A Review of Analysis into the Ethics of Artificial Intelligence

Connor Goddard

Department of Computer Science, Aberystwyth University
Aberystwyth, Ceredigion, SY23 3DB
Email: clg11@aber.ac.uk

I. Introduction

As machines progress towards greater autonomy and intelligence, the ideas and arguments relating to their ethical implications for the future are becoming ever more complex, insightful and controversial.

In its simplest form, ethics defines a system that enables us as humans to describe, discuss and categorise the behaviour we exhibit as either morally right or wrong. While it is generally accepted that all animals possess the ability to appreciate *"conscious experiences"* such as pain from an injury or the taste of food, the ability to perceive and subsequently act on these experiences using some form of *judgement* has remained a capacity exclusive to humans - being the only species currently holding a generally higher-form of intelligence [1].

By lending our intelligence towards complex problem-solving, reasoning and communication, we have not only successfully guaranteed our own position at the top of the food chain, but now also seek to protect the survival of other species, fuelled by our capacity for complex emotion and the ability to demonstrate empathy and compassion towards others - even those not of our own kind.

Over the last half-century, computer technology has developed at an exceptional rate, with artificial intelligence (AI) playing an increasingly important role in today's 'computer-age' society [1]. Algorithms developed as a direct result of AI research are now found in applications spanning from healthcare and manufacturing, to big-data analysis: allowing companies across the world to gain a level of understanding and insights into their customers that would otherwise be impossible to acquire [2].

At the present time, AI technology has yet to begin demonstrating behaviour that is indicative of 'true' cognitive ability [1]. As a consequence, many of the ethical challenges associated with these current systems have already been addressed by pre-existing agreement on ethical standards: e.g. "*designing a robot arm to avoid crushing stray humans is actually no more morally fraught than designing a flame-retardant sofa*" [1]. However, as AI systems move closer towards becoming quite literally 'thinking' machines, the ethical issues raised as a consequence become ever more complex and in a number of cases, unprecedented.

This report provides an insight into the latest research within the current ethical arena for artificial intelligence, before subsequently critiquing a selection of the most recent research publications compiled from a range of academic sources.

II. Latest Research & Analysis

A. Employment, Economic & Socialistic Issues

In their latest consensus report published in August 2015, the Pew Research Centre present analysis into the debate over what potential impact advanced AI technology could have on future employment. In response to the question - *"Will automated AI systems and robotic devices have displaced more jobs than they have created by 2025?"* - the authors report that 48% of the experts canvassed believed humans would lose significant numbers of jobs to AI and robotic systems, with the other 52% giving the opposite response [3].

This near-equal divide in opinion indicates what is generally perceived to be: a high degree of uncertainty as to how well human life could adapt to a world where machines have since become the dominant work force [3].

Analysing responses from those on both sides of the question, the report highlights a series of recurring themes identified across the arguments presented by experts in support of their opinion.

Theme 1 - *Predicting the balance of jobs between man and machine:* In response to the proposed question, a number of experts put forward the case: that while future AI and robotic systems would, in all likelihood, be responsible for the loss of many *existing* job roles, these losses would in turn, be counteracted by the creation of new job areas relating to the design, implementation and maintenance of these very same machines.

While no-one claims to accurately predict what these new jobs will entail, a number of respondents looked to previous historical examples as a basis for predicting how expected advances in technology may affect future employment prospects. Making reference to previous waves of industry automation and technological advancement, a number of experts noted how following the introduction of new technology innovations, the overall number of jobs for human workers has always recovered [3]. John Markoff, senior science writer for the New York Times alluded to how 15 years ago, the idea of search engine optimisation had never even been considered, but yet today had now become a highly-competitive and profitable job sector [3].

Theme 2 - *A future guarantee of jobs for humans:* Within their responses, another selection of experts put forward their belief that the majority of jobs in the future would continue to rely on the *"unique characteristics"* that we possess as human beings (e.g. empathy, creativity, judgement and critical-thinking) [3].

While most deemed it highly likely that many tasks undertaken by humans today would eventually be delegated over to automated systems (e.g. switching to self-driving cars), they argued that in spite of this, human input would continue to play a vital - albeit perhaps more 'backseat' - role in the safe, and ultimately successful completion of day-to-day tasks. For example, in the case of self-driving cars, human intervention would continue to be paramount in situations where the machine is presented with a situation analogous to the well-known behavioural dilemma known as the *"trolley problem"* [4].

Theme 3 - *Income inequality and social imbalance:* A key concern for many warning against further AI employment, is the expectation that, in addition to driving the continued displacement of lower-paid "blue-collar' jobs (e.g. truck drivers and factory workers [5]), intelligent machines will inevitably begin to also penetrate industries traditionally reliant on higher-paid, specialist workers (i.e. 'white-collar' jobs).

Experts argue that by moving further into complex, yet routine job roles - such as accounting and law [3] - robots and AI would leave their human predecessors with no option but to find work at the lower end of the labour market, where jobs are both less financially rewarding and less secure. While it is acknowledged that some higher-paid roles would remain - predominantly complex and less-routine work, requiring highly specialist knowledge or creative flair (e.g. software developer, music composers etc.) - experts warn that this would only exacerbate the increasingly large income gap between the lower and upper social classes, and would for all intents and purposes, *"hollow out the middle class"* to leave only the highest and lowest paying roles left for human workers [3].

This scenario leads to further implications for future social and economic landscapes, with many respondents raising deep concerns as to what the reduction in overall income - caused by the massive shift in job class - will mean for the supply and demand for good and services, and the potential rise of an *"elitist"* approach to accessing 'vital' services such as advanced healthcare and legal support [3].

Writing in the science journal Nature, Russ Altman, professor of bioengineering, genetics, medicine and computer science at Stanford University, considers the importance of equally and fairly distributing new AI advancements in healthcare technology so as to ensure all patients, regardless of social class, are able to benefit [6].

In support of his comments, Professor Altman refers to the present day US healthcare system, reporting how currently people in work generally receive a different standard of care to those who are unemployed [6]. He goes on to argue in the future case of access to advanced AI-based healthcare, adopting a similar system to that of the US approach would be both *"unjust and unfair"*, before calling on governments to ensure that the benefits of AI technologies be fairly distributed to all who require them, regardless of an individual's social status [6].

In response to the concerns presented over dramatic pay inequality, other groups of experts make the case - that *"our social, legal and regulatory structures will minimise the impact on employment"* - by wishing to continue inciting economic growth, and preventing at all costs, the risk of social unrest. A good example presented for this argument considered; while businesses would naturally wish to employ advanced automated technology with the aim of reducing running costs, they would at the same time understand that by displacing large swathes of their workforce, this would in all likelihood, prevent large portions of their target consumer market from having money available to purchase their goods [3].

B. *Militarised AI & Automated Weapons*

With many experts firmly believing that lethal autonomous weapons systems (LAWS) [6] will become available within a matter of years, researchers and experts are beginning to turn their attention towards what the possibility of truly autonomous weapons could mean for the future of human safety and political stability.

Writing in an article for Nature [6], Professor Stuart Russell, one of the most recognisable figures in the AI research community, wrote about his fears and concerns over what the future of warfare may look like, when the decision over life and death is taken fully out of human hands.

Russell begins by commenting on how a number of military projects aiming to utilise advanced machine intelligence are very much established in present-day warfare preparations. He draws attention to two present-day DARPA programmes; Fast Lightweight Autonomy (FLA) and Collaborative Operations in Denied Environments (CODE) that both describe the extensive use of LAWS within their operations.

In a later point, the researcher raises questions over the jurisdiction that existing humanitarian laws would hold over the actions conducted by future autonomous weapons. While it was recognised that certain aspects of current international humanitarian law would in theory apply to the behaviour of autonomous weapons - in particular, criteria relating to military necessity, collateral damage and discrimination between combatants and civilians - Professor Russell argues that in many cases these laws rely on *"subjective judgements"* that current AI systems are unable to satisfy [6].

The last section of the article pays special attention to the need for a strongly regulated and carefully considered approach to developing LAWS over the coming years, so that society can help to prevent what a number of experts describe as a future *"AI arms race"* [6] [7]. In order to accomplish this, the author stresses the importance of strong working relationships and clear communication between the AI research community and those developing autonomous weapons. He pleads to the AI and robotics communities, arguing that they must take a definitive position on what constitutes acceptable use of their technology. Taking no position, the author argues, is the same as taking a position *"in favour of continued development and deployment"*.

C. Domain-Specific vs. General Artificial Intelligence

In their current state, AI-based systems have proven on multiple occasions to be capable of excelling above human capability within a number of finite domains [1]. While the systems behind such achievements possess incredible insight within their specific domain, they are restricted to fulfilling only one type of role (e.g. 'Deep Blue' possessed the capability to defeat world chess champion Gary Kasperov [8], but would have failed entirely with any task other than playing chess) [1].

Within their paper focussing on a variety of ethical implications spawning from advanced artificial intelligence, Bostrom and Yudlowsky discuss how the development of "Artificial General Intelligence" - machines capable of applying new or existing knowledge to multiple domains - introduces unique ethical challenges that will require an altogether radical new approach to ensuring such technology can remain with the bounds of existing ethical frameworks.

They propose that, in order for AI systems to be capable of conforming to more than one set of ethical standards, they must be able to understand and *extrapolate* the *"distant consequences of actions"*, without relying on explicit instructions as to what may constitute good or bad behaviour within the confines of any specific task. Put succinctly, the authors write *"...we require an AGI that thinks like a human engineer concerned about ethics, and not just a simple product of ethical engineering"*.

D. Super-intelligence

In 2003, Bostrom discussed how the creation of AI 'super-intelligence' - machines intellectually superior to even the very best human minds - would not just simply present another technological breakthrough, but would in fact become *"the most important invention ever made"*, by promising to revolutionise our scientific and technological endeavours through their ability to conduct research and design at a rate humans could only ever dream of matching [9].

In their 2011 paper, Bostrom and Yudlowsky are quick to highlight the immense potential for good such technology holds, noting that should super-intelligence be put to work tomorrow, it would address all or nearly-all of the main existential risks facing present-day society (e.g. climate change, global disease pandemics and catastrophic asteroid impacts [7]).

While the potential benefits of super-intelligent machines may surpass anything seen before, the risks associated with an intelligence beyond that of our own become equally unprecedented [1]. Public fears over the dangers linked to such machines have been well capitalised by films and books, depicting worlds where 'evil' robots strive to end the human race in order to become the dominant force. Although recognising such scenarios as works of fiction, Bostrom and Yudlowsky highlight the very real dangers of failing to ensure super-intelligent machines adhere to human ethical standards, warning *"it is crucial that [super-intelligence] be provided with human-friendly motivations"* [1]. Other experts have classified the risks arising from super-intelligence to now be so profound, they have now become recognised as one of twelve existential risks threatening the very survival of human life, sitting alongside the irreversible destruction of our climate and the potential fallout from nuclear warfare [7].

One of the biggest challenges associated with developing an ethically-considerate super-intelligence, stems from the idea that in reality, ethical standards are inherently *unstable* within society. As time passes, human attitudes to certain behaviour changes, which in turn causes us to reconsider our ideas about what is deemed ethically acceptable within society (e.g. the abolishment of slavery within the UK and US) [1]. This presents a fundamental issues for developers of AI technology, faced with the mammoth task of designing a means of communicating these changes to machines following a rigid, and now-outdated set of ethical standards.

The severity of this issue is not lost on the authors, who describe it as "*perhaps the ultimate challenge of machine ethics*" [1]. Considering one potential solution, they suggest that rather than focussing on teaching machines our ethical 'rules' as they stand *today*, the AI community must instead look to developing algorithms capable of, over time, *recognising* changes in ethical direction, and then *adapting* their behaviour accordingly [1].

While undoubtably an *"exceptional"* task to undertake, the paper argues that it is one *"we must meet"*, if future plans for highly-advanced AI technology are to place it in positions of greater trust, power and responsibility within society [1].

III. Critique of Research & Analysis

This section provides critique of the selected papers with respect to the general consensus towards the arguments raised, the level of evidence provided in aid of supporting their claims, the level of bias demonstrated towards their arguments and finally any additional factors deemed to provide particular strength or weakness to the research work.

A. Paper 1 - *"AI, Robotics & The Future of Jobs"*

As part of their introduction, the authors use percentage-based statistics to report a near-equal split in respondent opinion [3]. While this provides the reader with a clear indication as to the overall opinion weighting, upon further investigation of respondent numbers it appears that these percentage values may in fact be masking what transpire to be relatively small sample sizes.

When later explaining their approach to gathering responses, the authors report that out of 12,000 individuals invited to take part, less than 1,900 actually responded to their open-ended question relating to the impact of AI on the future of employment [3]. This figure, equating to less than 16% of the target sample size, represents only a small proportion of experts from within the AI, robotics and business communities. Therefore, it is important to realise that the equal distribution in expert opinion reported within the paper, does not necessarily describe the opinion trend of the *majority*, given that the responses from which these percentages are calculated, have in fact come from only a minority of expert representatives (i.e. <50% of the target sample size).

With regard to the focus group demographic, the decision was taken to pre-select those who would be invited to participate in the canvas questioning. While the authors make a concerted effort to inform readers at the beginning of the report, this naturally raises some trepidation over the viability of the findings published in the paper. This concern is compounded by the apparent lack of independent verification of survey responses. The reader must therefore place their trust with the authors in assuring that the published responses are genuine, and contain no untoward bias or sway in line with the needs or expectations of the authors.

Within the main body of the report, the authors provide analysis for each of the key arguments identified in the 'Key Findings' section at the beginning of the document. For each of the arguments discussed, the report publishes supporting quotations from the answers of at least two different respondents.

While the publication of respondent quotations provides context and supporting evidence towards the arguments identified, the paper fails to provide any kind of further statistical analysis, that may have potentially brought to light additional patterns in opinion and reasoning that would not otherwise be obvious through simple comparison and extrapolation of common arguments. By conducting additional qualitative analysis on respondent data, the authors would provide more weight towards their published results by allowing users to scrutinise their analysis methodologies and results before subsequently drawing their own conclusions over the viability of the report findings.

Overall this report provides readers with a detailed view as to the current state of how varying groups of experts believe AI will affect future employment prospects, and what implications these may have for future economies. It is recognised that whilst the published responses provide compelling reasoning and support to the arguments discussed, some doubt is raised over the extend to which these responses provide a truly representative picture of current expert opinion, and that more work could be done to conduct further objective statistical analysis of the qualitative data.

B. Paper 2 - "The Ethics of Advanced Artificial Intelligence"

In their paper, Bostrom and Yudkowsky present their predictions as to the main ethical issues set to arise following the future introduction of advanced artificial intelligence into mainstream society. In addition to discussing how and why these issues may come into existence, the authors propose their ideas as to how they may be addressed, although they take great care to highlight that these solutions are unlikely to be easy, and indeed in some cases may in fact be impossible to implement - at least with the technological insight that we hold presently.

Unlike many of the other literature examined for this investigation, Bostrom and Yudlowsky specifically focus their attention towards tackling the more 'fundamental' - and naturally more difficult - ethical issues relating to future AI technology (e.g. the idea of machine super-intelligence and the moral status of machines). As a consequence of predicting scenarios destined for some point in the future, the authors could only base their discussions, ideas and arguments purely upon speculative reasoning, rather than on any factual evidence. It is therefore impossible to gain sufficient evidence to either prove or disprove the author's claims, however the arguments that are presented do appear on face value to build on top of plausible ideas and concepts.

As the paper is no doubt led by personal opinion, one would assume that there is significant risk for the introduction of bias into the points the authors raise. However upon reading, the paper does appear to provide generally well-balanced arguments, with the authors appearing to pay careful consideration over the choice of language used, so to not infer that any specific points should be taken as direct fact by the reader. One considers whether the existing backgrounds of the two authors - that of existential risk in the case of Bostrom, and advanced AI research in the case of Yudlowsky [1] - may have contributed to ensuring that the paper remained balanced in its predictions over both the positive and negative ethical impacts that future AI technology may present.

With this said, one particular argument appears to incorrectly suggest that little research focus has being drawn to the development of machine learning algorithms capable of exhibiting robustness to manipulation [1]. Upon further investigation, one concludes that the authors have failed to recognise a number of previously published papers [10] [11] relating to research specifically focussing on improving the security and robustness of advanced machine learning algorithms (known as *"Adversial Machine Learning"*) [12].

Although at times portions of the paper's content could certainly be described as 'abstract' (e.g. when using terms such as *"Principle of Substrate Non-Discrimination"* in discussions over possible criterion for identifying a machine as a moral entity [1]), the overall considerations put forward by the authors show deep insight into what are without doubt, highly complex and difficult ethical challenges likely to face humanity in the not too distant future. While never suggesting that a relationship between continued AI development and existential consequence is certain, the authors do provide compelling reason to suggest that serious considerations over how such a situation could be avoided need to begin now, if we as a society wish to secure our continued dominance and survival for future generations to come.

C. Paper 3 - "Comment: Ethics of Artificial Intelligence"

Writing in the science journal Nature [6], renowned AI specialist Stuart Russell is invited to provide his own predictions as to what consequences the introduction of advanced AI-based weaponry may have for the future state of human warfare.

From examining the article, one gains an impression that a considerable proportion of the author's discussion is based upon an assumption that in the future, intelligent weapons would not just simply act upon instruction from human commanders, but would in all likelihood, begin acting upon their own tactical decisions over the selection and engagement of targets within an active war-zone environment.

Inspecting from the article one example in particular, Russell makes reference to a proposed DARPA project titled: *"Collaborative Operations in Denied Environments"*, whose aim the author describes is to develop *"…teams of autonomous vehicles carrying out all steps of strike mission - find, fix, track, target, engage, assess in situations in which enemy signal-jamming makes communication with a human commander impossible"*. While the official information material published by DARPA does appear to corroborate this description, the author fails to bring appropriate attention to the fact that all CODE-enabled aircraft would only ever engage targets aligning to *"established rules of engagement"* defined explicitly via human operators prior to beginning a mission [13]. By failing to include this information, it would not be unfeasible for a reader to become mis-guided over the extent to which autonomous weapons will be responsible for making critical tactical decisions within a conflict environment. This becomes particularly important when noting that the readership for the journal stands at "more than 3 million unique users every month". [14].

Despite these concerns, credit must be given to Prof. Russell's article for no doubt providing readers with an insightful and thought-provoking view into the both the local, and wider ethical implications surrounding autonomous warfare, as technology comes ever closer towards fighting battles entirely on our behalf.

IV. Conclusion

In examining the literature selected within this report, it has become abundantly clear that the issues relating to the ethics of advanced artificial intelligence are both highly complex and controversial.

While of course nobody can be certain about what ethical challenges lay ahead, this report has attempted to recognise some of the visionaries currently working in the hope of inciting the important discussions to begin now, so that we as humanity will have the opportunity needed to consider very carefully, the issues that could ultimately spell the future triumph or demise of human civilisation.

References

[1] N. Bostrom and E. Yudkowsky, *The ethics of artificial intelligence*. Cambridge University Press, United Kingdom, 2013.

[2] "2015 State of Artificial Intelligence & Big Data in the Enterprise," https://www.narrativescience.com/state-of-ai, Narrative Science, 2015.

[3] A. Smith and J. Anderson, "AI, Robotics, and the Future of Jobs," Aug. 2014.

[4] T. Jaipuria, "Self-driving cars and the Trolley problem," https://medium.com/@tanayj/self-driving-cars-and-the-trolley-problem-5363b86cb82d, May 2015.

[5] S. Santens, "Self-Driving Trucks Are Going to Hit Us Like a Human-Driven Truck," May 2015.

[6] S. Russell, S. Hauert, R. Altman, and M. Veloso, "Robotics: Ethics of artificial intelligence," *Nature*, May 2015. [Online]. Available: http://www.nature.com/news/robotics-ethics-of-artificial-intelligence-1.17611

[7] D. Pamlin and S. Armstrong, "Global Challenges - 12 Risks that threaten human civilisaiton," Feb. 2015.

[8] F.-H. Hsu, *Behind Deep Blue: Building the computer that defeated the world chess champion*. Princeton University Press, 2002.

[9] N. Bostrom, "Ethical issues in advanced artificial intelligence," *Science Fiction and Philosophy: From Time Travel to Superintelligence*, pp. 277–284, 2003.

[10] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?" in *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*. ACM, 2006, pp. 16–25.

[11] P. Laskov and R. Lippmann, "Machine learning in adversarial environments," *Machine learning*, vol. 81, no. 2, pp. 115–119, 2010.

[12] L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *Proceedings of the 4th ACM workshop on Security and artificial intelligence*. ACM, 2011, pp. 43–58.

[13] J.-C. Ledé, "Collaborative Operations in Denied Environment (CODE)," http://www.darpa.mil/program/collaborative-operations-in-denied-environment, U.S. Department of Defense, Aug. 2015.

[14] "Announcement: A new iPad app for Nature readers," http://www.nature.com/news/announcement-a-new-ipad-app-for-nature-readers-1.12002, Dec. 2012.