Creating Something Useful for the Town

Conceiving Artificial Intelligence Machines for the Current and Future Society

Vlad-Stefan Tudor

2020

***Abstract*—The article aims to explore the questions of innovation and utility with regards to the development of technology, in particular Artificial Intelligence (AI). The article aims to investigate motivation, consequence and underlying phenomena and draw a set of conclusions on top of which several ideas are to be proposed and discussed.**

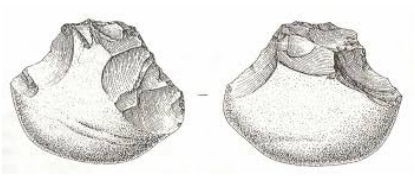

Fig. 1: Silex stone tools from Paleolithic

I. What is usefulness

Cambridge Dictionary defines *useful* as the property of an object to *help one do or achieve something*. The intrinsic incompleteness of this definition is not something anyone will complain about. However, a mathematical modelling of such a characteristic is not a trivial thing to do. When dealing with an AI, the level of patience required is infinitely higher than when answering to a child's question of *Why is that*. One cannot simply assume that there will be any intuition to spare the task of a complete definition.

The usefulness of an object is fundamentally tied to its means of exploit. How useful was gunpowder for digging up mines and how useful was it for preserving human life? How useful is a computer for an illiterate person? If one sets out to find something useful one has both to investigate the problems one's peers face but also their capabilities - and how said peers may re-purpose the invention.

II. The universe of effect and not of intention

The more revolutionary an invention the more potential it has for miss-use. If one sets out to find a way to dramatically improve the life of his fellow humans, one may easily create the means for its destruction. The underlying pessimistic viewpoint of humanity is that global optima is not only achieved by over-performing neighbourhooding optima but also by *eradicating* it. We have to remember that the first nuclear reactor ever built was only as a part of the Manhattan project, whose main purpose was building the Allied atomic bombs.

III. Writing a Scope of Work for an AI

The means by which automation can be of use to humans are in the reach of imagination both for the individual educated in the field and for the one who only has a few basic ideas. However, the most crucial aspect is not identifying a problem to be solved, nor is the implementation of a solution, as is correctly formulating the objective. An AI will rapidly identify by trial and error eventual loop-holes in its objective. It is not the difficulty of modelling the task to be executed as is the thoroughness required in avoiding frustratingly logical work-arounds.

In the paper *Concrete Problems in AI Safety*[7] multiple researchers tackle the most common short-cuts an AI can find when given a poorly formulated problem. In short, when evaluated against a score that should represent the correct accomplishment of a task, an AI can very efficiently identify loop-holes that, even though maximize the said score, totally deviate from the desired behaviour.

In the article they suggest the example of a cleaning robot. It easily goes to show that optimizing the speed of this task without proper constraints, can lead to the destruction of some of the objects to be cleaned, since keeping them intact was not defined as part of the objective. Moreover, AI is prone to *Reward Hacking* meaning that the score of a clean room can be achieved also by disabling vision when confronted with a messy room. (It will not see that it is dirty, hence the cleanliness score is high). There are practical examples of AI trained to play arcade games only for them to discover certain glitches that allow for artificially increasing the score without progressing through the game itself.

Usefulness is defined with regards to one particular task. However, this is sufficient only because humans are fairly good at common-sense reasoning. A hammer can be considered intrinsically useful only because we don't suppose to use it at anything besides construction work. But sometimes we do consider things for tasks that are totally remote from their common usage and we call it 'improvisation'. But there is a certain compatibility between the 'useful tool' and any task for which we think of using it. And that intuition is dangerously hard to model for an AI.

IV. The paradox of the relief from work

A. *Employee of the month*

The scientific community does provide some approximate milestones in the development of AI [8] but is not totally in agreement of the necessary time for each one. They give as a 50% chances that AI:

- codes simple algorithms by 2025
- works as Salesperson in Retail by year 2030
- writes a Best-Seller by 2050
- performs surgery by 2053
- researches AI by 2100
- replaces all human labour by 2150

These are approximate predictions that may likely be exaggerated (in the same study it was given as a 50% chance that an AI will win a Poker Tournament by 2020). But it is clear that the study illustrated one of the main fears of general population regarding AI: replacing human workforce.

It's unlikely that any level of labour automatizing will not further require humans. As an example of a possible career in the future, let us consider that a however computationally powerful AI needs rigorous prior training on the task it will perform. That is not to be confused with researchers programming the architecture and the objectives. The AI will need tutoring from humans that formerly performed the same job the AI would perform.

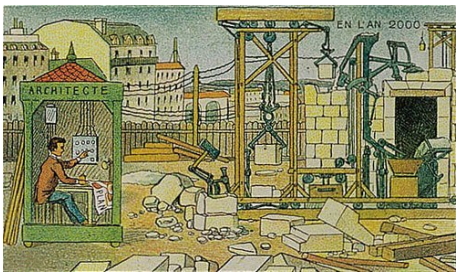

Fig. 2: 1910 Illustration: The World in year 2000 [5]

We may see that a considerable fraction of former labourers will continue to perform their work but rather as a simulation, in order to provide necessary data for the training of the AI. We may even see the re-appearance of craftsmen that currently struggle to compete with efficient mass-production. Training an AI for a job not only requires recordable data but also real-time feedback and guidance on how things should actually be done.

With the current progress in autonomous vehicles we might think that the job of a driver is in great peril. And indeed is not a profession that will see the same level of job opportunity as in the present. But it will most certainly not be eradicated. We have to consider that there are limit cases like avoiding an accident for which an AI's original work-around avoiding the common-sense objective is not to be desired in the least. The examples are countless for other similar *blue-collar* jobs. The formidable resource is that any ordinary man, even the totally uneducated, still posses a level of common sense that is very hard to train into an AI from scratch.

B. *Orwell's Argument*

In the aftermath of the Second Industrial Revolution, the 20^{th} century has seen technology both as an ideal and as a dystopia. It was the time of the automobile and the aeroplane and the sky was the limit. And it was a limit to be conquered as well, not long after, but with the same rockets that in the 40s ravished London and made the dreamers hide in fear.

In his 1937 book 'Road to Wigan Pier'[11], author George Orwell proposes some fundamental dilemmas regarding technology, and it was not regarding its evident destructive capabilities but rather it's more subtle consequences. He argues that machines ultimately bring *sub-human softening* and *hedonistic decadence* for the individual they have set free of labour. Humans will be left without the means to express their creativity as they will realise that they are easily outperformed by machines, he states. To remember their humanity, they will *exercise to harden muscles which they will never be obliged to use*.

Orwell puts forward the paradox that work is subjective and that what is effort to one can be occasional leisure to other. Accordingly, no-one would be able to enjoy doing something while conscious that all can be more efficiently done automatically. He predicts that art will, as well, be performed by machines (the term deep neural networks was not yet coined), further stressing on the future futility of activities traditionally considered specifically human. He sees *mechanical progress* as having the ultimate goal of *something resembling a brain in a bottle*, acknowledging for the non-literary form of this comparison but presumably pointing out to the trans-humanistic movements of today.

Orwell overlooked, however, a great deal of the future capabilities of machines, focusing on those that wore immediately similar to a human's, like it would be expected from presumptions made almost a century ago. The reality is that not all AI research focuses on copying human behaviour and abilities. There is a section of the spectrum that would allow even a sceptic like Orwell to see the true potential of machines - not in being like us, but in being exactly *unlike* us.

V. Types of AI Innovation

In a field that can unanimously fall under the category of innovation, such as AI, one has still to differentiate some nuances. This aims not to act as a definite categorisation, nor as a immutable viewpoint on existing research. That being said, there seem to be 2 types of usage for AI:

1) **Replacement** AI, like the examples provided in the previous section - tools for substituting human labor with automated systems, referring not only to physical tasks but also to tasks considered too computationally complex or too time-consuming. The main goal of this category is *efficiency*.
2) **Augmentation** AI, meaning the applications whose results produce surprise, help formulate new conclusions or provide unexpected viewpoints.

A more precise explanation is clearly needed. Novelty is something that is consistent in presence but inconsistent in form. Every generation has the set of tools, or using the modern term - *gadgets* - that follow an ascending and descending trend pattern, that grow exponentially in popularity until slowly fading off into the common and the everyday, being gradually replaced by others, newer inventions. We can conclude that this type of tools don't necessary bring information into the system - not in the overall sense. A telephone may help *communicate* information but the respective information has to already be part of the system, to have been already created. (Note: 'created' meaning 'being discovered' so as not to create side-tracking discussion on Information Theory).

Much like these perpetual novel tools, some AI is just a very powerful contraption for getting a desired result faster, cheaper and, in general, more efficiently. This is what falls under the first category, AI that ultimately seeks human commodity. It's indeed remarkable the level of closeness technology can now bring between the user and the desired product, and this ultimately, in a sense, benefits the individual. It is certainly regarded as such by the individual himself and most of it can even be objectively regarded as useful.

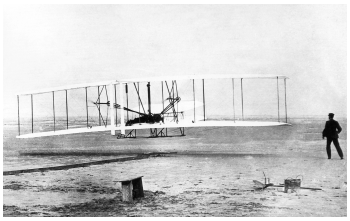

Fig. 3: First Flight of the Wright Brothers

Augmentation AI is the type of AI, however, that is not simply a tool but a true agent of information. It is not a replacement for a human nor is it intended to be alike one, but complementary to it. There exists certain tasks that allow for a individual's most valuable trait - *intuition* and there are tasks that frustratingly fail due to the individual's most personal weakness - *subjectivity*. There is no need to infinitely imitate humans, as there are known limits to our abilities. There are discoveries that await us, but they are incompatible with the human in its intrinsic form. They require *objectivity*, like any universal truth and how hard is for a human to detach oneself from one's own biases and prejudices? It is mere childlike curiosity that drives us to imagine human-like AI, that is with human-like perception and human-like actuators. Indeed the power of AI lays not in competing and outperforming humans but in providing the perspective humans irreversibly lack and, in return, humans providing the one thing AI may subsequently never truly *understand* - intuition, gut feeling.

In short the arguments refers to this: however useful it seems for a machine to perform a task faster and better than an individual, it should be considered of greater utility for the machine to perform a task that humans would not be able to perform given any number of individuals working for as many hours. It is not a matter of increasing efficiency but of crossing unimagined barriers in understanding.

VI. Looking forward

As an answer to the article's title, several ideas are proposed. They are given with regards to the discussion in Sections II - V, namely:

- It is admitted that no idea should be judged only by the means it can be miss-used since any new idea is prone to such a risk.
- A shift in the job market is natural and has always been - it is not by itself an argument for opposing technological progress.
- In the discussion of usefulness, a distinction will be made between machines increasing efficiency and creating information.

A. *Learning and Personal Tutoring*

It is regarded that there are seven main learning styles[3]. This may account for how conventional education sees an obvious decline even though general interest in knowledge does not follow a similar trend[4]. Younger generations increasingly demand from their teachers an adaptation of teaching methods, understandable examples and applications and in general, a subjective translation of the objective information. This may be caused by the habit of highly personalised goods and services that modern consumer market offers which schools contrast by remaining relatively constant and generic.

It is understandable that education has not yet been done with much regard to the proportion of learning styles in each individual student. It is simply unfeasible. However that can change with the ease by which an AI can learn and adapt to very particular traits and habits once a general framework is developed. Thus, the quality of education may be increased to the level of a personal tutor (one-to-one highly personalised education) while the quantity, or better said, the availability, simultaneously grows to the point that education truly becomes a right for every individual. It is possible to hope for this kind of goals once the main resource of education, the teaching power, becomes infinite.

Wide availability of information is already achieved without any truly intelligent algorithms (search engines should be considered as such only to a certain degree), but the initial effect is the creation of noise. There are simply too many sources of information for a novice to filter rigorous knowledge. Moreover, without some kind of guidance, people are inconsistent and prone to frustratingly give up. An AI Tutor that would select desired information, arrange it in the order that assures maximum continuity and represent it accordingly to each one's specific learning style is the level our hopes can rise to with the present progress in the field.

B. *Pattern Analysis in Big Data*

Statistics and probability are fields in which humans optimistically consider themselves potential experts. However even the simplest experiments can illustrate faulty rationale even in educated individuals. One may acknowledge highly probable scenarios but only for anyone except oneself. The degree by which this counter-intuition affects us is shown in generic slogans such as *It can happen to you*, as if such a reminder is necessary for such a self-evident assumption. This extends to how people so oftenly miss-label marginal but probable situations as *coincidence*. Media fake-news is mainly based on poor understanding of Bayesian Logic [6], proving that, in general, individuals are bound to subjective interpretation more than to rational reasoning. With larger data sets, these effects will only amplify. Holding on to one's personal believes against all factual data has been the ruin of uncountable many stockbrokers. This just goes to show that understanding patterns and mechanisms above the level of palpable matters is not intrinsic human.

Big Data analysis often leads to discovery of patterns that are invisible at lower scales [1]. Correlations cannot be made on small samples and concluding a social study has relevance only if the researchers account for diversity and un-biassness. This is almost impossible to achieve without AI that can look through large data-sets. Moreover, such studies are done with the objective of either confirming or disproving a prior presumption. This can act as a bias factor in itself. As opposed to this, an AI can be trained to look for patterns in however random data. There is no problem of subjectivity dragging conclusions in a certain direction.

The possibilities are indeed endless, and algorithms are already processing immense amounts of data. It is true that most of these applications are strictly commerce-oriented and their utility is closely tied to profit but the door is open for much more than that. As an analogy, think of how the first calendars were developed. They might have been looked upon with the same skepticism that these algorithms are today by almost half of the population [10], [12]. But the reality is that long-term usefulness is not defined by the first impression. One has to see past the fear of the new, even though a level of skepticism is generally a safe approach.

C. *Causality and Decision Micro-management*

Another classical weak-spot for humans is correctly interpreting cause and effect and following the consequence tree down the line. As a brief introductory example, note how early humans considered meteorological phenomena to be consequences of their own actions judged by a divine power. It goes to show that humans are bound to find causality even if their is none in order to produce some sort of explanation aligned with their current understandings. The reciprocal is also true and its best illustrated by present-day self-destructive habits and vicious cycles to which so many obliviously commit.

A chaotic system is a kind of system whose evolution in time is strongly dependent on initial conditions [2]. Such a system, even if it is a cumulus of deterministic laws and operations, is usually considered unpredictable, referring not to it being impossible to compute and analyse but unpredictable by human intuition. This a edge case for the lengths of our judgement, showing its limitations in the most direct manner. Most of the complex systems that govern our everyday life have emergent chaotic results, that cannot be trivially grasped. This may come as a burden to the individual that, faced with a choice, feels the need of broader understanding of the situation.

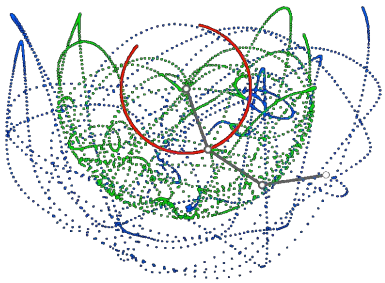

Fig. 4: Evolution of a Chaotic System, a Triple Pendulum

The effect of the inability to efficiently process cause and effect is the phenomena known as *Decision Fatigue*. Choice is a fundamental goal of humanity, seen as an indicator for liberty and it is also argued that it was the main reason for evolution of cognition. However, there seems to be a trade-off: *effortful choice is costly — and that the cost is paid in the same coin that pays for effortful self-regulation*[9].

There is an emerging paradox for the modern individual. Historically, the effort of choice was substantially less than the effort of accommodating an external decision (following one's orders), motivating an underlying ideal for liberty of decision. This manifests in the desire of wider and wider range of products and services. This, in turn, creates the premise of choice not as liberating, but as tiresome and undesirable. In short, people desire a customized, unique product but they don't want to actively take part in the necessary decisions.

This in where an AI may play an interesting part - managing micro decisions with low potential effect but high cost in terms of decision fatigue. This is an idea that is already in development and we can see the results in the algorithms that are able to suggest music and movies that the user may enjoy. In a similar sense, an AI can build a model of behavior and likelihoods for its user, thus compensating for the infinitely large range of choice that the modern individual faces daily.

This may not immediately appear as a highly useful application. However, we have to consider that proliferation of options and an overall increase in choice has been a strong trend in the world history, so it is to be expected that with the technological possibilities of the future this will only increase. At the same time, there are signs that *having too much choice can be detrimental to satisfaction* [9]. Moreover, decision fatigue causes people to give up faster on both solvable and unsolvable tasks [9].

D. *Distributed Architectures*

Centralised Control Architectures are a today's standard. It is so because it seems to be the most natural manner of resource management. Result is seen as of a higher dimension than the summed parts, thus execution that leads to the result is to be done by agents coordinated by a hierarchical superior entity. This encapsulated approach limits the level of understanding necessary in the lower levels and maximizes control of the output for the decisional agent.

Needless to say this system may refer just as well to a network of computers as to an operational structure of a business or to the social organisation of a government. There are underlying benefits of such a model and for a large period of time it was viewed as the single possible manner of organisation. It was only after technological systems with certain degrees of agency were developed that researchers started to postulate on the possible advantages a decentralised architectures. The discussion can follow the following rationale:

- Processing Power: the simpler the executive agents, the more complex the central authority has to be for a given task,
- Robustness: the system is robust only to change in the lower levels - it is unstable for changes at the top of the hierarchy,
- Quality: the output of a system is only as good as the performance of the central authority and is only as favorable as its intention.

An example of the shift in perspective is illustrated with the evolution of Road Routing Software. Such an application could not be just an optimisation of distance, but also of time. This requires data acquisition on the status of traffic at immense scale - and this is effectively impossible for a centralised architecture. Instead, data is acquired by the end users themselves which practically makes them active agents in the system. The system is robust, it will function just as well even if any or as many (roughly speaking, of course) of the agents are removed from the network.

It is to be understood that distributed systems have not yet gain too much popularity, compared to centralised ones. People are way too unpredictable to expect unconditional and perpetual cooperation between them and computers have only recently had the processing power to gain the minimum necessary levels of autonomy. This is to change with the development of AI. Systems of collaborating agents, while not requiring exceptionally capable individuals are still able to produce complex *emergent behaviours* that can tackle otherwise impossible problems.

Fig. 5: Emergent Behaviour for a System of multiple Agents - Swarm of Birds

VII. Conclusion

Technology is a double-sided sword. In a discussion of utility both of these sides need to be addressed. It's easy to imagine how technology can improve one's life by replacing one's need of labour. But this was an argument for slavery, as well. It is not a fruitful discussion and it's a matter of imagination to propose humanoid machines that simply outperform humans at arbitrary tasks. Therefore, the article tried to focus on alternative ideas in the quest of identifying *something useful for the town*.

References

[1] Big data shows people's collective behavior follows strong periodic patterns. https://phys.org/news/2016-11-big-people-behavior-strong-periodic.html.

[2] Chaotic Systems. https://www.sciencedirect.com/topics/computer-science/chaotic-systems.

[3] Overview of Learning Styles. https://www.learning-styles-online.com/overview/.

[4] The Decline of Education. https://knowledgeone.ca/the-decline-of-education/.

[5] The World in 2000 as Predicted in 1910 - Retro Art. https://thisisthestoryof.wordpress.com/2012/06/18/the-world-in-2000-as-predicted-in-1910-retro-art/?fbclid=IwAR0XlG6JLLmDZOqfSr-2VjNQkpAmrCgaC2GMziRKGWS3fxC6OcqjwUdwISQ.

[6] Understanding Bayes' Theorem. https://towardsdatascience.com/understanding-bayes-theorem-7e31b8434d4b.

[7] J Steinhardt P. Christiano J Schulman D. Mane D. Amodei, C. Olah. Concrete problems in ai safety. 2016.

[8] A. Dafoe B. Zhang O. Evans K. Grace, J. Salvatier. When will ai exceed human performance? 2018.

[9] J. Twenge B. Schmeichel D. Tice J. Crocker K. Vohs, R. Baumeister. Running head: Self-regulation and choice. 2004.

[10] O. Khasawneh. Technophobia: Examining its hidden factors and defining it. 2018.

[11] George Orwell. *The Road to Wigan Pier*. 1937.

[12] C. Romm. Americans are more afraid of robots than death. 2015.