Why a Special Issue on Machine Ethics

By H. JOEL TRUSSELL
Editor-in-Chief

In early 2019, the PROCEEDINGS OF THE IEEE will publish a Special Issue on Building Ethical Autonomous Systems. This topic is a departure from the usual reviews and surveys of highly technical and mathematical areas on the cutting edge of electrical and computer engineering research. You may have to search for the equations in this issue, but the logic can be challenging. The ethical questions that are addressed, while framed as dealing with autonomous systems, are important and applicable to dealing with all research areas under the IEEE umbrella. The methods of addressing these questions draw mainly from artificial intelligence and machine learning.

The decision to do this issue was not without problems. Some on the editorial board worried that it was not technical enough. Others thought that promoting the societal virtue of ethics was an appropriate topic for the PROCEEDINGS OF THE IEEE. We noted that the IEEE Code of Ethics was just recently revised to include specific mentions of sustainability and intelligent systems. Aside from the virtues of discussing ethics, the technical problems of designing ethical machines are extremely difficult and the area is in its infancy. There are only a few labs that are actively writing software to address this problem. In order for this effort to succeed, or at least to determine if success is possible, more effort, more ideas, and more innovation are needed. We hope that this issue inspires more people to get involved in this task.

My first thoughts on the possible success of this effort were pretty negative. There are many factors that one can cite that indicate severe problems. Ethics, itself, is not universally well-defined. Ethics may be a cultural or professional set of rules. Naïve subjects and ethicists are not in unanimous agreement on the ethical action for a specific situation. A classic example is the runaway trolley dilemma where one person must die to save five others. How can researchers decide what is the ethical action for a machine when humans cannot agree among themselves?

One answer is that it is not possible, that the best we can do is design a machine that is ethically aligned with humans, that is, behaves as an ethical human would. This might be measured by an ethical Turing test. Alan Turing proposed a test of artificial intelligence that if a subject could not tell the difference between the answers of a computer and the answers of a human that the machine was behaving in an intelligent way. Likewise, if a machine is asked ethical questions about various situations and gives answers that are indistinguishable from an ethical human, it is behaving ethically. In such a test, we would not expect the human and the machine to give exactly the same answers every time. But they would agree sufficiently often to be judged equivalent.

A related problem occurs when we consider real-world

Digital Object Identifier 10.1109/JPROC.2018.2868336

scenarios. The runaway trolley is an ideal fictional and unrealistic case. In the real world, people have to deal with uncertainty. Consider a case of an automated vehicle (AV) encountering an accident situation in traffic. A car ahead stops suddenly for some reason, for example its driver thinks that an object suddenly appeared in its path. While the automated vehicle should be following at a distance that would allow safe stopping, its brakes do not respond as expected. The AV must now decide on alternative actions. There are many possibilities, including: continue applying the brakes with the possibility of hitting the car ahead, veering left into an oncoming lane with the possibility of a head-on collision if there is a vehicle approaching in that direction, veering right with the possibility of leaving the road and going into an unknown terrain. The costs of the options are not clear: A human in such a situation might choose any of the ones mentioned or some novel alternative, but would not be held to a specific standard for his/her action. In such a case, would we expect the AV to perform with this latitude of variation? Would its actions be evaluated by a stricter standard that would be influenced by the harm to humans done by its decisions?

From my experience as a signal processer, I am familiar with probabilistic uncertainty. The information from sensors is always subject to noise and thus uncertainty. I know that every estimation or classification system produces both false positives and false negatives. While I teach this simple concept, I can never give the students a solid rule for determining the relative cost of the two errors. Yet, actual numbers must be used for an operational system in the field. Does the choice of the value of the relative costs reflect an ethical decision of the designer or of the designed system? If it reflects on the design, can the machine be considered ethical?

It is clear that ethics is very situational. Thus, it would appear that ethical machines, as humans, would need to be adaptive. The natural question then becomes, "What are the parameters necessary for adaptation?" This, in turn, depends on the exact situation. It is unlikely that machines will be able to handle a wide variety of situations in the near future. The research will start with relatively restricted cases like those used in the papers in the special issue.

The problems that are being addressed now may seem unrealistically simple, but they are the ones that will allow us to learn how to address the more complicated ones in the future. Artificial intelligence has come a long way from Minsky and Papert's perceptron of the 1960s [1]. There was the initial promise and the failure to deliver on that promise. With the advent of newer, faster computers with memory that most of us of the older generation never thought possible, artificial intelligence and machine learning have solved many problems that we thought would take many more decades to approach a reasonable engineering solution. We have quite acceptable voice recognition and machine translation of languages. While machine translators are not capable of handling philosophical concepts, they allow tourists to navigate through foreign countries without the frustrations of travel in the mid-20th century. Self-driving cars are now being used in limited environments and are likely to be on the public roads in significant numbers within a decade or so.

It is hoped that this issue of the PROCEEDINGS will start to lay the foundation for the design of ethical machines. Studying how to make machines ethical has some distinct advantages over dealing with humans. Machines provide us with a well-defined and controlled environment in which to conduct experiments. We can determine the complexity of the situation and increase that complexity as we learn how to deal with simpler cases.

Machines, as mentioned previously, are subject to uncertainty from noise in the sensors that they use to interpret the environment. Once the physical parameters are estimated, a defined algorithm computes the action that is to be taken. This result will be consistent for the same situation or, at least, the same estimated parameters. In [2], Anderson *et al.* use an example of a medical care robot. The robot checks on the patient for apparent health (frequent movement), checks the time for periodic medication, reminds the patient to take the medicine, and reports on the patient's behavior. The ethical requirements are simple: maximizing the health of the patient by monitoring physical movement and medication. Since it is also ethical to let the patient maintain his autonomy, the robot does not force him to take his medicine, but reports the event to a physician. The rules for the ethical actions are adaptive in that they depend on the values of machine-measured parameters that reflect the patient's current state.

In this case, the robot will behave the same for any patient, any doctor, and any social environment. A human caretaker would be subject to many emotional, social, financial, and physical conditions that might change his behavior. For example, if the patient were quite rude to the caretaker when he reminded the patient to take his medicine, and the caretaker's company had just cut staff salaries because of financial problems, he might feel justified in skipping the notification of the doctor. He would do it later, if the old buzzard refused it again. In a minor change in the situation, the caretaker's response would be quite different if the patient were a loving grandmother instead of an unknown curmudgeon.

We might think it advantageous that the robot makes ethical decisions based on physical facts. The robot is not affected by emotions, such as anger, friendship, and stress. The robot does not get hungry, a condition that is known to affect judgment. The robot cannot be bribed or threatened with losing its job. However, we might ask if there are cases where we want such intangible conditions to be considered. Would we consider our loving grandmother

patient and the unknown curmudgeon patient to be equivalent lives? If we design a machine accordingly, will we regret it later? In recent surveys, people preferred automated vehicles that favored saving them as riders over pedestrians or passengers in another vehicle.

The paper by Ema *et al.* [3] gives a good example of the complexities of defining appropriate ethics, both for the AI system used "to detect obscene expressions in order to eliminate information that might be 'harmful' to young people by filtering the text," and for the researchers who develop the system and publish the method. The authors use the example of a research paper that describes how to detect offending material used in stories contributed by members of a community on the internet. While the community was open, it was restricted to adults. The problem arose when the authors of the research paper identified the stories used to train their method, along with the authors' names and links to the online stories. Because of the nature of the stories, this caused some distress among the story writers, several of whom removed their stories from the web. To evaluate the ethical implications brought up by the situation, a panel was formed that included both humanities researchers and science and engineering researchers. As expected, the two groups demonstrated significantly different views.

The scientists and engineers tended to believe that the information about the stories used to train the AI filter was needed to enhance the value of the research and make a significant contribution to the technical field. This should hit home with the readers of the PROCEEDINGS, who are mostly "hard core" engineers and are data driven. There is a movement within the IEEE to promote reproducible research, which includes open access to data and computer programs.

The social scientists emphasized that the authors' contributions were submitted with the intent of publication to a limited community and the identity of the authors of the examples should be protected. Some in this group questioned whether a decision of "harmful judgement" should be left to the machines, which would seem to negate having machines making ethical judgments at all. In this case, both groups were trying to be ethical but their values were not aligned. It remains to be seen how such problems can be effectively handled in designing ethical machines.

As an aside, while the IEEE code was updated to include sustainability and intelligent systems, it does not explicitly address the problem of protecting data confidentiality. Perhaps we should consider an addition related to this. The paper by Spiekermann *et al.* [4] discusses the problems engineers have dealing with designing privacy and security into the systems. The authors conclude that too few engineers take the responsibility for insuring privacy and organizations give them too little time and too few resources to address the problems. The IEEE is currently preparing a position statement supporting the discipline of Privacy Engineering and co-sponsored a conference on privacy engineering. However, to my knowledge there is no explicit discussion of the ethics of privacy.

It may seem that designing ethical machines is impossible. Perhaps we need to consider the practical problem of designing systems that are as ethical as most humans, or as ethical as the ideals that we would use as examples. Finding a human example for ethical behavior appears impossible. But should we not dream the impossible dream? Should we not work toward the goal? This is the encouragement that I hope the Special Issue on Ethical Considerations in the Design of Intelligent Autonomous System will achieve.

In a final analogy, current polls show that over 60% of those surveyed would not ride in an autonomous car [5]. My guess is that within ten years that figure will decrease dramatically, as the technology improves and the statistics show that autonomous vehicles are actually safer than human-driven cars. It will take a longer time for us to develop trust in an ethical machine, but we need to start the effort now in order to achieve that same level of performance in other autonomous systems in the future. At least that is why we are publishing a special issue at this time. ■

REFERENCES

[1] M. Minsky and S. A. Papert, *Perceptrons: An Introduction to Computational Geometry*, 2nd ed. Cambridge, MA, USA: MIT Press, 1972.

[2] M. Anderson, S. Anderson, and V. Berenz, "A value-driven eldercare robot: Virtual and physical instantiations of a case-supported principle-based behavior paradigm," *Proc. IEEE*, to be published.

[3] A. Ema *et al.*,"Clarifying privacy, property, and power: Case study on value conflict of a fan fiction research paper," *Proc. IEEE*, to be published.

[4] J. Korunovska, S. Spiekermann, and M. Langheinrich, "Inside the organization: Why privacy and security engineering is a challenge for engineers," *Proc. IEEE*, to be published.

[5] [Online]. Available: https://www.brookings.edu/blog/techtank/2018/07/23/brookings-survey-finds-only-21-percent-willing-to-ridein-a-self-driving-car/