Fundamentals of Artificial Intelligence Assignment 1- Ethical, Emotional and Social issues of AI

Craig Heptinstall Crh13- 110005643
SEM6120
Institute of Computer Science
Aberystywth University

***Abstract*—The ethics of Artificial intelligence is an ever increasing debate amongst scientists and those that try to enforce new laws (e.g. privacy) or rights, with machines becoming more and more involved with the day-to-day practises and activities of people around the world. Though true AI (there is debate about that too) is something that may take many years to conceive, the ethics will always be an issue that will need addressing. So what is AI? Why are ethics important? What ethics are already considered today? Will AI be able to be created without the need for emotions?**

I. Introduction

Although this paper looks generally at ethics in Artificial intelligence, to understand the topic in more depth, ethics can be broken into two questions that can be asked for both humans and machines alike. Can the subject (person or machine) be ethical? And can the subject be effected by the ethics of others? The first of these questions looks at the responsibilities of AI, and whether AI should be used for such purposes as weaponry or health care. The second question looks at AI rights, laws that could one day effect how they work, and if AI can even be considered to be under the same social circumstances as a person. Can an AI have feelings? This is something the second and especially third paper concentrates on towards understanding how emotions can invoke good ethics. This paper uses sources from a number of journals and news articles to discuss current (less than five years) research, and provides a critique for each.

A. What is Artificial Intelligence?

Before answering questions about the ethics of AI, it is important to define to a certain extent what the term 'Artificial intelligence' means. To break this down further, understanding the word intelligence can help define the greater meaning. Alan Turing [1] breaks intelligence into five major components, all of which should be fulfilled in order to be classed as intelligent:

1) Learning- The simplest form of this is trial and error, with more complicated forms such as generalisation meaning the learner can perform better in situations not encountered before.
2) Reasoning- Using evidence from a set of given statements to deduct a conclusion.
3) Problem solving- Special and general-purpose methods exists, where the special means a method of solving the problem is tailor made, while the latter means the method can be applied to a larger pool of general problems.
4) Perception- To be able to process and analyse scenes into objects, features and relationships.
5) Language- To use a system of signs, or sounds to communicate or send information to others.

Knowing the general requirements of intelligence, Artificial Intelligence should allow machines perform operations or actions that require the intelligence listed above in humans.

B. Why Ethics are important

In a world where humans are becoming more and more dependent on machines, the need for AI is exponentially increasing. Due to this circumstance, ensuring that any machines or computers used that are safety critical, are dependable on (e.g. healthcare), or cases where sensible decisions are required should be ethically correct. This relates back to the reasoning component of the AI definition.

One example from the Cambridge Handbook of AI [2], could be that in the near future where for instance a machine that decides on mortgage decisions is found to be being unfairly handing out successful applications based on discriminative terms, who is to blame? Why is the machine coming to such results? It could be said that as long as humans are imperfect, so machines will be.

Alongside these kind of ethics, there is also the other side to be considered, the ethical rights of the machines themselves. Though this topic requires more imagination about the future, machines that have feelings should still be considered. For instance, if a machine was working in bad conditions, or being treated badly, should they have rights? This will also take into account the second point in this paper, whether AI with emotion could benefit or suffer with such a trait.

C. AI ethics being considered today, and tomorrow

Although laws and rules have been around for the misuse of technology for some time now, no concrete laws regarding the use of Artificial Intelligence are under use at this time,

meaning uses of AI for weaponry or wrongdoing is not currently monitored. Because at this time AI is not fully developed, and may have some time to go, creating laws that manage unethical behaviour might not fit well with future and more advanced AI.

Aside from laws which are maintained by the courts, AI and machines used for public purposes are still judged by society, and general census around misusing AI can affect how they are used. Drones for example, which are now becoming more and more popular with the general public have meant they are used for many uses, from capturing images to performing deliveries. There are however circumstances where privacy is breached, or accidents from using autonomous drones have taken place.

Alongside current examples where ethics have had to been taken into careful consideration, fictional ethics and rules have been created, which have then become a part of current research. The Author and professor Isaac Asimov created three basic 'laws of robotics' (made famous by the fictional story of I, Robot) in [4] 1942 which were:

1) A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2) A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Although fictional, the laws have been deeply discussed in current AI advances, such as in an edition of the IEEE Intelligent Systems [5], where the laws were modified to try reflect todays current AI.

Relating this back to current and future uses of intelligent machines, the three laws are not always a necessity to engineers, for instance military uses of autonomous vehicles designed to kill enemies would not have such laws built into their designs. Creating a standard applicable set of rules for future AI could be integral to ensuring that future technologies are morally and ethically correct.

II. Recent research into Ethics of AI

Because ethics is such a broad topic, a great amount of research is already available. This part of the paper considers three papers from the past five years, looking at ethical uses of AI, through to considering the ethics of social and emotional AI. With each, a discussion of the research undertaken and findings will be referenced.

A. The Ethics of Artificial Intelligence

A paper already mentioned briefly earlier, this piece of research from the Cambridge Handbook of Artificial intelligence [2] written by Nick Bostrom and Eliezer Yudowsky (2011) looks at the ethical use of AI for a number of scenarios, discussing moral status of machines, the transparency of AI workings, and the ideas of superintelligence, referencing various examples of AI in good and bad circumstances. As discussed earlier, if AI is to be used more often for day-to-day activities, the ethical decisions taken by the intelligence must be justified by the engineers, creators of the intelligence, and the intelligence itself. The paper spoken about here starts by mentioning the main reasons for the need of ethical considerations when creating AI. One such example as the AI using its own advanced intelligence for good rather than ill. It then goes onto talking about the bank mortgage example, explaining how it could prove nearly impossible to understand how or why the AI algorithm could be judging applicants unfairly.

The paper points out that if AI was working for purposes such as this, it would enforce the need for complete transparency of the way the AI algorithm works. This would ensure that the system could be trusted more, and that finding out the reasons for failing behaviour or bad decisions are found out sooner. One counter-argument that the paper does not mention is that if every AI algorithm was transparent, would that be safe for all uses? For example, where AI is used for controlling nuclear reactors, allowing humans to neglect and exploit the algorithms could result in dangerous circumstances.

This leads to another point made by the authors of the paper, who speak of the need for AI to learn, and basing as little knowledge as possible from information given by designers or engineers. The paper includes a good example of the case of Deep blue, where programmers sacrificed the ability to predict the AI's next move, in order to get optimal results. This was shown when the system beat the world champion, Garry Kasparov. Although this method of creating AI can produce better results, it can be linked back to transparency, and getting the balance of predictability correct for today's society.

Another safety critical point the paper raises is the design aspect of AI. AI should be able to for see consequences of actions. The author does point out issues with trying to do this though, such as that trying to find out consequences is none-local, meaning that consequences are more precise specifications, and though the system may be safe, the purpose of the system may never be fulfilled. Take one example, where in order to not lose a game of Tetris, a game-learning AI[8] simply paused the game to avoid losing when the board filled up. On one hand, the ethical behaviour is something that could be strange to humans, but to the AI, it has acted in a way it thinks sensible.

To define the moral rights of the machines themselves, the paper then goes onto discuss what it means to have moral status, starting with an example of a rock, and how humans can subject it to any treatment without concern for the rock. It then goes onto to define why humans are subject to moral rights, while it is 'widely agreed that current AI systems have no moral status'. The paper then goes onto explaining the basic requirements for an AI which would have moral status:

- Sentience- The machine would have to have the capacity to feel pain or suffer
- Sapience- Having a higher intelligence such as self-awareness

So in theory, backed up by this paper, an AI could have some form of moral status if it could feel pain. The paper expands here, stating that it is morally wrong to harm animals or other living things if it can feel pain. So if an AI could feel pain, this would be in effect affecting its moral rights. The following two papers in this article will go into further detail around pain, suffering and emotions within AI.

A part of the paper which deserves highlighting is the section about what would happen if a human was 'uploaded' to a machine, capturing all neurons and synaptic interconnections in the process. If the upload was successful, would the machine then be that person? And more importantly, would the AI then be sentient? If so, moral rights could then be in place for that AI. A lot or articles agree that once emotion and pain are in the AI after an upload, then it does have moral rights [9], while others argue that if a machine became obsolete and not required [10], it would be important that the machine did not have moral rights in order to dispose of it.

There are also other considerations within this paper, such as reproduction. Would they reproduce? The paper explains the possibilities that the AI could simply clone itself, and each clone creating more clones until they run out of resources. In this case, the machines would have to be ethical in the way they chose to either remove old AI's or manage the amount of children they have.

The final interesting but important point in the considerations of creating AI, is the speed at which this intelligence would operate. Due to the vast power of machines in today's world, they can calculate solutions to vast problems at thousands of times the speed of a human. In one example, if the AI was one thousand times faster at perceiving the world than humans, one second of our time could correspond to seventeen minutes of AI time. In all that extra time, the AI could develop much faster than humans, until their idea of ethics could be considerably different to humans'. The paper uses one example where an AI is created with a fixed moral code of Ancient Greece. Would the ethics seem ethical today?

The overall critique of this paper is a small one, the examples used and references cited throughout the examples give stable ground for recent ideas, and ones based on great ideas from earlier times. Both authors, who study in the field of 'future humanity' have written many previous papers in the field. There are some examples given though where the authors have used the term 'commonly proposed'. The paper appears very opinion based where instead some facts or references to others could have been provided, to allow their own opinions to be backed up. For example, in the section looking at whether animals have moral status, the authors can 'commonly' they are regarding as not having them due to not having sapience. By looking at various sources, such as the WDC research and charity [13], it can be seen that animals such as dolphins and whales can be considered to have social skills, cultures and some forms of communication.

Although this paper does have issue with representing some of the points it makes, it is generally well written and gives a good insight into today's (though in a number of years time the ethical opinions could have changed for humans') ethical problems when approaching the creation of AI.

B. Emotion, Artificial Intelligence, and Ethics

The second paper [6], titled as above and written by Kevin LaGrandeur looks more directly at emotional Artificial Intelligence, and how instantiating emotions in machines will allow them to act more ethically. The paper starts with using the term 'affective robotics', meaning the robots would have forms of human emotion, with the aim that they would have moral and ethical codes. The paper states that attempts to invoke emotions into AI has stretched back a decade. To expand further on this, the paper outlines three main motives for the need for emotion in AI:

- Facilitate believable human-robot interaction
- Provide feedback to users
- Reflect how the robot is affected and adapts to factors over time

Each of these requirements shows the need for AI to give as much feedback as possible. The paper confirms this concept by stating safety can be ensured when the machine gives more emotional and 'real' feedback.

An interesting statement that came without research or evidence was that 'by having more attractive robots by appearance, the human user would likely trust the robot more'. The author should have referenced or performed some research here to support their opinion though. The next highlight of the paper uses this opinion from the author to further state the need for emotions in AI to allow robots to be more attractive to end users. Statistics gained from IEEE [14] considering 'World robot population' shows that over ninety percent of robots currently are service robots rather than industrial ones, again helps the authors prove the need for robots general and emotional appearence to be attractive due to the vast use of machines in day to day use.

Otto Rosler, a German theorist hints at the idea of robots emitting emotions as well as having a friendly appearance in order to cause humans to develop a real bond with them. Some of his research included how a robot could in theory be brought up as a child-like being. This way, on first sight of the human, the AI could be automatically socially bonded. By having a bond like this, the AI could learn from its human parent, giving more chance for the AI to become better ethical. Of course, there could be a reverse effect if the human is not ethical towards the AI.

By living with robots, companionships could also be made, with which the paper provides another good example, LIREC (Living with robots and Interactive Companions project) [12]. This project aimed at designing robots to not

only fulfil their purpose, but to also interact with their users on a social and emotional manner. Take the Roomba example, a vacuum cleaner which manoeuvres the floor by hitting objects then turning direction. An attachment could be made with these quite easily in humans, because of the gratitude they receive.

Matthias Scheutz warns about cases such as this though, stating that attachments to AI could mean that humans could in fact be manipulated by the robots because of the vulnerability of trust and care for it. It was reported in an article by Scheutz that some people gave the Roomba day's off from work [11], strengthening evidence that humans can begin to care for robots. A final example that the paper uses to good effect in this circumstance is what would happen if a military person was reluctant to trade a bomb disposal robot for a newer and safer one, but had grown a bond with that robot? In the worst case, the bond could actually cause lives to be lost due to the old and obsolete robot in use. Scheutz shows in his own research that soldiers can become very devoted to these kinds of robots [11].

The authors speculate a lot about the future of emotions in the following section of the paper, going on to state that generally emotions between humans and AI raises some problems, though because of recent successes replicating cognitive features, some level of emotional contact is possible. Again, the same issue with missing out key research examples applies here. Without references to outside research, or statements from others, the author is only stating an opinion, and other opinions can counter this one. One such example counter argument comes from Eyal Reingold and Johnathan Nightingale from the university of Toronto [15] who outline the advantages of AI, saying that emotional AI will help machines 'learn about people, and the world in which they live'. The author fails to actually point out the suggested problems with emotional AI.

A final remark that needs mentioning as a highlight of the paper is where AI consciousness is talked about relating to military robots again. In an example taken from Ronald Arkin, part of the US Army, mentions that with the need for emotional AI with consciousness is very important, for instance where without a human operator available the robot would have to decide whether to decide to shoot at a target.

With consciousness, the AI might hesitate, or create reason for not shooting, while without a consciousness the AI would just be following straight procedures and could result in wrong-doings. Arkin, an author referenced numerous times in this paper, points out that an artificial affection function that provided forms of guilt, grief or remorse would motive the robot to review its decisions.

Although overall the author creates a good deal of examples to get their opinion across, some more research into the advantages and disadvantages of emotional AI could have been made. With this, then the opinion of the author could have been supported for when they pointed out the need for, but careful implementation of such feelings in an AI.

C. Considering Social and Emotional Artificial Intelligence

The third and final paper that was researched as part of this assignment takes a similar opion to that of the previous paper's on emotional Artificial Intelligence [7]. Instead, both social and emotional AI is looked into, considering what it would mean for a machine to have either types of intelligence. The paper explains both concepts in interesting ways, using examples from recent research. The authors, Marc Schroeder and Gary McKeown begin by explaining the need for both socially and emotionally intelligent machines.

Alan Turing's famous test as mentioned previously tests an AI with the question of 'can machines think?". Emotional and social intelligence are both highly important prerequisites for passing the test, for example attempting to fool a human into thinking they are speaking to another human would be very difficult if the machine responded with no emotional substance in its responses.

To understand why and how social and emotional intelligence would work in an AI, the paper defines what both terms mean first. The authors define social intelligence as the ability to interpret others mental behaviour, and to interact is social groups or close relationships. The paper highlights an important aspect of social intelligence, which is the ability to communicate with others. Rather than responding based off known information, replies should be sent from reading the others feelings, and situational context.

As for Emotional intelligence, this is described as the ability to "monitor one's and other's feelings and emotions". Although the paper fails to give an example here, one could be suggested of such that a person that has been through a very recent traumatic event can usually be picked up on easily, whereas a machine could struggle with this simply by text input.

In describing how humans perceive different objects or machines as social entities, the author uses the "tool" and "social" entity example. For instance, a simple object like a hammer would be a tool, because it has predictable behaviours. The more complex the object or machine, the more humans begin to think that object or machine has a mind. The paper explains that with unpredictability, comes the feeling that you are interacting with a social entity.

Following on from mentioning social machines, the paper then describes a modern day example with the aim to build a Sensitive Artificial Listener (SAL). This machine would focus more on real time social interactivity with a user rather than focusing on understanding context. The paper mentions the software to be used in the planned machine, from OpenSMILE to MARY. The first of these is an analysis tool for recognizing emotion in the user's voice and face, while MARY aims at sounding more natural when replying with speech.

Mentioned earlier in this assignment, because SAL must learn its way to becoming more natural to a human, the system is not currently running from its own data due to the time frame it has had. The paper mentions here that the project

is currently relying on data provided directly from voice and video recordings.

All of this of course then can be tied to the question, does emotion make the creation of fully fledged AI a step closer? The paper speculates how well SAL would perform in the Turing test, and concludes not very well due to the limited world knowledge, though contrasts this by stating other AI would not be able to address the criteria which SAL fulfils.

This paper concludes by questioning the practicalities of having emotion in AI, and why emotions should be present to allow for greater understanding of situations. The paper speculates that AI should be able to 'interpret a user's behaviour'. 'Predict how the user feels' and more importantly 'how it should think and behave in a given situation'. This could lead back to the example of and AI in charge of an autonomous gun, where if it did choose to kill, emotions could prevent it, by predicting future emotions of the humans around it, and even its own emotions.

III. Overall findings and conclusion

From researching and evaluating the papers used in this assignment, a good grasp of how ethics are an integral part of society to be considered in Artificial Intelligence, and how emotions are becoming a need in modern machines. By using examples provided in the papers critiqued, and external sources to compare and contrast research, an array of scenarios has been provided to show why ethics, emotion and is a prerequisite before true AI can be formed. Looking at the initial questions asked in the introductory stage of this report, ethics will become stronger and stronger issues in the future development of AI, with social and emotional skills a 'must' according to the previous paper's evaluated in this report.

References

[1] B.J. Copeland, *What is AI?*, www.alanturing.net/turing_archive/pages/referencearticles/whatisai.htmll, 2000

[2] N. Bostrom, E. Yudkowsky, *The Ethics of Artificial Intelligence*, Cambridge Handbook of Artificial Intelligence, 2011

[3] J.F. Weaver, *Robots Are People Too*, www.slate.com/articles/technology/future_tense/2014/07/ai_drones_ethics_and_laws_if_corporations_are_people_so_are_robots.single.html, 2014.

[4] I. Asimov, *Isaac Asimov's "Three Laws of Robotics"*, http://www.auburn.edu/~vestmon/robotics.html, 1942

[5] R. Murphy, D. Woods, *Beyond Asmiov: The three laws of responsible robotics*, http://www.computer.org/csdl/mags/ex/2009/04/mex2009040014-abs.html, 2009

[6] K. LaGrandeur, *Emotion, Artificial Intelligence, and Ethics*, New York Institute of Technology, 2011

[7] M. Schroeder, G. McKeown, *Considering Social and Emotiona lArticial Intelligence*, 2010

[8] I. Steadman, *This AI 'solves' Super Mario Bros. and other classic NES games*, *http://www.wired.co.uk/news/archive/2013-04/12/super-mario-solved*, 2013

[9] H. Edge, *When does Artificial Intelligence begin to have rights?*, *http://www.ethicsofthefuture.com/2010/09/at-what-point-does-artificial.html*, 2010

[10] L. McGovern, *Ethics of AI - Legal Rights*, *http://www.flickspin.com/en/artificial_intelligence/ethics_of_ai_legal_rights*, 2008

[11] M. Scheutz, *The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots*, 2010

[12] LIREC, *Exploring and designing our future robot companions*, *http://lirec.eu*, 2015

[13] Whales UK, Sentient and sapient whales and dolphins, *http://uk.whales.org/issues/sentient-and-sapient-whales-and-dolphins*, 2014

[14] E. Guizzo, World Robot Population Reaches 8.6 Million, *http://spectrum.ieee.org/automaton/robotics/industrial-robots/041410-world-robot-population*, 2009

[15] E. Reingold, J. Nightingale, Can Computers Possess Emotional Intelligence?,. http://psych.utoronto.ca/users/reingold/courses/ai/emotional.html, 1999