Module 4: A Very Brief Introduction to Ethics in AI

AI ethics is an enormous topic, at least as large as AI itself. You could take an entire university-level course on the subject. In the time we have allotted for this content, we will barely scratch the surface of the topic. Still, it is vitally important when working with AI tools that you keep ethics at the forefront of your mind. At some level, all AI tools represent offloading human decision making to a machine, which then in turn affects other humans.

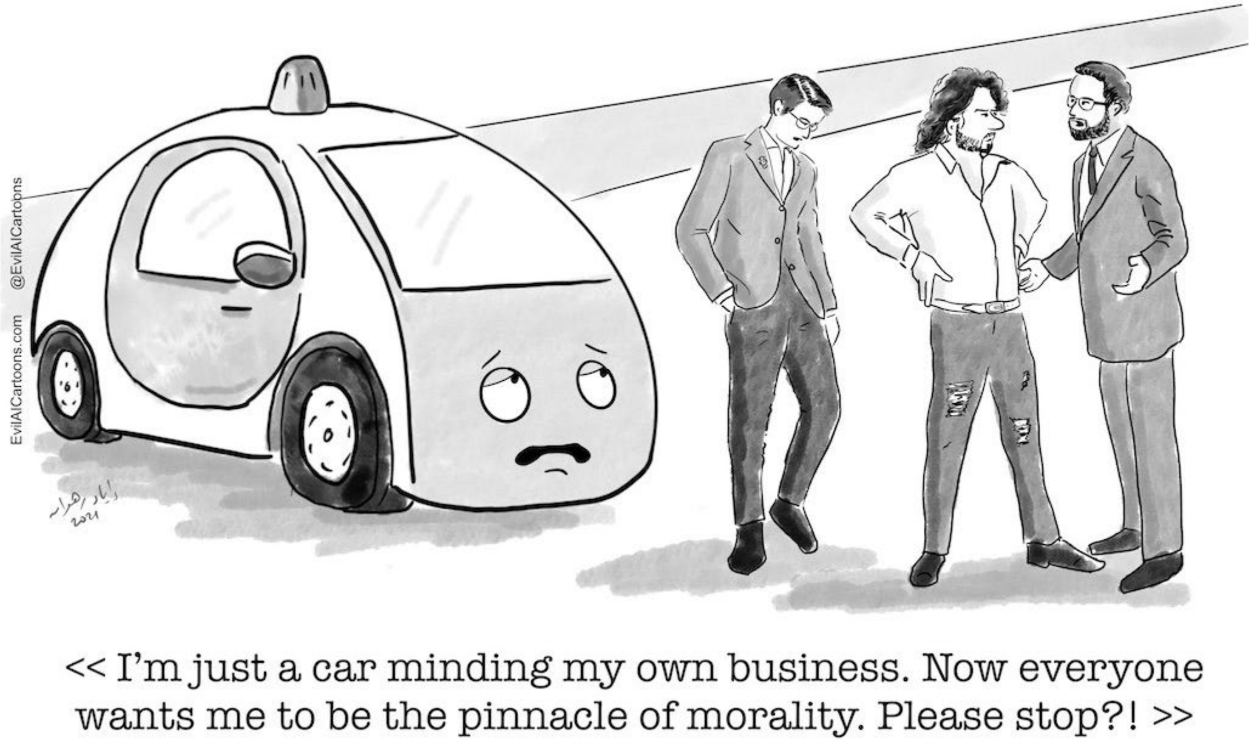
EvilAICartoons.com @EvilAICartopns
<< I'm just a car minding my own business. Now everyone wants me to be the pinnacle of morality. Please stop?! >>

Figure 1: Evil AI Cartoon. Credit - EvilAICartoons.com [1]

Module 4 Objectives:

By the end of this module, you will be able to:

1. Recall some ethical issues related to developing and using AI applications.
2. Identify key federal or departmental regulations and guidance for the responsible and ethical use of AI.
3. Recognize what the Rome Call for AI Ethics is and its impact on AI development.
4. Explain the importance of responsible and ethical use of AI applications.

Federal and Departmental AI Regulations and Guidance

On December 3, 2020, President Joe Biden signed Executive Order (EO) 13960, *Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government*. This EO outlines nine principles to guide the use of AI in government to ensure that such uses of AI are consistent with our Nation's values and are beneficial to the public. These principles include...

> **(a) Lawful and respectful of our Nation's values.** Agencies shall design, develop, acquire, and use AI in a manner that exhibits due respect for our Nation's values and is consistent with the Constitution and all other applicable laws and policies, including those addressing privacy, civil rights, and civil liberties.
>
> **(b) Purposeful and performance-driven**. Agencies shall seek opportunities for designing, developing, acquiring, and using AI, where the benefits of doing so significantly outweigh the risks, and the risks can be assessed and managed.
>
> **(c) Accurate, reliable, and effective**. Agencies shall ensure that their application of AI is consistent with the use cases for which that AI was trained, and such use is accurate, reliable, and effective.
>
> **(d) Safe, secure, and resilient**. Agencies shall ensure the safety, security, and resiliency of their AI applications, including resilience when confronted with systematic vulnerabilities, adversarial manipulation, and other malicious exploitation.
>
> **(e) Understandable**. Agencies shall ensure that the operations and outcomes of their AI applications are sufficiently understandable by subject matter experts, users, and others, as appropriate.
>
> **(f) Responsible and traceable**. Agencies shall ensure that human roles and responsibilities are clearly defined, understood, and appropriately assigned for the design, development, acquisition, and use of AI. Agencies shall ensure that AI is used in a manner consistent with these Principles and the purposes for which each use of AI is intended. The design, development, acquisition, and use of AI, as well as relevant inputs and outputs of particular AI applications, should be well documented and traceable, as appropriate and to the extent practicable.
>
> **(g) Regularly monitored**. Agencies shall ensure that their AI applications are regularly tested against these Principles. Mechanisms should be maintained to supersede, disengage, or deactivate existing applications of AI that demonstrate performance or outcomes that are inconsistent with their intended use or this order.
>
> **(h) Transparent**. Agencies shall be transparent in disclosing relevant information regarding their use of AI to appropriate stakeholders, including the Congress and the public, to the extent practicable and in accordance with applicable laws and policies, including with respect to the protection of privacy and of sensitive law enforcement, national security, and other protected information.
>
> **(i) Accountable**. Agencies shall be accountable for implementing and enforcing appropriate safeguards for the proper use and functioning of their applications of AI, and shall monitor,

> audit, and document compliance with those safeguards. Agencies shall provide appropriate training to all agency personnel responsible for the design, development, acquisition, and use of AI.

On October 4, 2022, the White House Office of Science and Technology Policy published *Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People*. This document identifies “five principles that should guide the design, use, and deployment of automated systems to protect the American public in the age of artificial intelligence. The Blueprint for an AI Bill of Rights is a guide for a society that protects all people from these [AI-related] threats—and uses technologies in ways that reinforce our highest values." These five principles are:

Safe and Effective Systems: You should be protected from unsafe or ineffective systems.

Algorithmic Discrimination Protections: You should not face discrimination by algorithms and systems should be used and designed in an equitable way.

Data Privacy: You should be protected from abusive data practices via built-in protections, and you should have agency over how data about you is used.

Notice and Explanation: You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you.

Human Alternatives, Consideration, and Fallback: You should be able to opt out, where appropriate, and have access to a person who can quickly consider and remedy problems you encounter.

On October 16, 2023, USDA’s Office of the Chief Information Officer released a new memo providing interim guidance for the use of generative AI models and technologies by USDA employees. This guidance is intended to address concerns about misinformation, privacy protection, and potential misuse, among other issues. The memo establishes a review board for uses of generative AI but includes a broad exemption for scientific research use cases. This is an evolving issue and further guidance is expected in the future.

On October 30, 2023, President Biden signed a new executive order, *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. This wide-ranging EO includes many directives relevant to AI ethics. In particular, it outlines eight guiding principles and priorities for the development and use of AI. It also requires the establishment of new guidelines and best practices to ensure: a) the safety and security of AI technology; b) that the use of AI advances equity and civil rights;

and c) that the development of AI technologies does not infringe upon individual privacy. (Please note that this EO addresses many other topics, too - we are only including the details most relevant to AI ethics here.)

The Rome Call for AI Ethics

The Rome Call for AI Ethics [3] is a document that was signed by a group of organizations in February 2020. The signatories include the Pontifical Academy for Life, Microsoft, IBM, FAO (Food and Agriculture Organization of the United Nations), the Italian Ministry of Innovation, and several universities.

The Rome Call for AI Ethics outlines six principles for the ethical development and use of AI. These principles are:

1. **Transparency:** AI systems must be understandable to all.
2. **Inclusion:** AI systems must not discriminate against anyone because every human being has equal dignity.
3. **Responsibility:** There must always be someone who takes responsibility for what a machine does.
4. **Impartiality:** AI systems must not follow or create biases.
5. **Reliability:** AI must be reliable.
6. **Security and Privacy:** These systems must be secure and respect the privacy of users.

The Rome Call for AI Ethics is a significant document because it is one of the first high-level calls for an ethical approach to AI. The signatories to the call represent a wide range of stakeholders, including religious organizations, technology companies, and governments. This broad support for the call is a sign that there is a growing consensus that AI needs to be developed and used in an ethical way.

We should point out a key criticism for this framework: Its vague. Very vague. The Rome Call for AI Ethics gives us a list of things to consider but lacks firm guidance on implementation. *How* you go about following these principles will require decisions made at every level of the AI application development process.

For each principle of the AI Bill of Rights and the Rome Call listed above, consider the following:

1. Imagine you are a Supervillain. What would you do to take advantage of AI, or how would you use AI to do evil if this principle were *not* in place? Feel free to be creative!
2. Now imagine you are just a terribly busy, very tired data scientist and you have a deadline looming. What would some potential consequences be if you rolled out an AI-enabled project without applying this principle?
3. What are some safeguards you can think of to ensure that this principle is implemented?

AI as Moral Agents

To make the issue a little more concrete, consider the case of driverless cars. Driverless cars have already been executing decisions with ethical implications for years now, and disagreements over their moral frameworks are part of why the technology has stalled in its development.

1. Watch the TED-Ed video "The Ethical Dilemma of Self-Driving Cars" by Patrick Lin: https://youtu.be/ixIoDYVfKA0 [5]
2. Recalling the principles listed in the EO and AI Bill of Rights, consider each of the following:
   a. **Transparent**
      i. How can we make the decision-making process of a self-driving car understandable to the passengers and other road users?
      ii. Should there be a way for users to see the process the AI uses to make decisions in real-time? How could this be implemented?
   b. **Algorithmic Discrimination Protections**
      i. How can we ensure that AI systems in self-driving cars do not discriminate against individuals based on their age, race, gender, disability, or any other factor?
      ii. How can we make sure the development teams for these AI systems are avoiding unconscious biases?
      iii. How can we ensure that the algorithms used in self-driving cars do not inadvertently favor certain groups of people or locations?
      iv. What safeguards can be put in place to prevent AI systems from adopting biases present in their training data?
   c. **Accountability**
      i. Who should be held responsible when a self-driving car makes a mistake that leads to an accident? The developer of the AI system? The owner of the vehicle?
      ii. How do you think laws should be adjusted to accommodate the increased usage of AI in areas of responsibility and liability?
   d. **Safe, Secure, and Resilient**
      i. What measures should be taken to ensure that AI systems in autonomous vehicles are reliable and safe?
      ii. What would be considered an acceptable error rate for self-driving cars?
   e. **Security and Effective and Data Privacy**
      i. How can we protect the AI systems in self-driving cars from malicious attacks?
      ii. How can we ensure the privacy of the passengers, considering that these cars might need to process personal data (e.g., regular routes, driving habits)?
      iii. As self-driving cars may gather large amounts of data about their environment, how do we ensure this doesn't infringe on the privacy of individuals or property captured in that data?

Conclusion

As stated previously, this short lesson is only meant to prime you on the topic of AI ethics. When working with AI applications as an ARS employee, we suggest you familiarize yourself with the AI Bill of Rights and any other departmental regulations, executive orders, or guidance pertaining to the use of AI, including the recent memo on the use of generative AI. We recommend taking a deliberate, methodical approach to implementing AI ethics in your work.

Optional Activity: A Crash Course in AI Ethics

Let’s look at MIT’s Moral Machine. The Moral Machine is part survey, part tool designed to help guide and educate people on AI decision-making.

Here’s the link to the site: https://www.moralmachine.net/ [2]

Watch the video on the main page, “Moral Machine - Human Perspectives on Machine Ethics”, then click the red button labeled “Start Judging”. An exercise will launch that will have you select the actions of a self-driving car in a variety of scenarios. **Be sure to use the “Show Description” button to see the details for each scenario.** Make your selections until you get to the Results screen. Review the Results screen and consider the following questions:

1. How did your answers compare to “Others”? Do you agree with the current consensus?
2. Do you think that this crowdsourced method is the right way to make these ethical decisions? If not, how should they be made (panel of ethicists, the discretion of programmers, elected officials, etc.)?
3. How do you feel about offloading potentially lethal decision-making to machines? While these scenarios seem contrived and there are likely alternatives to avoid killing anyone, at some level, AI-controlled systems are already making decisions now that may result in harm to some people and benefits to others. Who programs those systems and how they are programmed to act is already relevant to the discussions around AI.
4. What other ethical considerations should be taken into account when thinking about the behavior of self-driving cars?

References

[1] Rahwan, Iyad. “Use All Tools of Regulation.” *Evil AI Cartoons*, 17 Aug. 2022, www.evilaicartoons.com/archive/use-all-tools-of-regulation.

[2] Rahwan, Iyad. “Moral Machine.” Moral Machine, www.moralmachine.net/. Accessed 1 June 2023.

[3] “Ethics.” *Rome Call*, www.romecall.org/the-call/. Accessed 2 June 2023.

[4] “Ai Ethics.” *AI*, ai.ufl.edu/about/ai-ethics/#:~:text=The%20Rome%20Call%20for%20AI%20Ethic’s%20principles%20ask%20for%20transparency,research%2C%20education%20and%20workforce%20development. Accessed 2 June 2023.

[5] “The Ethical Dilemma of Self-Driving Cars.” *YouTube*, YouTube, 8 Dec. 2015, https://www.youtube.com/watch?v=ixIoDYVfKA0&ab_channel=TED-Ed. Accessed 24 July 2023.