Data, A.I., and Ethics: An (Informal) Companion to the Slide Deck

Jen Schellinck, Ph.D.
Sysabee / Data Action Lab

Patrick Boily, Ph.D.
Idlewyld Analytics and Consulting Services / Data Action Lab

As adults, we all have some understanding of ethics and what it entails to be ethical. Principles of behavior that we have adopted at one time or another could include:

- be honest
- be fair
- be objective
- be transparent
- be responsible
- be compassionate?
- be empathetic?
- etc.

Underlying these individual principles, there are **ethical systems** that drive us, and that we adopt and consider when we have to decide what actions to take (whether or not we decide to reject or accept them in the moment of action).

There's nothing new about any of this. This is all familiar territory.

So, if we're already ethical, and we already have systems in place to help us behave ethically, where is the need for yet another ethical discussion in the context of **artificial intelligence** (A.I.)?

However, A.I. itself is a new technology. And as such, it would allow us to do new things and to carry out the things that we can already do in a more powerful fashion: more **easily**, more **comprehensively**, more **broadly**, more **forcefully**, more **subtly**. On a very practical level, this is why we have to re-open the ethics discussion.

To understand this reasoning, let me tell you about a student in my lab who has just completed a Masters' thesis on the building ethical artificial intelligence and is now starting a PhD thesis on the same topic.

As a part of that thesis, he considered how humans themselves understand ethics and behave ethically, and he discussed one strategy that people often use – a strategy where their ethics boils down to a list of situations or scenarios, plus broader principles or categories that these situations fall under and ethical judgments about these. For example: taking something from a store without paying for it (scenario) is stealing (broader principle or category) and stealing is wrong (ethical judgment).

Pragmatically speaking, this trio of **scenario** + **broader principle** + **judgment** is a decently effective strategy for learning and carrying out ethical behaviours in particular contexts. But for it to work you first need to be able to draw that map between a particular scenario, the general principle or category that scenario falls

under, and the ethical judgment. Thus, if a person wants to behave ethically, and responsibly when using A.I., they must know what scenarios they are dealing with.

This sounds promising, but it leads to another question: To successfully come up with the situations – and then, in turn, the categories and the ethical judgments about these situations – d oyou need to understand how the technology itself works?

The answer to this question, unsurprisingly, is both yes and no.

Consider the ethics of using cars. In some cases, understanding the technology is not important. We don't need to understand much about the technology or inner workings of a car to know that if it runs over someone and kills them, a great wrong has been committed.

On the other hand, there are some situations where technological knowledge seems more critical. If we didn't know enough about the technology of cars to know that cars that use leaded fuel cause air pollution *via* the lead, would we be able to determine, in turn, that the car exhaust pollution was harming the cognitive development of children exposed to it, and thus letting people use cars that do this was, ethically, wrong? On a related topic, isn't it also possible that we might need to know about the technology of the car, and how people use it, if we want to **design** cars that are less likely to cause this problem, and to teach people how to use cars so that they don't harm anyone.

Even more broadly, if you had never before seen a car and didn't know what it was used for and how it ran, could you come up with all of the necessary situations and scenarios that would be required for consideration to determine how they can be used ethically? If you genuinely did lack an understanding of the technology you were using, and of its implications, and you did not create it, would that mean that you were off the hook, ethics-wise?

Related to this, the environmental attorney Andrew Kimball has recently proposed the concept of "**cold evil**", which he defines as a systemic evil in which we are all complicit. As he writes in the essay *Cold Evil: Technology and Modern Ethics*, "A synonym for the work "cold" is "distant," and a vital component in the success of modern cold evil is the physical and psychic distance that technology creates between the doer and the deed."

If any of this rings true, then we might want to say that ignorance, lack of intent, and acting in error provide no excuse for unethical behaviour – 'I didn't mean to' is not, and should not, and will not be used as an excuse. Consequently, it may be necessary for us to gain some measure of understanding of the technologies we use before we can hope to safely use it in a way that is consistent with the ethical principles to which we already adhere.

From General Ethics to Data Ethics

Broadly speaking, ethics refers to the study and definition of right and wrong conducts, "not [...] social convention, religious beliefs, or laws". (R.W. Paul, L. Elder). Influential Western ethical theories include: **Kant's golden rule** (do onto others...), **consequentialism** (the ends justify the means), **utilitarianism** (act in order to maximize positive effect), etc. Other influential ethical theories include: Confucianism, Taoism, Buddhism, Ubuntu, First Nations' principles of OCAP®, as well as common principles such as "Do No Harm" and "Informed Consent", which are substantially harder to achieve than might be hoped.

We can stray from our chosen ethical path in two ways – the first is to do so with the **knowledge** that an action will have consequences that go against our ethics, but which we still chose to carry out. The second is through **error**, where our actions or their consequences go against our ethical principles, but we remain ignorant of this fact because we don't have sufficient understanding of the consequences. It is worth noting that, when technology is involved, this second situation is most likely to occur when the technology is new – when we lack the ability to fully understand not only the technology itself but also all of the possible outcomes and implications of its use.

How do such general principles about ethics and types of possible ethical transgressions get translated into specific principles relating to data and artificial intelligence?

No answer can be forthcoming without first studying the **specifics** of data analysis and artificial intelligence and the **functionality** they provide. This will enable us to become aware of the situations and scenarios we need to build our ethics in the A.I. context. Then, once we've completed this investigation, we will be able to step back and categorize these situations in a broader fashion, and draw conclusions as to how to proceed.

How A.I. And Machine Learning Work

In their most basic form, A.I., machine learning, data science, data analysis, or even modern automation techniques require data to flow in, operations to be performed on said data, and for a result to come out.

Broadly speaking, we can think of data as the **fuel** which allows all of these technologies to run. Thus, for the purposes of this discussion we view data as the component of these technologies which enables their functionality. At the same time, like the oil that enables a car to run, we recognize that raw data has power **in and of itself**, even before it is used by the technology, and must be dealt with ethically in this isolated context.

To give you a flavor for the operations component of these technology, and how it interacts with the data component, lets very briefly consider the example of **deep learning** – or artificial neural nets more broadly. The names make them sound rather mysterious, but there is little magic behind the curtains.

Neural networks consist of inputs, an internal architecture consisting of nodes, connections between nodes, and functions within nodes. Depending on the inputs functions and the nature of the connections, different nodes fire and send a signal to connected nodes; the signal propagates through the network as a whole, resulting in a particular output. The only technically demanding part lies in determining the importance of each node in the network relative to the desired output (this is referred to as **training the network**).

This architecture allows appropriately-trained neural networks to answer **sophisticated** questions, based on the received input, with the (major) caveat that reasonable answers are only forthcoming when the networks are complex enough and trained properly, using a dataset that consists of **large numbers** of specific instances for which the correct answers are **already known**. This is a crucial part of the process – without high quality labeled data (which can be costly to acquire), the entire process is doomed to produce incorrect or inapplicable conclusions.

However, and very importantly, even in the **absence of large amounts of high-quality data**, the network structure can still be generated and will still produce outputs given a particular input. Deep learning is **operationally flexible** in that given a partial input, or input of an entirely different sort, an output will still be

produced, and may have the superficial veneer of **accuracy** and **precision**, even if it does not, in fact, reflect anything substantial or correct.

This is but a quick gloss of a single type of A.I. technology. Even with this gloss, however, it may start to become apparent that this technology can be used in a variety of problematic ways (for some additional technical details and specific examples of neural net behaviours, please see the Appendix). Given this risk, why use such technology at all?

Considering Functional Goals

The question of **risk vs. reward** can only be answered by considering what we hope to gain by using these technologies. One way to put this question is: what is the overarching functional goal behind adding artificial intelligence/data science/automation technology to a system?

Answers might include:

- to keep up with the Joneses
- to increase or add to the capabilities of the system
- to increase the power of the system
- to make the system better (better how? for whom?)

When Things Go Right

How might these new artificial technologies we are considering achieve these broad goals above? Broadly, we can say that they have the potential to increase the **efficiency**, **effectiveness**, **consistency**, **reliability**, **speed**, **accuracy**, **precision**, **range**, and **scope** of our current capabilities in a variety of areas.

More specific possibilities in the case of A.I./Automation technologies include:

- reaching conclusions/making predictions more accurately than humans analysts
- coming to useful novel conclusions, discovering new knowledge
- performing functions that humans don't want to carry out (e.g. menial, repetitive functions)
- increasing the extent of our situational awareness in range and degree

Potentially, as a result of successful application of this technology, we could then imagine that the system:

- becomes more capable (efficient, powerful, etc.), fair, consistent, trustworthy
- increases the agency, capabilities and dignity of participants
- empowers participants
- becomes more secure, less vulnerable to exploitation

All of these would seem to be well in line with positive ethical principles. To underscore this, let us look at a concrete example of using these types of technologies to improve people's situations by considering the case of the *Danish Medical Study*.

[Slides: Denmark, positive health example]

When Things Go Wrong

We stated earlier that we could stray from our ethical path either **deliberately** or **unintentionally**. If we start by focusing on the second of these two options, we can ask: what happens from a functional perspective when our technology goes wrong? what are some situations we may encounter?

From a strictly functional point of view, if we implement technology poorly, we may suffer **negative functional consequences** – a decreased functionality or capability of the system in some way (transparency, accuracy, flexibility, and options). In the case of A.I., the most concrete manifestation of this would be for the technology to come to **erroneous** or **inaccurate conclusions**.

We may also ask if an increase in capability itself can lead for a once ethical activity to become either **unethical** or **problematic in some other way**. For example, if these technologies allow for greater predictive ability, will the use of this increased ability itself become questionable ethically? Or will it possibly lead to a set of new behaviours that were not previously possible and which would themselves prove unethical?

Keeping these issues in mind, we can recognize that, as a consequence of either unintentional or deliberate actions taken through the use of these technologies the system may:

- cease to be equitable and fair
- decrease autonomy of people participating in the system
- decrease dignity
- increase alienation or objectification
- decrease the ability to be transparent or trustworthy
- decrease flexibility
- increase the potential for abuse (of all participants? of vulnerable participants?)

And potentially, this could lead to:

- people losing trust or confidence in the system
- people rejecting, rebelling against, defying, or refusing to participate in the system
- the system becoming (even more) vulnerable to exploitation or misuse

Interestingly, if some academics are to be believed, Denmark may also serve to illustrate these consequences.

[Slides: Denmark, negative example]

What could such failures look like at the CRA specifically?

- harm to the reputation/trust (CRA & GOC), leading to information being withheld in future interactions (refusal to grant data sharing consent)
- reduced compliance, leading to an increase in enforcing efforts/costs
- inequitable administrative/program outcomes – original data is incomplete (e.g. indigenous people)
- costly errors in program administration
- negative privacy impact on Canadians
- reduced Agency service ratings

- impact on self-service channels (Digital Government) – decline in trust
- increase in call centre volume (public unlikely to distinguish between front line automation and web-based applications)

Best Practices: Building Success Out of Understanding Technology

Based on an understanding of the technology, and subsequent functionality, what specific technological practices will **lead to success** (or **failure**)? Some of these practices focus on the data being used to power these technologies, some on the structuring and design of the technologies themselves, and some still on the uses to which the results of these technologies are put.

Successful conditions and practices might include:

- representative data
- appreciation for the limitations of the data obtained, in terms of proper uses and relevant conclusions
- consent respected, including circumstances in which data is re-purposed
- respect for privacy
- acceptable rate of false positives and negatives for the CRA and the clients
- explainability (neural nets), monitoring and algorithmic renewal/ adjustment standards (new data meets older models)
- model validity
- appropriate model choice
- accessibility/transparency (access to the algorithm that is being used)
- finesse
- granularity of results
- correct use and interpretation of results

Practices and conditions leading to failure might include:

- unavailable data, missing data, complex, unstructured, miss-transformed data, analyst training, unconscious and conscious biases (gender, racial, social and data-driven)
  ignorance or disregard for ethical use of data
- mismatch between algorithm and data
- inadequate data to power algorithm, lead to 'good' results
- lack of understanding of what the results of the algorithm mean

Conclusion

As the sophistication of A.I. technology increases, the effort required to understand its capabilities, functioning, and the consequences of its use will also increase. However, the fact that this presents challenges does not absolve us of our responsibilities to behave ethically. If we wish to reap the benefits of these new technologies, we must also take the time to develop standards of ethical use for these technologies and then put these standards into active play. By doing so we may move forward in way that is consistent with both our personal ethical principles and those of the organizations we support.