NOTE FOR NATIONAL DEFENCE:
Canadian Design Concepts in Autonomous Systems

Authors: S. Shahkar[1] and K. Khorasani[2]

[1] Graduate Student, Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada

[2] Professor, Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada

SUMMARY

- Robotics and autonomous systems ethics known as "roboethics" is concerned with policies and regulations that guarantee ethical behaviour of robots.
- The development of AI systems must always be responsible and advance towards optimal sustainability.

CONTEXT

- Potential harms that may yield unethical behaviour in robots might be caused by AI system developers as listed in [2].
- AI system designers may willingly or unwillingly influence the decision-making process of the AI through architecture design, insufficient or non-representative samples that are used for training, and/or the algorithm that is used for making decisions.
- Certain situations may arise where individuals who are interacting with robots experience unsatisfactory outcomes. "Take as an example a case of injuries, or a negative consequence such an accountability gap, which may harm the autonomy and violate the rights of the affected individuals". In such cases there should be a human(s) accountable and responsible for negative consequences.
- In some cases, machine learning models may generate their results by operating on high-dimensional correlations that are beyond the interpretive capabilities of human reasoning. In applications where the processed data could harbor traces of discrimination, bias, inequity, or unfairness, the lack of clarity of the model may be deeply problematic.

- Artificial intelligence is developed based on large clusters of data that may include personal data. Very often such private information might be deployed without consent of the owner which can invade the privacy of individuals.

- Irresponsible data management, negligent design production processes, or questionable deployment practices may result in implementation and distribution of AI systems that produce unreliable, unsafe, or poor-quality outcomes. Such outcomes can also undermine public trust in the responsible use of beneficial AI technologies to the society.

PARTICULAR GOALS

- "AI ethics is a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in their development and use of AI technologies" [3].

- The following provides a guideline for an ethical AI project and is referred to as "Ethical Platform for Responsible Delivery of AI":

- Ethical Permissibility: It is required to ensure that the AI project is considering the impacts of the outcomes on the wellbeing of affected individuals and communities.

- Non-discrimination: Evaluating the potential to have discriminatory effects by carefully understanding biases in deployment of data and algorithms.

- Worthy of Public Trust: Ensuring safety, accuracy, reliability, security, and robustness of the products to the possible extent.

- Justifiability: Transparency of the process, and the transparency and interpretability of its decisions and behaviours.

FUTURE DIRECTIONS

- Considerations of the moral scope of ethical impacts of the AI project and the criteria to evaluate its ethical permissibility is based on the following: Respect, Connect, Care and Protect.

- Need to ensure non-biased, non-discriminatory, fair and publicly trustable AI outcomes.

- Setup transparent design and implementation that ensure justifiability and the resulting products of AI.

REFERENCES

[1] Turkish Informatics Foundation, Ethics of AI: Benefits and Risks of Artificial Intelligence Systems, August 2020.

[2] David Leslie, The Alan Turing Institute of London, https://www.turing.ac.uk.

[3] David Leslie, The Alan Turing Institute, Understanding Artificial Intelligence Ethics and Safety, 2019.