IRIE
International Review of Information Ethics Vol. 34 (10/2024)

Author(s): Peter Andes, Robin S. Lau, Geoffrey Rockwell, and Tammy Mah-Fraser

A Rossian Method for Applying Principles in AI: The Missing Link Between Principles and Policy

Agenda:

Author(s):

Dr. Peter Andes

- University of Alberta
- 🖂 andes@ualberta.ca

Dr. Robin S. Lau

- Alberta Innovates
- 🖂 robin.lau@albertainnovates.ca

Professor Geoffrey Rockwell

- University of Alberta
- 🖂 grockwel@ualberta.ca

Dr. Tammy Mah-Fraser

- Alberta Innovates
- 🖂 Tammy.Mah-Fraser@albertainnovates.ca

ISSN 1614-1687

Introduction

In the past several years there has been a rapid development of new technologies, applications, organizations, and institutions in the area of artificial intelligence (AI). At the same time, ethical reasoning about AI has not been able to keep up with the speed of these advances. As a result, developers are left to rely on existing rules, professional codes, policies and personal ethics which may not provide the appropriate guidance about ethical conduct and may require greater specificity (O'Leary). Many commentators have acknowledged the need for a clearer understanding of ethical values and principles to guide AI research. As Resseguier and Rodrigues emphasize, the purpose of ethics is to be

> a powerful tool against the cognitive and perceptive inertia that hinders our capacity to see what is different from before and in different contexts, cultures or situations and what as a result calls for change in behaviour. The turn to ethics is to ensure that AI [and digital technologies are] deployed in a manner that respect dearly held societal values and norms and puts them at the heart of responsible technology development and deployment. Ethics is about navigating murky and risky waters, and to do so, it needs to be watchful. The value of ethics is its constant renewed ability to see the new and it is critical to keep ethics alive and agile (Resseguier and Rodrigues).

In this paper we take up this need for careful ethical thinking about AI. We present the growing consensus on key principles of AI ethics as a sign of progress and offer a way to apply these principles to particular cases that appears missing so far in much of the scholarly discussion of AI ethics. Drawing on insights from the formation of the field of biomedical ethics, we argue that AI ethics should make use of the method based on prima facie duties derived from W.D. Ross' approach to ethics. A Rossian approach has proved influential in biomedical ethics. We further propose a modification to the list of principles proposed by Floridi and Cowls, arguing that the principles of explicability and accountability should be separated for ease of application. We argue that this method of applying principles is just what has been missing in AI ethics and is the crucial link between the now common lists of principles and putting them into practice in a way that can inform actual developments on the ground. In order to actually do their job, principles must have this connection to actual cases in order to have any effect in properly orienting policy.

Ethical Principles: Lessons from Biomedical Ethics

To inform the development of digital ethics around AI, ML (machine learning), and other data-enabled digital technologies, principles of medical ethics have been applied to guide the development of digital technologies (Mittelstadt). The moral obligation of medical practitioners is to advocate for their patients interests against institutional rights, with a common goal of promoting the wellbeing of the patient (Mittelstadt). This fiduciary relationship stems from the Hippocratic Oath, the Declaration of Geneva, and the Declaration of Helsinki, which all serve as a basis for ‘good’ professional conduct (Mittelstadt). Subsequently, the development and conduct of professional societies and boards, ethics review committees, accreditation and licensing schemes, peer self-governance, and codes of conduct in medicine, enable the translation of medical ethics principles into practice (Mittelstadt). This accountability mechanism helps health professionals fulfil their professional duties and establishes forms of self-governance by defining and enforcing good behaviour (Mittelstadt).

The development and implementation of digital technologies differs from medicine and traditional professions. The challenges are

- Digital technology development and implementation involves expertise from varied disciplines and professional backgrounds with diverse cultures, incentive structures and moral obligations (Mittelstadt). Digital technologies are largely developed by the private sector for application in public and private sectors (Mittelstadt). Companies can have relationships with their shareholders, where commercial interests are often prioritized (Mittelstadt), or traded off over public interests (Madaio et al.). In the absence of a fiduciary relationship, trust can erode when companies are not acting in the public's best interest (Mittelstadt).

- The increase complexity in development, implementation of digital technologies and their impacts requires organizations and society to understand the challenges at different points in the development lifecycle. The consequences and impacts of data-enabled digital technologies are difficult to predict and understand due to their system complexity (numerous interactions, numerous data sources) and distance (time, network) from their consequences and may result in unintended unethical behaviour (Mittelstadt).
- The short history of digital technologies and the lack of legal and professional frameworks are challenges to understanding the impact and consequences of digital technologies and to the translation of high level principles into ethical conduct in development and implementation of digital technologies (Mittelstadt).

Ethical guidelines alone are not a complete system for ethical decision making and not suitable for solving complex ethical problems (Canca). Where ethical principles have been developed by policymakers, companies and research institutions as a form of guidance for the development of digital technologies, merely having a set of agreed principles does not bring about actual change in the design of algorithmic systems (Hagendorff). A shift in culture from competition to cooperation, from self-regulation to sustainable system pathways to impact, and from technological solutionism to an ethical process of development and implementation of digital technologies may be required (Mittelstadt).

Given the challenges, this could make translating a set of high level abstract ethical principles into practice difficult (Mittelstadt). There are a lack of governance and systems supports for translation and implementation of the principles. A unified system approach to the development of digital technologies is needed. Ethical principles provide the foundation for this development. We argue that their uptake can be facilitated through the use of the Rossian method to address conflicts in AI principles.

The Principles of AI Ethics

There has been a recent convergence on key principles of digital ethics (Floridi and Cowls; Jobin et al.). Although there are differences in implementation challenges between the principles of biomedical ethics and those of digital ethics, this recent convergence is a crucial step in applying ethics to the digital sphere. Even as more ethical principles may be added to the core set of principles in the future, we argue that the core set are as follows, with a brief description of the main idea behind the principle. These principles largely follow the list offered in Floridi and Cowls, however we differ from their list in separating the principles of explicability and accountability.

1. Beneficence: Benefit others (including by preventing harm from befalling them)
2. Nonmaleficence: Do not harm others
3. Autonomy: Do not do things to another without that person's informed consent
4. Justice: Don't discriminate unfairly and distribute resources fairly
5. Explicability: Ensure the parties involved understand in general the way the technology at issue works
6. Accountability: Assign and accept responsibility for technology

These six principles form the core principles of digital ethics. Even as more may be added in the future, that does not prevent working with these widely agreed on principles now. We argue that explicability and accountability should be separated out, unlike the joint principle offered in Floridi and Cowls because this will allow for greater clarity. Explicability and accountability are two distinct concerns and it seems that any shortcomings that result from having 6 principles instead of 5 are more than outweighed by keeping these concerns independent. Better two singularly focused principles than one principle that really involves checking for two separate concerns.

A Rossian Method of Application

What is lacking in the surveys of principles that have been done is much discussion of how they should be applied. This is where consideration of the biomedical ethics literature is useful. The influential four principles approach, or "principlism", of Beauchamp and Childress employs an approach inspired by the work of W.D. Ross which sees multiple principles as important in our ethical reasoning, yet sees each as a prima facie (at first glance) or pro tanto (obtaining to a given extent) principle (Beauchamp and Childress). The Rossian method has us heed principles, and then offers a special approach to resolving conflicts of principles. In a particular ethical case where two principles conflict, we are to reflect on the principles and recognize one intuitively as of stronger weight. We then follow that principle as our actual duty in the situation, recognizing it as outweighing the weaker duty. The same approach can be employed for these six principles of digital ethics. Having an approach that can weigh multiple principles at once and show us how to approach conflicts of principles is important because so often in analyzing ethical cases principles will conflict. Without a way to resolve conflicts there will be no clear action to take.

To illustrate, consider the following case from the Markkula Center for Applied Ethics. Some jurisdictions have proposed that smart lampposts be installed on streets. The lampposts would have face recognition and audio recording capabilities that could be used to keep people safe and solve crimes, such as by detecting gunshots and identifying perpetrators.

When considering this case, applying the principle of beneficence would involve recognizing that smart lamppost technology has the potential to benefit people by helping to prevent and solve crimes. This counts in its favor. However, we also have an obligation to help others by preventing harm to them. The potential for abuse through loose control of the information gathered is one such potential harm that could result from smart lamppost technology.

In addition, the potential benefits, recognized in applying the principle of beneficence, must be weighed against autonomy concerns. Many of the choices we make concern the control of information about ourselves. Some information we may want to keep private. The smart lamppost technology threatens this by listening in on people and acquiring facial imagery. Depending on the implementation of the technology, this could be undertaken without the awareness of passersby. The principle of autonomy obligates us to respect people's choices about the extent to which they want to share information including images of them and recordings of their speech with law enforcement. Without a clear way to only record those who give informed consent (that is, who agree to be recorded and understand what they are agreeing to) this technology seems in violation of this principle. There is thus a conflict between beneficence and autonomy in addition to questions regarding the weighing of avoiding harms and benefits that fall under the realm of beneficence.

Employing the Rossian method, we now consider which duty seems intuitively stronger in this case in order to resolve the conflict. We argue that the smart lamppost technology seems morally impermissible to implement primarily due to privacy concerns, which fall under the duty to respect autonomy. Thus, on the Rossian method here, the principle of respect for autonomy seems intuitively stronger and outweighs the principle of beneficence.

Without a method for handling conflicts of principles, lists of principles are not going to be implementable. However, if we take the Rossian approach into digital ethics, as we argue ought to be done, the lists of principles become readily applicable. Even as there may be challenges in implementation in digital ethics not faced in biomedical ethics, we argue this set of 6 principles and the methodology explained here are readily applicable to ethical issues in digital ethics.

The Rossian approach is not without its critics. Some have opposed the idea that duties or principles should be the primary focus of ethics as opposed to other concepts such as virtue (Garcia), care (Carse), or consequences (Sinnott-Armstrong). However, we argue that one can embrace a Rossian method in AI ethics without committing at some deep level to the normative tradition of deontology as opposed to virtue ethics, the ethics of care, or consequentialism. This is in line with how Beauchamp and Childress defend their four principles approach. It is not meant to be a rival to normative theories but something that people endorsing different normative approaches can embrace. One can see the principles of digital ethics as reminders of ethical concerns that tend to be important in many contexts without being committed to the claim that the principles have a deeper status, or that duties or principles are the most important ethical concepts.

A different worry is that the method is liable to be too subjective, or based on personal values, with individuals simply deciding conflicts of principles based on their own personal opinions (Clouser and Gert). Some may aspire to an ethical approach that will eliminate all space for individual judgment and deliver unambiguous verdicts in all cases, leaving no room for reasonable disagreement. We argue that we should remain open to the possibility that disagreements among individuals will be an ineradicable feature of ethical reasoning, even if only in some cases. What the Rossian method for resolving conflicts offers even in the face of this possibility of ineradicable disagreement is much needed guidance on how to approach conflicts of principles when there would otherwise be none. If there is disagreement on which duty seems most intuitive, then the discussion must proceed to consider what can be said to try to draw out the plausibility of one intuitive response over another. Whether in the end agreement can be produced or not, the discussion can at least be narrowed down and the conflict clarified, so that discussion can proceed in a focused manner. The Rossian method is not necessarily the end of the discussion, but it is needed to ensure a focused beginning.

Although we suggest six principles instead of five, it should be clear that there should be no significant differences between using our list rather than Floridi and Cowls' list other than, if we are right, differences of clarity and ease of application. The disagreement is minor and does not suggest that there still remain deep and intractable disagreements on the principles of digital ethics. The principles laid out above can also be simplified to make application easier in non-academic contexts by using everyday language, with a simple system of imperatives, keeping in mind that they are to apply prima facie or pro tanto:

1. Help Others (Beneficence)
2. Don't Harm (Nonmaleficence)
3. Respect Choices (Autonomy)
4. Be Fair (Justice)
5. Be Clear (Explicability)
6. Be Accountable (Accountability)

Valuable time may be saved and efforts at knowledge translation accelerated through the ready adaptability of the academic ethical principles to non-academic contexts. Thus, we argue we should not be so wedded to the exact language that these principles have been articulated in throughout the scholarly literature but rather to the ideas that they capture, which can be communicated in more way than one.

Conclusion

Digital technologies (i.e., AI, ML, algorithms), their development and application is challenging knowledge structures built around specialized knowledge and siloes. The increasing complexity between the individual and society is changing at a pace that cannot be resolved by siloed knowledge. Currently, there is insufficient attention to the context or ecosystem where digital technologies operate (Gerhards et al.). What is needed is a shared knowledge space, for individuals to interact, reflect and learn. The greater the shared knowledge space the less context is needed to share knowledge, while small existing shared knowledge paces requires the greater need for contextual information and the more effort is needed to exchange information (Alavi and Leidner).

Ethics provides a perspective to tease out these complex issues into productive dialogue and a starting point to incorporate other diverse perspectives. Innovation happens at the knowledge boundaries, and innovation depends on the quality and number of conversations happening in the ecosystem. There remains much work to be done in taking up the Rossian method for applying principles and bringing it into conversation with the most pressing issues today. Vesnic-Alujevic et al. outline key ethical and society issues currently under discussion as a starting point for dialogue around the impacts of digital technologies on individuals and society. These issues have not been addressed in a comprehensive way or shown how they are connected and intertwined (Vesnic-Alujevic et al.). We propose that the Rossian approach the principles of AI ethics can serve as a framework for tackling these pressing issues and that further work needs to be done to systematically work out how the Rossian approach would handle each of them. Only then can we have begun the much-needed task of taking the principles of AI ethics from largely effete abstractions to concrete and lived policies for action.

References

Alavi, Maryam. and Dorothy E. Leidner. "Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues." MIS Quarterly, vol. 25, no. 1, 2001, pp. 107-136.

Beauchamp, T.L. and J.F. Childress. "Chapter 1: Moral Foundations." Principles of Biomedical Ethics, Oxford University Press, 2019.

Canca, Cansu. "Computing Ethics: Operationalizing Ai Ethics Principles." Association for Computing Machinery. Communications of the ACM, vol. 63, no. 12, 2020, p. 18, Business Premium Collection, doi:http://dx.doi.org/10.1145/3430368.

Carse, A. L. "The 'Voice of Care': Implications for Bioethical Education." J Med Philos, vol. 16, no. 1, 1991, pp. 5-28, doi:10.1093/jmp/16.1.5.

Clouser, K. D. and B. Gert. "A Critique of Principlism." J Med Philos, vol. 15, no. 2, 1990, pp. 219-236, doi:10.1093/jmp/15.2.219.

Floridi, L. and J. Cowls. "A Unified Framework of Five Principles for Ai in Society." Ethics, Governance, and Policies in Artificial Intelligence, vol. 1, no. 1, 2019, pp. 5-17.

Garcia, JLA. "The Primacy of the Virtuous." Philosophia, vol. 20, no. 1-2, 1990, pp. 69-91.

Gerhards, H. et al. "Machine Learning Healthcare Applications (Ml-Hcas) Are No Stand-Alone Systems but Part of an Ecosystem - a Broader Ethical and Health Technology Assessment Approach Is Needed." Am J Bioeth, vol. 20, no. 11, 2020, pp. 46-48, doi:10.1080/15265161.2020.1820104.

Hagendorff, Thilo. "The Ethics of Ai Ethics: An Evaluation of Guidelines." Minds and Machines, vol. 30, no. 99-120, 2020.

Jobin, Anna. et al. "The Global Landscape of Ai Ethics Guidelines." Nature Machine Intelligence, vol. 1, 2019, pp. 389-399.

Madaio, Michael A. et al. "Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in Ai." CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (April 25-30), 2020, pp. 1-14.

Mittelstadt, Brent. "Principles Alone Cannot Guarantee Ethical Ai." Nature Machine Intelligence, vol. 1, 2019, pp. 501-507.

O'Leary, Daniel E. "Ethics for Big Data and Analytics." IEEE Intelligent Systems, vol. 31, no. 4, 2016, pp. 81-84.

Resseguier, Anais. and Rowena. Rodrigues. "Ai Ethics Should Not Remain Toothless! A Call to Bring Back the Teeth of Ethics." Big Data and Society, vol. July-December, 2020, pp. 1-5.

Sinnott-Armstrong, W. "How Strong Is This Obligation? An Argument for Consequentialism from Concomitant Variation." Analysis, vol. 69, no. 3, 2009, pp. 438-442.

Vesnic-Alujevic, Lucia et al. "Societal and Ethical Impacts of Artificial Intelligence: Critical Notes on European Policy Frameworks." Telecommunications Policy, vol. 44, no. 6, 2020, EconLit, doi:http://dx.doi.org/10.1016/j.telpol.2020.101961.