Solving the ethical issues of AI

AI has captured the fascination of society tracing back to the Ancient Greeks: Greek mythology depicts an automated human-like machine named Talos defending the Greek island of Crete. [1] However, the ethical issues of artificial intelligence only started to be seriously addressed in the 1940s, with the release of Isaac Asimov’s short story “Runaround”. Here, the main character states the “Three Laws of Robotics” [2], which are:

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 the orders given it by human beings except when 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 Laws.

The rules laid out here are rather ambiguous; B. Hibbard in his paper “Ethical Artificial Intelligence” [3] provides a situation that conflicts these laws – his example situation being “An AI police officer watching a hitman aims a gun at a victim” which would necessitate, for example, the police officer to fire a gun at the hitman to save the victim’s life, which conflicts with the First Law stated above.

Thus, a framework to define how such artificial intelligence would behave in an ethical manner (and even make some moral improvements) is needed; the other factors this essay would discuss (mainly with the help of N. Bostrom and E. Yudkowsky’s “The Ethics of Artificial Intelligence” [4]) are transparency to inspection and predictability of artificial intelligence.

Transparency to inspection

Engineers should, when developing artificial intelligence, enable it to be transparent to inspection. [4] For artificial intelligence to be transparent to inspection, a programmer should be able to understand at least how an algorithm would decide the artificial intelligence’s actions.

Bostrom and Yudkowsky’s paper gives an example of how this is important, using a machine that recommends mortgage applications for approval. [4] Should the machine discriminate against people of a certain type, the paper argues that if the machine was not transparent to inspection, there would be no way to find out why or how it is doing this.

In addition, A. Theodorou et al. in the document “Why is my robot behaving like that?” [5] emphasizes three points that dictate transparency to inspection: to allow an assessment of reliability; to expose unexpected behavior; and, to expose decision making. The document takes this further by implementing what a transparent system should be, which includes its type, purpose and the people using the system – while emphasizing that for different roles and users, the system should give out information readable to the latter. [5] While the document does not specifically mention artificial intelligence as a separate topic, the principles of a transparent system could be easily transferred to engineers developing artificial intelligence.

Therefore, when developing new technologies such as AI and machine learning, the engineers and programmers involved should ideally not lose track of why and how the AI performs its decision-making process and should strive to add to the AI some framework to protect or at least inform the user about unexpected behaviors that may come out.

Predictability of AI

While AI has proven to be more intelligent than humans in specific tasks (e.g. Deep Blue's defeat of Kasparov in the world championship of chess [4]), most current artificial intelligence are not general. However, with the advancement of technology and the design of more complex artificial intelligence, the predictability of these comes into play.

Bostrom and Yudkowsky argue that handling an artificial intelligence, which is general and performs tasks across many contexts are complex; identifying the safety issues and predicting the behavior of such intelligence is considered difficult [4]. It emphasizes the need for an AI to act safely through unknown situations, extrapolating consequences based on these situations, and essentially thinking ethically just like a human engineer would.

Hibbard's paper suggests that while determining the responses of the artificial intelligence, tests should be performed in a simulated environment using a 'decision support system' that would explore the intentions of the artificial intelligence learning in the environment – with the simulations performed without human interference. [3] However, Hibbard also promotes a 'stochastic' process [3], using a random probability distribution, which would serve to reduce its predictability on specific actions (the probability distribution could still be analysed statistically); this would serve as a defence against other artificial intelligence or people seeking to manipulate the artificial intelligence that is currently being built.

Overall, the predictability of artificial intelligence is an important factor in designing one in the first place, especially when general AI is made to perform large-scale tasks across wildly different situations. However, while an AI that is obscure in the manner it performs its actions is undesirable, engineers should consider the other side too – an AI would have to have a certain unpredictability that, if nothing else, would deter manipulation of such an AI for a malicious purpose.

AI ethical thinking

Arguably, the most important aspect of ethics in AI is the framework on how the artificial intelligence would think ethically and consider the consequences of its actions – in essence, how to encapsulate human values and recognize their development through time in the future. This is especially true for superintelligence, where the issue of ethics could mean the difference between prosperity or destruction.

Bostrom and Yudkowsky state that for such a system to think ethically, it would need to be responsive to changes in ethics through time, and decide which ones are a sign of progress – giving the example of the comparison of Ancient Greece to modern society using slavery. [4] Here, the authors fear the creation of an ethically 'stable' system which would be resistant to change in human values, and yet they do not want a system whose ethics are determined at random. They argue that to understand how to create a system that behaves ethically, it would have to "comprehend the structure of ethical questions" [4] in a way that would consider the ethical progress that has not even been conceived yet.

Hibbard does suggest a statistical solution to enable an AI to have a semblance of behaving ethically; this forms the main argument of his paper. For example, he highlights the issue of people around the world having different human values that they abide by – thus making an artificial intelligence’s ethical framework complex. He argues that to tackle this problem, human values should not be expressed to an AI as a set of rules, but learned by using statistical algorithms. [3] However, he does concede the point that such a system would naturally be intrusive (which conflicts with privacy) and that relying on a general population carries its risks, using the rise of the Nazi Party through a democratic populace as an example [3].

Overall, enabling an artificial intelligence to act in an ethical manner is a process with huge complexity; the imbuement of human values into the artificial intelligence’s actions would almost certainly give it moral status, which could ease the ethical confusion of some advanced projects (e.g. where the responsibility lies after a fatal accident involving a self-driving car). However, such an undertaking is itself complicated and would require self-learning, which holds its own risks. Finally, an artificial intelligence, to be truly ethical, would need to (at the least) be open to ethical change and will most likely need to consider what parts of the change are beneficial.

For engineers to address the ethical concerns stemming from creating artificial intelligence and using machine learning, they should:

Ensure transparency to inspection by considering the end-users of such a machine, and provide safeguards against any unexpected behavior that is quickly readable to a person using it. They should use algorithms that offer more predictability and could be analyzed by at least a skilled programmer, even if this sacrifices the efficiency of the machine learning of its environment – this would reduce the chance of its intentions being obscure.

Consider the AI’s predictability; testing it in a different, simulated environment would allow the observation of what the AI would do, although not necessarily in an environment that models the real world. Predictability is somewhat linked with transparency to inspection in that engineers could track the intentions of a predictable artificial intelligence. However, to make the artificial intelligence resilient against unwanted changes, it is important for a random element to be added to the AI’s learning algorithm too.

Make efforts to study what underpins the ethics and different human values that modern society has, and start considering how an AI would be capable of continuing ethical progress (instead of simply looking at this progress as an instability).