (Bostrom, 2014; Everitt et al., 2018; Gurkaynak et al., 2016; Kaplan & Haenlein, 2018). They will also have the capacity to control themselves autonomously; having their own thoughts, worries, feelings, strengths, weaknesses and predispositions (Muller et al., 2016; Pennachin & Goertzel, 2007). Whilst AGI systems do not yet exist, credible estimates suggests they could emerge by 2050 (Muller et al., 2016).

Given their projected capabilities, AGI systems could revolutionize humanity. Potential benefits discussed include curing disease, revolutionising the nature of work, and solving complex environmental issues such as food security, oceanic degradation, and even global warming. However, it is widely acknowledged that a failure to implement appropriate controls could lead to catastrophic consequences. Various adverse impacts have been discussed, with the most pessimistic viewpoints suggesting that AGI will eventually pose an existential threat to humanity (Bostrom, 2018).

The intelligence explosion is a much-discussed scenario whereby rapidly self-improving AGI systems become far more advanced than their human counterparts (Bostrom, 2014). This ‘super-intelligent’ AGI is the source of most scholars’ concerns. Bostrom (2014), for example, argues that the intelligence explosion will eventually lead to humans becoming obsolete and then extinct. The Future of Life Institute (FLI) identify two worrying scenarios. First, that super-intelligent AGI systems will be designed to do something devastating, such as kill (e.g. an AGI-based autonomous weapons systems), and second, that they will be designed to do something beneficial, but develop a destructive method to do so (e.g. a cancer prevention AGI system that decides to kill everybody who has a genetic predisposition to cancer). The latter argument is based on the notion that an AGI with a goal of some sort will seek more efficient ways of achieving its goal as well as more resources to do so.

Regardless of whether the singularity is ever reached, there are other significant risks associated with AGI. These include the malicious use of AGI for terrorist and cyber-attacks, population control

and manipulation, a replacement of the human workforce, and mass-surveillance to name only a few. Across the literature there is widespread agreement on the need for urgent research into how best to design and manage AGI systems so that these kinds of risks are minimised (Amodei et al., 2016; Bostrom, 2014; 2017; Brundage et al., 2018; Omohundro, 2014; Steinhardt, 2015).

The potential role of HFE in supporting the design and operation of safe AGI is compelling. Leading figures in the field of AI have discussed the need for designers to more fully consider the risks associated with AGI and to place more emphasis on how AGI will interact with humans and align with our goals. Of course, these considerations are precisely what HFE focuses on in many areas. It is therefore our view that HFE has a critical role to play in the design and operation of safe AGI (Salmon et al., In Press). Based on our work in safety and risk management, we recently identified three forms of AGI system controls that urgently require development and testing (Salmon et al., In Press):

Controls to ensure AGI system designers and developers create safe AGI systems;

Controls that need to be in-built into the AGIs themselves, such as "common sense", morals, operating procedures, decision-rules, etc; and

Controls that need to be added to the broader systems in which AGI will operate, such as regulation, codes of practice, standard operating procedures, monitoring and maintenance systems, and infrastructure.

Based on this Salmon et al. (In Press) reviewed fifteen categories of HFE method to determine whether they could be used in AGI design and management. These categories of HFE method ranged from task and cognitive task analysis, workload and situation awareness assessment to risk assessment and systems analysis and design methods. All 15 categories of HFE method were deemed to be suitable for use in support of AGI system design and management. Critical areas where HFE can contribute included: risk assessment, the design of risk controls, human-AGI interactions, teaming, standard operating procedures, dynamic function allocation, usability assessment, AGI errors and failure, and also aspects of AGI cognition, such as decision-making, situation awareness and cognitive workload. In addition, systems HFE methods such as Cognitive Work Analysis (Vicente, 1999), the Systems Theoretic Accident Model and Process (STAMP; Leveson, 2004), the Networked Hazard Analysis and Risk Management (Net-HARMS; Dallat et al., 2018) and Agent-Based Modelling (ABM; Bonabeau, 2002) were deemed to be particularly suited to the design and testing of appropriate AGI controls.

In the recorded workshop, we will work through an AGI case study scenario designed to showcase how HFE methods can be used to support the design of safe AGI.

References

Amodei., D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mane, D. (2016). Concrete problems in AI safety. AI, 1-29.

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Inc. New York, NY, USA

Bostrom, N. (2017). Strategic Implications of Openness in AI Development. Global Policy, 8:2, 135-148.

Bonabeau E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. Proc Natl Acad Sci, 99:3, 7280-87.

Dallat, C., Salmon, P. M., & Goode, N. (2018). Identifying risks and emergent risks across sociotechnical systems: The NETworked Hazard Analysis and Risk Management System (NET-HARMS). Theoretical Issues in Ergonomics Science, 19(4), 456-482.

Everitt, T., Lea, G., Hutter, M. (2018). AGI safety literature review. IJCAI. arXiv: 1805.01109.

Gareth, J. (2013). An Introduction to Statistical Learning. Springer: New York.

Gkikas, N. (2019) Allocation of Function in the era of Artificial Intelligence: a 60 year old paradigm challenged. Contemporary Ergonomics 2019. Taylor and Francis: London.

Goodfelow, I., Bengio, Y., and Courville, A,. (2016). MIT Press: Boston, MA.

Gurkaynak, G., Yilmaz, I., Haksever, G. (2016). Stifling AI: Human perils. Computer Law and Security Review, 32:5, 749-758

Kaplan, A., Haenlein, M. (2018). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons, 62:1, 15-25

Leveson, N. G. (2004). A new accident model for engineering safer systems. Safety Science, 42:4, pp. 237—270.

Müller, V. C., Bostrom, N. (2016), 'Future progress in artificial intelligence: A survey of expert opinion', in Vincent C. Müller (ed.), Fundamental Issues of Artificial Intelligence (Synthese Library; Berlin: Springer), 553-571.

Omohundro, S. (2014) Autonomous technology and the greater human good, Journal of Experimental & Theoretical Artificial Intelligence, 26:3, 303-315.

Pennachin, C., Goertzel, B. (2007). Contemporary Approaches to Artificial General Intelligence. In B. Goertzel & C. Pennachin (Eds.), Artificial General Intelligence, Springer.

Steinhardt, J. (2015). Long-Term and Short-Term Challenges to Ensuring the Safety of AI Systems. https://jsteinhardt.wordpress.com/2015/06/24/long-term-and-short-term-challenges-to-ensuring-the-safety-of-ai-systems/