Applied AI ethics: report 2019

De Bruijn, Berry; Désillets, Alain; Fraser, Kathleen; Kiritchenko, Svetlana; Mohammad, Saif; Vinson, Norm; Bloomfield, Peter; Brace, Hayley; Brzoska, Katarzyna; Elhalal, Anat; Ho, Katy; Kinsey, Libby; McWhirter, Ross; Nazare, Madalina; Ofori-Kuragu, Evelyn

For the publisher's version, please access the DOI link below./ Pour consulter la version de l'éditeur, utilisez le lien DOI ci-dessous.

https://doi.org/10.4224/40001806

Applied AI Ethics

Report 2019

This work was supported by the National Research Council of Canada (NRC), United Kingdom Research and Innovation (UKRI) Pilot Programme.

Digital Catapult

Digital Catapult is the UK’s leading advanced digital technology innovation centre. It drives the early adoption of digital technologies: to make UK businesses more competitive and productive, and to grow the country’s economy. Its AI and machine learning stream consists of a team of applied technology specialists, and the provision of facilities and programmes that support innovation and facilitate collaboration across large organisations, startups and academia.

National Research Council Canada

National Research Council Canada (NRC) is the Government of Canada’s primary national research and technology organisation (RTO), specialising in science and technology research and development. It has approximately 80 machine learning research staff and is internationally recognised for expertise in machine translation and sentiment analysis, machine vision, and biological and medical informatics.

Contents

1. Foreword

From Theory to Practice: Applied Ethics for AI Systems

Lord Clement-Jones, former Chair of the House of Lords Select Committee on AI and Co-Chair of the All Party Parliamentary Group on AI, and Professor Luciano Floridi (University of Oxford), member of the Centre for Data Ethics and Innovation and of the High-Level Expert Group (HLEG) on Artificial Intelligence of the European Commission, Chair of the Alan Turing Institute's Data Ethics Group, and of the Digital Catapult Ethics Committee, and former Chair of the AI4People project.

It is now commonplace to say that Artificial Intelligence (AI) has a huge potential range of applications, from biotechnologies to cybersecurity, from smart cities to health care. Indeed, AI is already having a major impact on our lives. Because of AI's wide and deep impact, the House of Lords Select Committee on AI[1] took the view—shared by nearly all witnesses and subsequent commentators—that our societies need a common and applicable ethical framework, so that we can develop AI policies, regulations, technical standards, and business best practices in ways that are socially beneficial and environmentally sustainable.

A wide range of initiatives have sought to establish ethical principles for the adoption of beneficial and sustainable AI. This is a healthy sign of interest and engagement. These shared frameworks will not guarantee success. Mistakes and illegal behaviour will continue to happen. But their availability does increase our chances of getting a clear idea of what ought to be done, how to evaluate competing solutions, and which ones to implement. They can also be positive drivers of innovation.

In a recent article, we wrote that "without an ethical framework, 'better safe than sorry' becomes the only guiding rule, excessive caution overrides innovation, and we all lose out."[2] A huge number of codes or sets of principles have been proposed over the past couple of years. Last time we checked, there were more than 70 proposals[3], suggesting a plethora of principles and guidelines, and more are in the making. When they are similar, the various sets of ethical principles currently available for AI are unnecessarily repetitive and redundant. And when they are different, they risk creating a "market for principles", where stakeholders can shop for the most appealing ones. Hopefully, this is only a stage in which many views are flourishing, which will soon begin to convergence around a shared, unified perspective. Past experience, in other sectors, from medical ethics to environmental ethics, show that this can be done. In very practical terms, for example, the food industry has

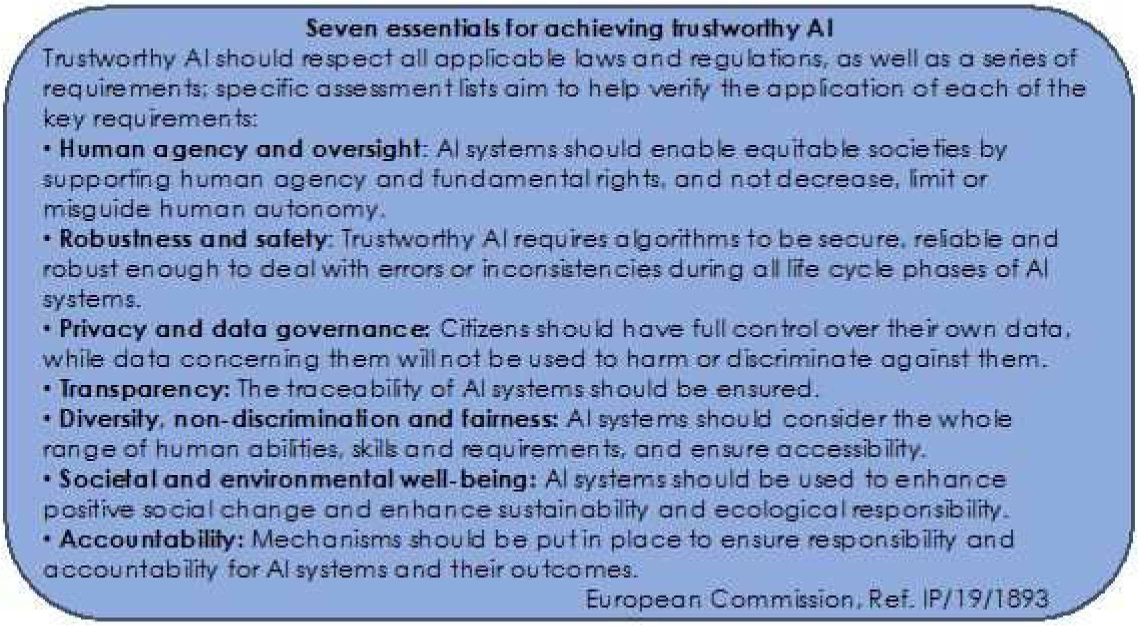
[1] AI in the UK: Ready Willing and Able? The Report of the House of Lords Select Committee on AI. https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

[2] See https://tech.newstatesman.com/policy/ai-ethics-framework

[3] See these two repositories: Algorithm Watch. The AI Ethics Guidelines Global Inventory (9 April 2019): https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/ and Winfield, A. (18 April 2019): An Updated Round Up of Ethical Principles of Robotics and AI. http://alanwinfield.blogspot.com/2019/04/an-updated-round-up-of-ethical.html

managed to adopt clear frameworks for supply chains and sustainability. And the good news is that we may already have the basis for agreement.

Last year the Select Committee, in the Report “AI in the UK: Ready Willing And Able ?”[4] suggested the adoption of a cross sector code of 5 principles. More recently, work at the European level—currently carried on by the High-Level Expert Group (HLEG) on Artificial Intelligence of the European Commission also on the basis of research developed by the AI4People project—has led to the adoption of similar fundamental principles, which can provide the ethical framework needed to support future efforts to create socially good AI across the European Union. They are summarised in Table 1. The HLEG principles represent a convergence of ethical thinking and we believe they are currently the best bet for international agreement.



Seven essentials for achieving trustworthy AI

Trustworthy AI should respect all applicable laws and regulations, as well as a series of requirements; specific assessment lists aim to help verify the application of each of the key requirements:

- **Human agency and oversight**: AI systems should enable equitable societies by supporting human agency and fundamental rights, and not decrease, limit or misguide human autonomy.
- **Robustness and safety**: Trustworthy AI requires algorithms to be secure, reliable and robust enough to deal with errors or inconsistencies during all life cycle phases of AI systems.
- **Privacy and data governance:** Citizens should have full control over their own data, while data concerning them will not be used to harm or discriminate against them.
- **Transparency:** The traceability of AI systems should be ensured.
- **Diversity, non-discrimination and fairness:** AI systems should consider the whole range of human abilities, skills and requirements, and ensure accessibility.
- **Societal and environmental well-being:** AI systems should be used to enhance positive social change and enhance sustainability and ecological responsibility.
- **Accountability:** Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes.

European Commission, Ref. IP/19/1893

Table 1 - The 7 ethical principles grounding the EU Ethics Guidelines for Trustworthy AI

There is no doubt that we all have a common focus on making sure the prejudices of the past are not unwittingly built into automated systems, and that individuals have greater personal control over their data. Fairness, transparency, explainability, and hence accountability, must be built into the design, development, deployment, and governance of AI systems. The G20 meeting in Osaka this June, with one of its key discussions on data governance, will be a good place to start, followed perhaps by incorporation in, or an annexure to, the Universal Declaration of Human Rights.

Senior figures in the AI field have sometimes rather impatiently repeated that consensus on an ethical framework for AI may be vital, but it is nothing without its real-world application and adoption.

[4] See footnote 1 above.

We do not disagree, but we do not see this as an alternative. It can and must be a win-win situation in which good ethics and great technology develop together. For this to happen, we sincerely hope that technology firms will drive good ethical practice, in everything they do, but we are also aware that we need to ensure that there are practical and regulatory tools that support their work.

In terms of the role for regulators, we are strong believers in ‘smart’ governance, which involves policymakers being well aware of technological developments (and the surrounding educational, scientific, social, economic, and industrial contexts) and anticipating their impact on society, before technologies are deployed at scale.

We see policy-making as a matter of design, and since we live in the age of design, we must ensure that it will be marked as the age of good design for all and for the whole environment. Good design implies having frameworks in place that are agile and adaptive, depending on the course of technology and its deployment in society; and good design requires policymakers to develop a robust understanding of the global AI development landscape, to assess the impact of AI on individuals and groups, and to deploy rapidly governance models that can intervene when technology deviates from societal values.

To achieve all this, we need to set standards for AI and consolidate the toolbox available for ensuring that AI applications meet ethical principles, at every stage, of design, development, and deployment. And this is where there is a key role for organisations like the Digital Catapult and the Centre for Data Ethics and Innovation.

There are quite a number of tools that need developing, but there are already existing regulatory concepts, such as those present in the General Data Protection Regulation (GDPR) and the Data Protection Act, including ‘privacy by design’ and ‘data protection impact assessments’, which we can already deploy. Their scope could be expanded, to cover the full range of ethical principles, ensuring that new uses of AI consider ethics from the start, and that any risks arising from the use of AI are considered in a structured, accountable and redressable fashion. This means that developers and those applying AI solutions cannot and must not shelter behind “black box” excuses.

The Ethics Framework[5] created by Digital Catapult’s Machine Intelligence Garage, is a very useful addition to our toolkit for companies that have considered the ethical implications of the products and services they develop. The Framework is a valuable contribution to the field of corporate governance and ethical AI, which is still in its infancy. Companies need to be transparent about the impact of AI solutions on their workforces and on decision making. They need to accept that they are fully accountable where the introduction of new technology (by which we mean the combined impact of several digital technologies, including AI, IoT, Blockchain, and Cloud) makes a significant impact on employees and customers. Companies should in particular consider whether independent, ethics advisory boards ought to be introduced. We are convinced that companies that will monitor, manage,

[5] See https://www.migarage.ai/ethics-framework/

and communicate effectively about the ethics of their AI products will have a competitive advantage against those that do not.

Companies need to educate themselves as to the ethical challenges of technology, such as those shortly due to set out in a briefing by the Institute of Business Ethics. It is vital that they set up their own AI audit processes—rather than waiting for some new AI regulatory function—in order to assess algorithms, related uses, and investigate complaints by individuals. This means that auditors hitherto used only for financial audit also need to step up and transform the profession and their practices. We need to develop appropriate audit tools too as a matter of urgency. Finally, it is really important to make progress in kite marking.

There is a strong case, particularly in the public sector, to introduce a new 'Certificate of Fairness for AI systems' with a kitemarking scheme to reflect it. The criteria would be defined at industry level, similarly to food labelling regulations. This could be informed by the 10 principles for public sector use of algorithmic decision making, set out last year by NESTA[6].

What we have suggested so far relies on corporate responsibility, industrial sector self-regulation, and a sincere commitment, on the part of business, to embrace good citizenship. But if the private sector fails to rise to the challenge and does not grasp that socially good AI is also good for business, then society and government need to be prepared to move quickly and firmly. Standards, requirements, and regulations may need to be transformed into law, not so much by trying to catch up with technological innovation, but by looking at the long-term values and goals that our society wants to see fulfilled by the development of AI as a force for good. Contrary to current narratives, policy and law making is a matter of steering, not accelerating or decelerating innovation. If we like the direction of travel, we will want to get there even more quickly.

At the same time, it is crucial to stress that governments themselves need to be aware that they cannot be immune from regulation. Some of the worst misuses of technologies have been caused by states and governments abusing their powers. Algorithmic decision making is increasingly prevalent in policing, justice, immigration, and social security systems. We need clear rules about this as well. Building competency for AI governance in the public sector is a crucial part of the agenda.

Self-regulation, design of the environment in which innovation takes place, public opinion, and legislation are four crucial tools for policy making. To these, two more can be added when it comes to AI: public procurement and regulatory sandboxes, at various levels that could be municipal, national, and global. Sandboxes could aggregate the problems into one place so they can be worked on and the solutions rolled out appropriately. A sandbox approach can assist in speeding up innovation and scaling up adoption of AI projects. It can deliver a model for competence in understanding the application's fundamental aspects and how to map, track and measure it in a given context using metrics and evaluation.

[6] See https://www.nesta.org.uk/blog/10-principles-for-public-sector-use-of-algorithmic-decision-making/

Of course, sandboxes will need to be carefully managed between the public sector, regulators and business. But they could help in improving public procurement as well, which remains a crucial way of exercising pressure on the private sector to deliver AI solutions that are both socially desirable and environmentally sustainable.

We must ensure that ethics is an enabling force, not a way of avoiding regulation, diluting the positive impact of nascent applications, or merely improving reputation and public relations. In fostering public trust in AI we need not only to get the ethical framework right but also to be seen to do so in a transparent way. This is an epochal opportunity. We cannot miss it. And it is in all our interests to get this right first time.

2. Background

UK Research and Innovation (UKRI) and National Research Council Canada (NRC) signed a memorandum of understanding in July 2018 to facilitate the delivery of collaborative jointly-funded research and innovation programmes[7].

The memorandum confirmed a joint aspiration to intensify and strengthen co-operation between Canada and the UK in the fields of science, engineering, research and innovation, including networking and relationship-building activities.

This report synthesises the findings of one such network relationship-building activity, supported by NRC and UKRI as a result of the memorandum of understanding. This activity, led by NRC and Digital Catapult, was focused on how to translate good intentions for the responsible development and deployment of artificial intelligence (AI) into practical action: 'applied AI ethics'.

Canada and the UK share common values and an explicit goal[8,9] to promote responsible and sustainable development of AI. NRC and Digital Catapult have complementary networks and expertise, ranging from fundamental research to adoption of AI, and related applied AI ethics [10,11].

2.1 Applied AI ethics

AI systems are moving out of the laboratory and into the real world at a rapid pace. The decisions they make now affect people's lives in diverse domains, such as medical treatments and diagnosis, hiring and promotions, loans and the interest rates borrowers pay, prison sentences and so on[12,13,14]. This widespread adoption has produced a corresponding increase in ethical concerns about how decisions are made using AI, particularly as these decisions can be inaccurate or unacceptable[15].

[7] https://www.ukri.org/news/uk-canada-collaboration-to-build-on-research-and-innovation-strengths/

[8] https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

[9] https://www.canada.ca/en/government/system/digital-government/modern-emerging-technologies/responsible-use-ai.html

[10] https://www.migarage.ai/ethics-committee/

[11] NRC consistently applies research ethics principles, and is in the process of re-examining them

[12] Pasquale, F. (2015). The black box society: The secret algorithms that control money and information (1st ed.). Boston, USA: Harvard University Press.

[13] Roshanov P.S., Fernandes N., Wilczynski J.M., Hemens B.J., You J.J., Handler S.M. et al. (2013) Features of effective computerised clinical decision support systems: meta-regression of 162 randomised trials. BMJ, 346 :f657

[14] Wexler, R. (2018). Life, liberty, and trade secrets: Intellectual property in the criminal justice system. Stanford Law Review, 70(May), 1343–1429.

[15] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 1–42.

In response to ethical concerns, a number of entities have created codes for the responsible development and application of AI. However, there is a gap between aspiration and viability, and between principle and practice. This is the gap that applied AI ethics seeks to fill.

2.1.1 Consultation

“*I’ve seen many ‘ethics codes’ focused on AI, and while many of them are very good they’re more directional than prescriptive – more in the spirit of the Hippocratic Oath that doctors are expected to live by. Meanwhile, many data scientists are hungry for something more specific and technical. That’s what we need to be moving toward.*”
-- Rumman Chowdhury, Accenture’s Responsible AI Lead.[16]

The Canada-UK collaboration brought together technologists, lawyers, standards bodies, academics, civil society groups and other interested parties to examine how to support practitioners in translating ethical aspirations into values-aligned AI; to build technologies with positive effects and avoid negative consequences from the technology they develop.

Two events were convened: first in Ottawa, Canada[17] (21 and 22 February 2019) and next in London, United Kingdom[18,19] (11 and 12 March 2019). At these events, the following questions were posed:

- What is the demand for applied AI ethics?
- Is there supply to meet demand?
- What support is required to encourage progress in applied AI ethics?
    - Who should be involved?
    - What is the funding or business model?

The methodology and findings of the two events are detailed in sections three and four, respectively.

2.1.2 Literature review

In addition to the network relationship building activities, Digital Catapult commissioned a companion piece of research to develop a lexicon for applied AI ethics and to survey its state of development.

The research found more than 800 papers, tools and processes targeted at understanding, operationalising and automating adherence to, or monitoring of, good ethical practices when developing and deploying AI-driven products and services. These were then grouped and

[16] https://www.sas.com/content/dam/SAS/documents/marketing-whitepapers-ebooks/third-party-whitepapers/en/ai-momentum-maturity-success-models-109926.pdf

[17] Symposium and workshop: https://sites.google.com/view/aiethics/home

[18] Symposium https://www.eventbrite.co.uk/e/canada-united-kingdom-symposium-on-practical-ai-ethics-registration-55343106722_

[19] Workshop https://www.eventbrite.co.uk/e/canada-united-kingdom-workshop-on-practical-ai-ethics-registration-56518802259

mapped to a grid, comprising ethical principles[20] on one axis, and stages of the machine learning development pipeline on the other.

This study is the first attempt to visualise the applied AI ethics field as a whole: to spark discussion and debate, and to highlight potential new research questions and challenges. The resulting research paper is called ‘Moving from the What to the How: An Overview of Applied AI Ethics Tools, Methods and Research’[21] and will be available shortly as a preprint on arXiv[22].

2.2 This report

Sections three and four of this report review the methodology and outcomes of the Ottawa and London networking activities respectively.

Section five identifies the four specific recommended activities that arose from these networking activities. These focal points are the ‘hub’ through which the network’s strengths will be utilised to advance applied AI ethics research, development and adoption; the development of specific products; continuation of the fruitful networking activity via an annual EIAI[23] event and the use of competition to stimulate activity and innovation in applied ethics.

Section six offers some conclusions.

[20] The five ethical principles - Beneficence, Non-Maleficence, Autonomy, Justice, Explainability - that ‘should undergird [AI’s] development and adoption’ (Floridi, L., Cowls, J., Beltrametti, M. et al. Minds & Machines (2018) 28: 689. https://doi.org/10.1007/s11023-018-9482-5)
[21] Morley, J., Kinsey, L., Elhalal, A. Moving from the What to the How: An Overview of Applied AI Ethics Tools, Methods and Research (2019) (in preparation)
[22] https://arxiv.org/
[23] Ethics in AI

3. Ottawa

3.1 Starting Point

On 21 and 22 February, 2019, the National Research Council Canada and Digital Catapult (UK) held a two-day symposium on Ethics in AI, which took place in Ottawa, Canada. The objective was to bring together and foster collaboration between various parties involved in establishing ethical AI in Canada and the United Kingdom. In recent years there have been several initiatives focused on identifying the ethical issues and challenges posed by the increasing use of AI. While these events have helped to improve our understanding of the problems, to date little progress has been made towards identifying clear and actionable solutions.

3.2 Ambition

The Ottawa symposium was the first of two such events, hence the goal was to establish a foundation for future discussions on practical tools for ethical AI. Presentations were given by experts across a range of related disciplines, including researchers, industrial players, legal experts, and policy-makers. The discussion was then further focused on actionable next steps.

3.3 Methodology

Over the course of the two-day event, a broad and comprehensive overview of issues and existing solutions were provided through presentations, panels, and posters (day one). This laid the groundwork for deeper and more focused discussion on practical AI ethics (day two).

Registration for day two was limited to 40 participants. The format was highly participatory, involving an agenda-setting brainstorming phase followed by breakout sessions under the guidance of an experienced facilitator. This format permitted more focused discussion and idea generation around the specific topics related to practical and ethical AI, such as accountability, privacy, explainability, fairness and inclusion. The workshop goal was to define actionable 'how-tos' that practitioners and policy-makers can use to ensure that their AI policies and applications are ethical. Both technical solutions to technical problems and non-technical solutions to more societal problems were discussed (such as automating while preventing mass unemployment).

During the eight breakouts,conducted as two sets of four parallel sessions, discussion among the self-selected participants was primed by seed questions. At the end of each session, one member from each group summarised their discussion.

3.4 Outcomes

There was a beyond-capacity interest in attendance, and over 120 people from over 80 institutions and affiliations attended the Ottawa event. Participants using Twitter sent 350 tweets using #EIAI19 on day one alone; all were positive towards the event, the talks and the discussions. Several people at the event praised the caliber of the invited speakers and the many affiliations represented by participants. Many had been to other AI-for-good events before and were pleased with how useful this event was to them. For many others, this was their first event of the kind.

3.4.1 Day 1: Symposium Presentations, panels, and posters

Day one of the symposium involved over 30 presenters, panellists and moderators from diverse backgrounds, such as computer science, law, ethics, policy making, business and healthcare. The slate of thirteen invited speakers included keynotes by Alison Paprica and Joelle Pineau. Panellists discussed privacy, transparency, and explainability (panel one) and governance and accountability (panel two). Student posters had been solicited and eight were presented. Details of presentations, panels and posters can be found in the workshop program[24].

In his opening remarks, NRC president Iain Stewart emphasised that ethics for AI technologies concerns not only the government or business, but also affects people across all segments of society. Speakers agreed that artificial intelligence can be a force for good, as it has the potential to significantly benefit our society and our environment. However, ethics violations can negate the benefit for AI users and for society. Multiple presenters addressed ethical issues within healthcare, human rights and language technologies, as well as policies, guidelines, toolboxes, and other practical solutions (existing or prospective) that accelerate ethical AI outcomes.

Healthcare: In their presentations, Jennifer Gibson and Alison Paprica described AI's great potential for transforming healthcare, including strengthening health policy and planning, addressing critical service gaps to meet care needs, improving health outcomes and addressing health equity gaps. Medical practice ethics, clinical research ethics and privacy regulations are well-established and generally well-respected in this field. Social licence (whether earned or lost) is a key ingredient in this mix. AI will test and probably drive changes to such boundaries as risks and benefits are rebalanced. Only through public engagement and transparency can this process succeed.

Privacy and Human Rights: David Van Bruwaene provided a compelling use case in SafeToNet, an app to ensure children's online safety and wellbeing while safeguarding their privacy. An AI algorithm monitors the child's online activity and makes decisions about disclosing information on behalf of the parent. However, to train ethical systems such as this may require the collection of sensitive or protected information about users in the first place.

[24] Full Ottawa Day 1 workshop program: https://sites.google.com/view/aiethics/program

Petra Molnar explored automated decision-making within Canada's immigration and refugee system ('Bots at the gate'), and discussed the human rights ramifications of deploying AI technologies. Risks include raising (existing and new) barriers to entry, tensions between the need to minimise and maximise the amount of collected data, and the (im)possibility of opting-out for applicants, a vulnerable population. Automated systems have a potential to strengthen the transparency, regularity and explainability of administrative decision-making. Faults, malfunctions, and (perceived) bias can sink public confidence. Transparency, accountability mechanisms, independent oversight and binding standards are urgently needed here.

Language technologies: Recently, powerful new language technologies have emerged, resulting in significant improvements in areas such as machine translation, personal assistants and question-answering. Graeme Hirst described use cases for both positive and harmful applications. For example, technologies that detect fake reviews can also generate fake reviews; tools to attribute true authorship can also unmask anonymous dissidents. Natural Language Processing (NLP) gravitates to English, creating inequities with other languages. This linguistic exclusion further extends into demographic exclusion, since systems tend to be trained on text written by or about white middle-aged males and reflect that bias. Well-known examples include Amazon's use of tools that were later found to downgrade resumés with 'woman-related words, and Google Translate was discovered to favour 'his' for 'doctor' and a 'her' for 'nurse', even when the source text indicated no gender. Similar gender bias was described by Saif Mohammad: more than 75% of 219 studied sentiment analysis systems consistently marked sentences involving one gender or race with higher intensity scores. Such biases are more common for race than for gender, and differ depending on the affect dimension involved. Research to counteract this (for example, on de-biased word embedding) has not yet led to effective solutions.

Existing guidelines and regulations: Several speakers and panellists referred to existing structures: the Montreal Declaration, the General Data Protection Regulation (GDPR), and the Health Insurance Portability and Accountability Act (HIPAA). In 2010, Ann Cavoukian designed the 'Privacy by Design' framework that was unanimously passed by international data protection and privacy commissioners, and helped to shape GDPR. Privacy by Design recognises that retrofitting systems for regulatory compliance is unsustainable. Instead, it prescribes being proactive: identify the risks and address them before any harm can be done, and embed active user-centric privacy by default, end-to-end. Following the success of the Privacy by Design guidelines, Ann Cavoukian created 'AI Ethics by Design' which includes seven principles:

1. Transparency and accountability of algorithms
2. Ethical principles applied to the treatment of personal data
3. Algorithmic oversight and responsibility
4. Respect for privacy
5. Data protection/personal control via privacy as default
6. Proactive identification of security risks to minimise possible harm
7. Strong documentation to facilitate ethical design and data symmetry

Anat Elhalal described an ethical framework that Digital Catapult has developed to help AI companies, especially startups, to design and deploy ethical AI products. The framework also consists of seven concepts which can be applied during different stages of product development:

1. Be clear about the benefits of your product or service
2. Know and manage your risks
3. Use data responsibly
4. Be worthy of trust
5. Promote diversity, equality and inclusion
6. Be open and understandable in communications
7. Consider your business model

Samantha Brown introduced Consequence Scanning by doteveryone – a new agile practice that prompts organisations to reflect on the (un)intended positive and not-so-positive impacts of their product at various stages of its iterative development process. Consequence Scanning requires a multidisciplinary and cross-functional team. Michel Girard summarised a recent Centre for International Governance Innovation (CIGI) report on the available standards in digital technologies and their impact on digital transformation. The report argued that the sector is mature enough to develop foundational standards (such as common definitions, standardisation of data taxonomy, data sharing models) and verification and certification mechanisms for AI products. Keith Jansa introduced the Canadian CIO Strategy Council, and its efforts in creating a forum for developing standards and guidelines for the ethical design and use of AI within national and global partnerships. Accreditation as a standard-developing organisation is in progress.

Demand for more practical tools: There is a thirst for assistive tools to connect practice guidelines and regulations with actual AI applications. Anat Elhalal pointed out that many companies have embraced the concept of responsible AI but have limited skill in its implementation. Assistive tools tend to have limited usability; testing requires real-life scenarios under the right conditions for their deployment. Christine Henry described her experiences with a range of existing practical screening tools, as DataKind UK was considering using a predictive model to identify the people most in need of support from a food bank. There are three types of ethics assessment tools:

- Awareness-building tools, such as model cards, nutrition labels and Digital Catapult's ethics framework, which bring ethical points to stakeholder discussions
- Enabling tools, such as Doteveryone's consequence scanning practice, which are agile approaches for transparent working at every reporting period
- Substituting tools, such as IBM's AI Fairness 360, which can make decisions without human critical assessment of the context and are therefore potentially dangerous - these should be used cautiously and mindfully.

Fiducia (meaning 'trust') is an ambitious platform that was pitched by Wilco van Ginkel, It aims to provide a tool for holistically assessing all ethical aspects of a system (including fairness, explainability, security and safety) for all components (data, infrastructure, and model). The envisioned platform will be a knowledge base covering technical, ethical and legislative perspectives. Joelle Pineau described the challenging ambition to translate ethical concepts like privacy, fairness and safety into mathematical objective functions that can be optimised during the AI training stage. Such functions would also enable quantitative evaluation of a system's ethical performance and provide a better understanding of its stability and predictability of output, even under the pressure of deliberate threats, such as perturbed input designed to confuse a model into inaccurate predictions.

3.4.2 Day 2: Participatory workshop

The agenda setting produced eight themes for discussion in the breakout sessions. In this section, we summarize these discussions. More detailed minutes can be found online[25].

Tools for data and algorithms:
Better tools are urgently needed to uncover bias, provide explainability, permit secure data sharing, track data provenance and maintain privacy through de-identification and/or encryption. System design should consider allowing users to have certain control over their own data, for instance by using data trusts. Benchmarking paradigms are also needed to track accuracy, bias, etc. This also supports replicability of AI system output.

Dos and Don'ts:
There is a general need for a repository of ethical AI best practices, providing concrete, practical solutions. It should be user friendly and searchable by topic (for example, GDPR, bias). A wiki-style barn-raising event (when an online community meets to build content together) might be a productive way to get such a repository started. The repository should be open, and led by an organisation with generally recognised legitimacy. It should contain best practices, possibly expressed as design patterns with comments, reviews, ratings and links to real-world applications. Taking into account that what's ethical in one culture may not be in another, these various perspectives could be accommodated by using a Wikipedia-like neutral point of view (NPOV) approach.

Certification and compliance:
Government plays multiple roles in the development and use of AI, each potentially shaping regulatory structures.

- Government is a **client** to AI solution providers, which means that ethics can be made part of the procurement process, enabling government to lead by example. At minimum, it can request an ethics statement from bidders, which can also be used to inform and shape future policies.
- Government also has a traditional **regulator** role and can use structures from Health Canada, CFIA, EMA, FDA and the NHS as models. Regulated markets, such as medical devices, can pioneer the inclusion of ethical requirements in their controls and legislation.

[25] Day 2 minutes: https://drive.google.com/file/d/1m5dmU5h2S3vgPfGAUyE4YHnM3cswKcKq/view

- Government as a **funder** of research can encourage/impose grant requirements around ethics.
- Government as an **all-in-house entity** (tools + data + applications all in-house) should lead by example, and adhere to (inter)national directives, such as the Directive on Automated Decision-Making, which is being introduced in Canada[26].

The process of arriving at AI ethical regulation would not necessarily be top-down, but may be partially **industry-**driven, as with the necessary pressures from investors and shareholders, commercial organisations can self-regulate to some extent.

Academia has a stronger track record of self-regulation, largely through the peer review process. The creation of a Hippocratic-style oath for AI practitioners has been proposed (coined as the 'Lovelace oath' in UK Parliament). As such an oath is taken when transitioning from education to practice, it should be supported by ethics courses as part of the curriculum.

Compliance can be driven by human rights commissions, privacy commissioners, or conceivably an AI-ethics ombudsman for a lower access threshold to correct what went wrong. Formal courts, but also the court of public opinion add powerful mechanisms. Babylon Health's controversial diagnostic chatbot led to its NHS accreditation being revoked. There are various (inter)national regulation initiatives under way, we expect to see an active convergence.

Acceleration and adoption:
Ethical practices in AI can be accelerated by cost and effort reduction, incentives, and by publicising the consequences of poor practice. Encouraging the development of shared cultural and legal norms and practices would also increase adoption. Increased public awareness of companies and organisations that comply (and do not comply) with ethical AI principles can apply pressure to those which are not yet compliant. We can learn from the adoption of other practices (such as cybersecurity) or the adoption of ethics in other sectors, such as medicine.

Literacy:
There is a need to improve public literacy and education about AI, and its current uses and limitations. An informed public discussion can only take place within a society that has some understanding of the subject, including opting in or out of products and services that use AI technology.

The need for incorporating AI education into both secondary and university-level education was identified. It is not only computer science curricula that should cover ethics, but future practitioners in any field, whether a professional domain or business studies, also need to have a basic understanding of the benefits and limitations of AI in the context of their own

[26] See https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592 and https://www.canada.ca/en/government/system/digital-government/modern-emerging-technologies/responsible-use-ai.html

domain. Some of the potential tools that were suggested to improve AI literacy, of varying ambition, included seminars and TED Talks, the introduction of standardised labelling for AI products (similar to nutrition labelling), and the creation of course curricula for universities and companies.

Communication:

Some aspects of this theme overlapped with the *literacy* theme, notably the need to inform the general public about AI technology, not only in relation to its hazards but also about its positive successes. This needs to happen in a way that is accurate, accessible, appealing, entertaining and targeted using the appropriate level of detail. Suggested channels included TED Talks, Wikipedia, museums and other cultural organizations; and (inter)national AI bodies such as the Big Data Value PPP in Europe, or the Partnership on AI.

Redress was discussed: people sometimes feel they are on the wrong end of an AI decision, and communication on explainable algorithmics is key, as are opt-out processes or correctable output.

Bias and inclusiveness:

Undue bias and lack of inclusiveness have the potential to create serious adverse consequences in the application of AI technology. Bias doesn't stop with gender or race, but can include characteristics such as age, geography, political beliefs, genetic make-up, (first) language, or medical/physical condition. Biases can originate from the input data, or from the implemented algorithm, or both in tandem. Good practice dictates that input data should always be assessed to ensure that it is representative for the problem or solution concerned. In practice, some populations are under-sampled, and this effect is amplified when under-represented aspects interact (for example, male nurses in rural Manitoba). Research led by Cynthia Dwork was mentioned in this context (https://arxiv.org/abs/1104.3913). Whether accidentally or deliberately, algorithms might favour one population sub-group over another. Examples of this can be found in banking, insurance, travel and recruitment. An algorithm might operate within its error margins overall, but with errors that are skewed towards one particular sub-population. Mechanisms to prevent or mitigate algorithm bias include ongoing validation, avenues for recourse and algorithmic interventions such as bias correction or cost-sensitive learning.

Lack of inclusiveness also occurs within population segments that tend not to use certain technologies and therefore do not contribute to source data. This means that they could be cut off from the benefits derived from that AI — the data-rich become richer. Technologies such as smart devices and improved connectivity to remote communities may be an equalizer, although affordability and availability remain a problem.

'Geneva Convention' of Ethical AI:

Global conventions and treaties - such as the Geneva convention, the Ottawa treaty on landmines and the Declaration of Helsinki - provided inspiration for this concept. Autonomous anti-personnel weapons were identified as one potential no-go zone. Numerous initiatives including the Centre for the Governance of AI (UK) and the International Panel on Artificial Intelligence (Canada and France) gain traction, and may be the more promising way for convergence. Local laws have the potential to create global impact, due to the relatively borderless nature of the technology market. GDPR is such a case in point, where North American enterprises strive towards compliance to ensure their continued access to the European market. Human rights, including the right to be forgotten, should be viewed through the lens of both individual rights and group rights.

3.4.3 Concluding remarks

At the beginning of the workshop, participants were asked what they hoped to get out of the session. The most salient answers were:

- Obtain a better understanding of the ethical issues posed by AI
- Obtain actionable information for making AI applications and programs ethical
- Create a vibrant community of people interested in ethics and AI and who will continue to interact after this event
- Generate up to three concrete initiatives with a clear action plan and - ideally - some funding

The general feeling was that the workshop achieved all of those objectives. Most participants, even those already well versed in the topic, felt that they came out of the event much more knowledgeable than when they went in. Within the notes of the various breakout sessions are several nuggets of wisdom that provide clear and actionable guidance on how to deal with specific ethical issues in AI. These, and other best practices, will be included in an open, collaboratively built repository of ethical AI patterns.

There was a palpable energy in the room, and all participants expressed an interest in continuing interaction on an ongoing basis. The London event which took place less than a month later, would help to maintain that momentum. This initiative is only one of many worldwide trying to shed light on the complex issues of ethics in AI, and many candles can in the end light up a dark room.

Slides and video recordings of the presentations were made publicly available shortly after the Ottawa symposium, as well as minutes of key discussions[27].

[27] Slides of Day 1 workshop: https://sites.google.com/view/aiethics/program
Minutes Day 2: https://drive.google.com/file/d/1m5dmU5h2S3vgPfGAUyE4YHnM3cswKcKq/view

4. London

4.1 Starting point

The events in Ottawa established that there is a problem with ethics and AI. This was illustrated by examples of unintended consequences arising from AI-enabled products and services, and from case studies describing the challenges involved in attempting to address ethical concerns in a principled and efficient way.

The role of technology in addressing ethical challenges was debated, and it was recognised that rather than being the entire solution, it can play a part in addressing challenges that have wider socio-cultural-economic contexts. However, in the interests of efficiency, technology tools promise to remove friction from ‘doing the right thing’ by operationalising and automating tasks that might otherwise consume time and resources where there are competing demands.

Participants were united in wanting to conduct AI research and create AI applications in a responsible way. The question is: how?

4.2 Ambition

The ambition for the London events was to develop concrete proposals for translating good intentions for responsible AI development and deployment into practical action. A broad range of stakeholders were invited to participate and to actively contribute and challenge ideas.

4.3 Methodology

As in Ottawa, the London event was comprised of a symposium and a workshop, with complementary content, and structured to facilitate progression from high-level discussion to practical outputs and recommendations.

4.3.1 Symposium

Day one of the event was an evening symposium attended by more than 120 attendees from academia, government, the wider public, and the third and private sectors (including both smaller and enterprise businesses). The evening introduced the project objectives and the collaboration between Digital Catapult and National Research Council Canada.

The keynote address was delivered by the Lord Clement-Jones, (Chair of the House of Lords Select Committee on Artificial Intelligence (2017-) and Co-Chairman of the All-Party Parliamentary Group on Artificial Intelligence) and examined the wider socio-economic landscape in relation to the development of AI and ethical principles, why it is important for the discussions to move forward to the more practical, and how collaborations such as this project play a key role.

After the keynote, Jessica Morley (Digital Ethics Lab, Oxford Internet Institute, University of Oxford) provided a preview of research to develop a lexicon for applied ethics and to survey the state of its development (see section 2.1). She provided examples of tools currently available and where they fit into the lexicon, and highlighted the substantial areas where research and tools are lacking. At present there is a lot of focus on inputs and outputs, but less focus on what happens in between.

The next two sessions sought to explore the supply and demand of applied AI ethics tools and resources. First, the demand for applied AI ethics was explored from different perspectives through a panel discussion. The panellists provided a balance of technology, product, ethics and society, and business expertise. They discussed whether there is demand for processes and tools to operationalise aspects of responsible AI development (and where it is coming from), what is currently available and the challenges involved. Following, four initiatives were presented by suppliers (full details of speakers and topics in these sessions are provided in Appendix 2.).

4.3.2 Workshop

Building on the discussions in Ottawa and the topics highlighted in day one of the London event, a full-day workshop took place on day two, based around structured, facilitated conversations that would lead to clear outputs. Over 40 experts and practitioners in the fields of AI and AI ethics actively participated. The workshop was designed to transition from big picture, high-level thinking to specific ideas and recommendations on how to move forward.

The workshop began with an exercise to imagine what an AI ethics utopia could look like. This exercise was designed to open up the participants' thinking without prescribing any themes, the discussions became sequentially more specific. Participants were encouraged to think about the current challenges in ethical AI, and consider potential opportunities for technological solutions for embedding ethical AI into product and technical development. The aim of this exercise was to succinctly map out the current landscape of demand and supply based on participants' experiences, expertise and interactions, and then narrow in on the technological solutions (current and potential). The exercise then provided the basis for the next session, where participants voted on the most important challenges in ethical AI and the most impactful technological opportunities. These were then mapped on a timeline for impact and development, and the section of society (from private to government sectors) which should be the key driver to tackle each specific challenge was identified.

From the priority areas identified by the participants, the following key themes were extracted for more thorough exploration by focus groups in the afternoon:

- Increase productivity in responsible AI product development
- Tools for privacy in practice
- Implementing and measuring compliance with standards
- Diversity in development

- Developing human agency
- AI for all

The morning sessions gathered insights into some of the important attributes of ethical AI, and the challenges that can be tackled using technology or process solutions as a key component. This led into the afternoon sessions, where the task was to develop technology product ideas in each of the themed areas. Three product ideas are provided for illustration in Appendix 3.1-3. The ideas were then used to motivate thinking about how to create an environment - a hub - to accelerate the development and adoption of such products.

Each group was asked the following questions:

- Who would need to be involved?
- Where should it be located?
- How should it be accessed?
- What business model would be appropriate and who should fund this?

4.4 Outcomes

At the outset, both events set out to answer the following questions:

What is the demand for applied AI ethics?
Is there supply to meet that demand?
What support is needed to encourage progress?

4.4.1 Demand

There is widespread demand for applied AI ethics. This is perhaps unsurprising in relation to government, academia, regulators and the subjects of algorithmic decision-making. However, for industry, a deluge of negative press[28] over the last year could be seen as evidence of the reverse - a disregard for the ethical implications of AI-driven products and services. Yet this would be an oversimplification of reality.

The demand-side event panel identified that companies of all sizes are grappling with how to build, integrate or buy value-aligned AI-driven products and services. This is motivated in part by the risk of negative press arising from careless design choices. In addition, a more positive narrative emerged relating the responsible development of AI to greater competitiveness and sustainability, and to the ability to attract and retain employees.

There is sufficient demand to have encouraged a number of commercially motivated initiatives to develop responsible AI tools in recent months, such as those auditing algorithmic fairness in certain contexts, or helping to ensure data privacy. These initiatives

[28] For example, see AI Now Institute's snapshot of significant news addressing social implications of AI and the tech industry in 2018: https://medium.com/@AINowInstitute/ai-in-2018-a-year-in-review-8b161ead2b4e

coexist with or productise non-profit (academic, open-source and community) initiatives, and are another indicator of an apparent shift in industry priorities, in favour of committing time and resources to the responsible development and use of AI.

4.4.2 Supply

A common complaint made against industry is that commitments to ethical AI are more PR than reality ('ethics washing') and are a cynical "substitute for stricter regulatory approaches" [29].

While this may be true in some cases, it is worth noting that what constitutes ethical AI and how to achieve it is far from agreed at present. Even where multiple stakeholders are coalescing around ethical principles (such as the recently published EU Commission's ethics guidelines for trustworthy AI[30], developed by a 52-member high-level expert group) there is a gap between those principles and practice: the 'how' of implementing and monitoring value-aligned technology. Building good - ethical - AI presents new challenges in what is already a fast-moving and competitive space.

Tools and methods that help embed and operationalise responsible AI practices to make it as frictionless as possible to do the right thing are necessary to turn ethical aspirations into reality. Some such tools already exist (for example, those profiled in Appendix 2) and there is active research being carried out in the field of responsible AI.

However:

- Where resources exist, they can be difficult to identify or access, or it may be difficult for non-specialists to ascertain their scope-of-use and areas of applicability
- There are areas of research and resource (white space) that correspond to applied ethics requirements that are not currently being addressed[31]

Tools and methods are a necessary but insufficient part of responsible AI development. Far from being substitutes for standards and regulation, they are complementary, being essential enablers of adherence to, and monitoring of, those requirements.

4.4.3 What support is required to encourage progress in applied AI ethics?

Reflecting the diverse stakeholders represented at these events, it was felt that similarly diverse stakeholders would need to work collaboratively to advance applied AI ethics.

[29] Wagner, B. (2018). Ethics as an Escape from Regulation: From ethics-washing to ethics-shopping? In M. Hildebrandt (Ed.), Being Profiling. Cogitas ergo sum. Amsterdam University Press https://www.privacylab.at/wp-content/uploads/2018/07/Ben_Wagner_Ethics-as-an-Escape-from-Regulation_2018_BW9.pdf

[30] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[31] Ibid. 21

- The challenges presented by applied AI ethics require expertise from many domains, including from ethicists, domain experts, lawyers, civil society and user groups, researchers, technologists, government as well as private sector companies.
- Solutions need to be developed in response to requirements, and to be challenged and tested. A shared purpose and trust are required to permit innovation, open dialogue and possible failure
- Diverse stakeholder collaboration is more likely to result in the utility and adoption - and hence, impact - of resulting advances in applied AI ethics

Given a diverse range of stakeholders, the motivations and incentives to participate and the resources available to do so will differ widely. It was suggested that those with the means to contribute should subsidise the participation of those without. Those with means include government and philanthropic funding bodies whose mandates are aligned with the goal of advancing applied AI ethics, and companies with commercial incentives to participate.

5. Future Directions

The overall aims of network relationship building activities described in this report were:

- To build a foundation for future collaboration between the United Kingdom and Canada to develop concrete proposals for translating good intentions for the responsible development and deployment of AI into practical action

Therefore, in this section, we translate the conversations and ideas forged through the network activities into a number of proposed next steps that have the potential to establish long-term partnerships with significant impact.

5.1 Creation of an applied AI ethics hub

5.1.1 What is it?

There is a need to bring together multiple parties to collaboratively define and address practical challenges in applied AI ethics. Multi-stakeholder collaboration is essential for inclusion and trustworthiness. Such a hub would implement or co-ordinate some, or all, of the following activities:

- Maintaining a directory of resources that are already available, along with an assessment of the state of their maturity, scope of application, and limitations
- Testing resources with pilot users and/or dedicated testers in a safe space
- Supporting open-source and community efforts with funding, space, developer time, promotion and channels to users
- Translating research into robust, documented, usable and accessible solutions
- Gathering evidence of demand from practitioners and co-developing solutions in response
- Development and dissemination of best practice, including the creation of an evidence base for responsible AI ROI so that practitioners can better make the case for the required investment of time and resources
- Providing physical space to showcase resources and facilitate knowledge exchange

5.1.2 How would it work?

The hub would bring together those interested to work on defining, developing and using practical resources (tools, processes, and research) for the responsible development and deployment of AI. By design, this is a diverse community which would be required to ensure representation, utility and adoption of hub outputs.

Workshop attendees suggested joining together pioneers of responsible AI and pilot groups adopting responsible AI practice, and considered international collaboration a boon. This would ensure alignment of research with practical requirements and tight feedback loops on new resources. They proposed roles that different types of organisations could play, such as providing funding, developer time or a testing ground for new resources.

Different participants will clearly have differing motivations and resourcing to participate in the hub, and this suggests a tiered membership model that could be supported with philanthropic or (bilateral) government funding.

5.2 Specific product development

The workshop activities identified a number of product ideas in response to particular needs, some of which may be worthwhile developing further. These could be developed as part of a hub, or separately. For example, one product idea is for a wiki-like website to capture best practice patterns for responsible AI[32]. (Another could be a collection of datasets for benchmarking and assessment of fairness and bias in algorithms, inspired by the original research by Mohammad, et al.[33])

5.2.1 What is it?

There is a need for some sort of easy-to-navigate and understand repository of best practices that can enable AI practitioners and policy-makers to quickly find practical solutions to specific ethical issues.

A wiki-like repository of best practices could allow a diverse community of knowledgeable people to co-create the repository with:

- Some agreed public and private sector examples of ethical AI that are recommended as exemplars and models for others to follow
- Information about common failures and near-misses
- Examples of when technology is not useful

5.2.2 How would it work?

The IT infrastructure requirements for a community website or database are well-defined. The challenges involved relate to building and supporting a community and achieving legitimacy, such as:

- How to incentivise contributions from diverse groups, especially sharing what didn't work
- How to ensure quality of content
- How to handle differences of opinion on what is ethical and in what context - is a single common set of practices possible and/or desirable?
- The need for the wiki to be trusted, suggesting some sort of independent body or partnership with an existing trusted entity

There are a number of existing wikis that have navigated analogous challenges and from which lessons can be learned (such as Wikipedia's NPOV). It is possible, therefore, to make a start by seeding the wiki with content from other organisations and/or via a barnstorming

[32] See Section 3.4.2

[33] http://saifmohammad.com/WebDocs/Saif-2019-EIAI-talk-web.pdf

event, and then expand. A number of workshop attendees volunteered to kick this process off.

5.3 #EIAI2020

5.3.1 What is it?

There is a need for a continued open discussion and knowledge-sharing relating to applied AI ethics. The success of the Ottawa and London events (‘Ethics in AI 2019’, #EIAI2019), which were both over-subscribed, suggests that a rerun in 2020 would be welcome, and would have the additional benefit of acting as a method to motivate and track progress against the aspiration of moving from principles to practice in AI ethics.

5.3.2 How would it work?

The calendar is filled with events relating to AI ethics, but a focus on *applied* AI ethics is much less well-served, while the use of a networking event to build a consensus around actionable next steps currently seems to be unique. The events could be run independently again, and funded through sponsorship. Alternatively, to achieve a wider audience or to marry up complementary content, the event series could be co-located with other established academic or industry conferences, or run as a workshop stream inside one of them.

5.4 Competition to stimulate innovation

5.4.1 What is it?

There is a need to incentivise innovation in applied AI ethics. Competition is one way to achieve this, and there have already been a number of successful examples targeted at evaluation of the ethical aspects of AI systems, such as:

- The Gendered Pronoun Resolution competition hosted on Kaggle, on behalf of the competition sponsor Google AI. This competition offered $25,000 in prizes for solutions addressing gender-bias in pronoun resolution[34].
- The NeuRIPS 2018 Adversarial Vision Challenge (Robust Model Track) pitted machine vision models against adversarial attacks. The prize was $15,000 of Paperspace cloud compute credits, while the top 20 teams in each track received $250. The competition attracted 382 participants and 1,995 submissions[35].
- The NeuRIPS 2018 Inclusive Images Challenge involved developing models that address distributional skew and increase global inclusion. This challenge was hosted on Kaggle and sponsored by Google (in partnership with NeurIPS) with a prize pot of $25,000, and attracted 109 teams[36].

[34] https://www.kaggle.com/c/gendered-pronoun-resolution
[35] https://www.crowdai.org/challenges/adversarial-vision-challenge
[36] https://www.kaggle.com/c/inclusive-images-challenge/

- Gender and race bias was incorporated as an additional evaluation criterion[37] alongside the regular performance-oriented SemEval[38] emotion prediction competition.

Competition has historically had a great impact on bringing the attention of the research community to a particular problem and in attracting new researchers to the areas. Created datasets are also a valuable resource for data-hungry machine learning fields.

5.4.2 How would it work?

Competition needn't involve monetary rewards, although there does need to be an incentive to participate.

Two approaches to competition development merit further work:

- Building stand-alone competitions in response to specific ethical evaluation criteria in different tasks and domains
- Partnering with existing performance-oriented competitions to add appropriate ethical dimensions - this would have the advantage of leveraging existing communities at the same time as establishing a common understanding that such issues should be addressed right from the start of building an AI system

In each case, there are outstanding challenges related to developing appropriate datasets and evaluation criteria and to avoid overfitting. Sponsors could be sought to cover competition overheads and, where appropriate, prizes.

[37] Kiritchenko, S. and Mohammed, M, S. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems, 2018 arXiv e-prints arXiv:1805.04508

[38] https://competitions.codalab.org/competitions/17751

6. Conclusions

This Canada-UK collaboration sought to bring together technologists, lawyers, standards bodies, academics, civil society groups, private sector companies and other stakeholders to examine how to support practitioners in translating ethical aspirations into value-aligned AI; to build technologies with positive benefits and to avoid negative consequences from the technology they develop.

The approach taken was to hold two networking and relationship-building activities: in Ottawa, Canada, and in London, United Kingdom, in February and March 2019 respectively. Measured on attendance, social media impact, connections made and feedback, these events were extremely successful. Both events were over-subscribed, and participants brought with them an abundance of goodwill and enthusiasm.

However, the most important measure of success for this collaboration was being able to move the conversation from the ‘what’ to the ‘how’ of responsible AI, specifically, producing well-defined, actionable post-project next steps with the potential to establish long-term partnerships and impact. This was an ambitious aim - it is hard to funnel and focus thinking from abstract principles to tangible ideas, and to do so in a multi-stakeholder environment - but it worked.

Section five of this report lists four recommended post-project activities. These vary in effort and impact, and therefore questions remain about resourcing and participation models. However, they represent a consensus from a broad group of stakeholders - many of whom are in a position to contribute tools, networks, research, or funding themselves - on where to focus now .

Effort should now turn to finding ways to make these recommendations a reality, exploring potential collaborators and sources of funding, and aligning with other related initiatives where they exist (such as Partnership on AI).

Future work should endeavour to preserve the collaborative and inclusive nature of these events to maximise the utility, legitimacy and impact of the relationships and connections built, and so ‘light up a dark room’.

Appendices

A1. NRC Ottawa Workshops - additional details

Day 1: Symposium Presentations, panels, and posters:

- **Keynote** - Alison Paprica, Vice President, Health Strategy and Partnerships, Vector Institute. *Social Licence, Health Data and AI*
- Anat Elhalal, Head of Technology and Tech Lead for AI and Machine Learning, Digital Catapult. *Moving the AI Ethics Conversation From the 'What' to the 'How'*
- Luciano Floridi, Professor of Philosophy of Ethics and Information, University of Oxford, directing the Digital Ethics Lab*, A message of support*
- Ann Cavoukian, Distinguished Expert-in-Residence, leading the Privacy by Design Centre of Excellence at Ryerson University. *AI Ethics by Design: An Extension of Privacy by Design to Artificial Intelligence*
- Christine Henry, Product Manager, Amnesty International. *Data Science, for Good? Adventures in Practical Ethics Implementation in the 'AI For Good' Space*
- David Van Bruwaene, Director of Research and Development at SafeToNet Ltd, CEO SafeToNet Canada. *Privacy, Machine Learning, and the Digital Parent*
- Jennifer Gibson, Sun Life Financial Chair in Bioethics, Director of the University of Toronto Joint Centre for Bioethics, Associate Professor in the Dalla Lana School of Public Health, Director of the World Health Organization Collaborating Centre for Bioethics at the University of Toronto. *AI for Health: Key Ethical Considerations*
- Graeme Hirst, Professor, Department of Computer Science, University of Toronto, Senior researcher in AI and NLP. *Ethical Issues in Natural Language Processing*
- **Keynote** - Joelle Pineau, Associate Professor and William Dawson Scholar at McGill University, Lead of Facebook's AI Research lab in Montreal, Canada. *Ethical Challenges in Data-Driven Dialogue Systems*
- Keith Jansa, A/Executive Director, CIO Strategy Council. *Key Ingredient to Implementing Ethical AI and Growing the Digital Economy*
- Petra Molnar, Lawyer, International Human Rights Program, Faculty of Law, University of Toronto. *The Human Rights Impacts of AI and Emerging Technologies: Experiments with Migration Management and Refugee Decision-Making*
- Saif M. Mohammad, Senior Research Scientist, National Research Council Canada. *Examining Fairness through Emotions in Language*
- Samantha Brown, Program Lead, Doteveryone, UK. *Responsibility in Tech Practice*
- Wilco van Ginkel, Founder and CEO, a3i. *Trust or not to Trust AI - That's the Question!*

Panel 1 - Privacy, Transparency, and Explainability in AI Applications,

- Sébastien Gambs, Canada Research Chair (Tier 2) in Privacy-preserving and Ethical Analysis of Big Data, Université du Québec à Montréal (UQAM)

- Patricia Kosseim, Counsel, Privacy and Data Management, Osler, Hoskin & Harcourt LLP, Co-Lead, AccessPrivacy by Osler
- Jocelyn Maclure, Professor of Philosophy, Université Laval, President, Commission de l'éthique en science et en technologie du Québec.

Panel 2: Governance and Accountability for AI Algorithms/Products

- Rob Davidson, Manager, Data Analytics and Research at the Information and Communications Technology Council (ICTC)
- Hessie Jones, Director for the International Council on Global Privacy and Security, by Design
- Mark Robbins, Senior Researcher, Institute on Governance (IOG)
- Sarah Villeneuve, Policy Analyst, AI + Society, Brookfield Institute for Innovation + Entrepreneurship.

Student Posters were prepared and presented by:

- Fateha Khanam Bappee, Dalhousie University. *Crime Pattern Detection and Prediction: Fidelity, Interpretability and Ethical Considerations*
- Jonathan Bowen, Western University and the Rotman Institute of Philosophy. *Non-Instrumental Reasons for Creating Artificial Persons*
- Chris Dulhanty, University of Waterloo. *ImageNet Demographics Audit*
- Atoosa Kasirzadeh, University of Toronto. *Ethics, Explanation, and Machine Learning*
- Nishila Mehta, University of Toronto. *Assessing Medical Trainees' Knowledge and Perceptions of Artificial Intelligence in Medicine*
- Victor do Nascimento Silva, University of Alberta. *Algorithms and Social Media: A Challenge to Democracy*
- Patricia Thaine, University of Toronto. *Perfectly Privacy-Preserving AI: What is it and How do we Achieve it?*
- Christine Wang, University of Toronto. *Incorporating Ethics of Artificial Intelligence Education into Medical School Curricula: A Call to Action*

Full agenda, speaker biographies, abstracts, slide decks, and minutes from the Ottawa symposium are available through the links below:

Day 1 workshop program: https://sites.google.com/view/aiethics/program
Day 2 minutes:
https://drive.google.com/file/d/1m5dmU5h2S3vgPfGAUyE4YHnM3cswKcKq/view

A2. London symposium - additional details

Applied AI Ethics demand panellists were:

- Ray Eitel-Porter, Accenture
- Christine Henry, Datakind UK;
- Louise Marston, DotEveryone

- Ben Blume, Atomico

Applied AI Ethics Supplier presentations:

- **Accenture Fairness Tool:**[39] Caryn Tan, Digital Strategy Consultant from Accenture, demonstrated the Accenture Fairness Tool, which enables a user to identify and fix some of the problems that result in unfair outcomes by analysing both training data and models. It also enables users to understand the trade-offs they may be making between model accuracy and fairness.
- **Ellpha:**[40] Stephanie Creff, CEO, spoke about the Ellpha Bias Detection Engine^{TM} that identifies gender-biased language by analysing text, providing feedback and making suggestions for more neutral language. Ellpha is currently focusing on deploying their product in the HR environment, and is undertaking research to detect bias in datasets.
- **Hazy:**[41] Alice Piterova, Head of Privacy, presented the company's work in synthetic data innovation. She addressed the new challenges and scenarios that organisations are facing in an increasingly data-reliant age. Hazy generates synthetic data that provides companies with the ability to enhance privacy, address class imbalance, act as a benchmarking tool and enable forecasting and simulation of future events.
- **OpenMined:**[42] Andrew Trask presented the work of OpenMined, a community focused on researching, developing and elevating open-source tools for secure, privacy-preserving and value-aligned artificial intelligence. Andrew described three techniques that can be combined to keep both data and models private: federated learning, secure multi-party computation, and differential privacy - all of which are implemented in OpenMined's PySyft library[43].

[39] https://www.accenture.com/gb-en/blogs/blogs-cogx-tackling-challenge-ethics-ai

[40] https://www.ellpha.com/

[41] https://hazy.com/

[42] https://www.openmined.org/

[43] https://github.com/OpenMined/PySyft

A3. Product Idea Generation workshop

In the final workshop in London, different groups spent time generating product ideas focusing on the following areas:

- Increase productivity in responsible AI product development
- Tools for privacy in practice
- Implementing and measuring compliance with standards
- Diversity in development
- Developing human agency
- AI for all

Here are three product examples, provided for illustration only (the one hour allocated was insufficient - and not intended - to produce fully stress-tested plans).

A3.1 Implementing and measuring compliance with standards

Background
The core of this offer was to provide a group of analytics tools (mainly in-app tools rather than process tools) to identify common approaches that yield undesirable results. The toolset would be built (or later tuned) to conform to any nationally or internationally agreed standards in place.

A version of the tools would be deployed in-house to address compliance. An additional proposition would be to send the product or service for review by a central organisation for approval or benchmarking, to provide independent assurance that the product or service developed is in line with required standards.

What is the product?
An auditing toolkit for continuous compliance to ensure that agreed standards have been met. Some of the tools discussed included:

- Code analyser - a review tool to identify potential hazardous code
- Content review tool - to read text or images in an application and provide feedback

Who are the users?
This tool would be used for quality assurance by product and software development teams. Consultancies and third-party auditors would also use it to ensure compliance with internationally agreed standards.

Why does it help?
Continuous compliance monitoring will offer organisations the opportunity to communicate regularly with end users and other key stakeholders, informing them of product benefits as well as risks, thereby building trust. The compliance tool would provide reassurance for companies working in high-stakes industries and make the process of compliance less

arduous. Such in app tools will also result in a faster response to errors or changes to the product.

What is the impact of the product?
The product will increase or maintain a company's reputation with customers and the general public, while aligning business needs with ethical needs. The key driver for this product would be that organisations using the tool see an increased return from being recognised as socially conscious, giving them a competitive advantage.

Does it present any challenges or opportunities?
Regulatory costs may act as a disincentive for companies to use the tool, especially if that use is voluntary. Organisations may be concerned about sharing information about proprietary services with external auditors and third-party consultancies. However, if there is an opportunity for companies to receive certification (e.g. B Corps), this can improve their reputation, and if adopted by key industry players, the tool could have a significant effect on driving an industry towards making more ethical business decisions.

Finally, an alert function for any non-compliance provides developers with the opportunity to improve their code and overall product.

A3.2 Diversity in development

Background
The idea behind the second product is to provide a way for AI-driven products to be built with and for diverse groups and requirements. The motivating use-case was an AI startup with a handful of founders, but even large organisations are unlikely to have sufficient diversity within their workforce or the structures in place for meaningful and non-exploitative discussion.

What is the product?
A diversity marketplace a platform to connect product developers with representative groups and individuals to facilitate discussion, garner advice and carry out user-testing.

Who are the users?
Buyers: developers, product teams, service providers, startups
Sellers: community and other representative groups, individuals
The platform would be used for consulting at all stages of AI development

Why does it help?
Despite efforts to increase diversity in the workforce, startups and other developer/product teams are usually not fully representative of the communities that could be affected by their design and technology decisions. Ad-hoc attempts to canvas diverse viewpoints can be time-consuming, ineffective and clumsy. A platform could help to implement best practice and reduce the cost of engagement.

What is the impact of your product?
By ensuring that diverse viewpoints are heard at all stages of planning, build, test and deployment, the resulting products and services will be more likely to meet the needs of a broad range of users and avoid negative consequences. Products and features that may not

otherwise have been built will get built.

Does it present any new challenges or opportunities?

- Uncertainty over what is best practice for engaging with diverse viewpoints, and acting on that engagement. For instance, what are the diverse groups that need to be engaged with? Is there a working minimum and what does it comprise? What questions should be asked? How should participants be incentivised or recompensed?
- In terms of accessibility, how can such a platform be prevented from exacerbating the digital divide? Would it need a non-digital component?
- In terms of recruitment and retention, any platform's utility is dependent on being able to match providers with consumers. Is it possible to acquire users in a cost-effective way? Is it possible to piggy-back onto existing platforms and community sites?
- How would the platform assure diversity and avoid identity fraud?

A3.3 Increase productivity in responsible AI product development

Background

The rationale for the third product was to develop a series of tools and resources to help reduce the amount of time needed to assess capabilities and make decisions to improve products and services. In relation to ethics in an AI environment, much of the time taken to assess and improve products is dedicated to reaching a shared understanding of the terminology that different teams and partners are using. To help address this issue, the workshop group designed a product to create a uniform process to help technical and non technical practitioners communicate more effectively with a shared language and set of definitions.

What is the product?

The product is a centralised online bank of definitions and case studies of both terminology and practical tooling related to responsible and ethical AI. This was suggested to be similar in format to a community wiki or GitHub-style repository to enable voting, contributions and collaboration. This style of tool could also eventually sit as an online front- end interface for selecting different tools and facilitate a community-sourced understanding of them. What would initially start as a repository should be developed so that it can be easily combined with other resources and tools for cross- platform communication and integration. Poor compatibility can have serious consequences for productivity.

Who are the users?

This type of product would be useful to anyone developing an AI system. An effectively implemented solution would allow product development teams, data scientists, developers and any interested parties to communicate effectively, and therefore enhance productivity and decision making processes.

The product would be used for:

- Assembling the correct team
- At the planning and initiation phase of a project and at review points to ensure ongoing adherence to a collaboratively decided plan

Why does it help?

The product would enhance collaborative working by developing a shared understanding of terms, reducing misunderstandings and avoid talking at cross-purposes. Over time, this tool could be expanded and further developed, similar to the way in which GitHubhas evolved.

What is the impact of your product?

This product will provide impact by improving cross team collaboration and improve productivity. Functionality will be key to uptake and wider industry impact.

Does it present any new challenges or opportunities?

This tool will incorporate vast amounts of information and definitions of ethical and technical terms relating toAI technology. The tool could also be used to engage clients and stakeholders in a collaborative way, through project or product design and management features.

End paper

Thanks to:
National Research Council Canada and Innovate UK, part of UK Research and Innovation

Contributors:
[National Research Council Canada] Berry de Bruijn, Alain Désillets, Kathleen Fraser, Svetlana Kiritchenko, Saif Mohammad, Norm Vinson.
[Digital Catapult] Peter Bloomfield, Hayley Brace, Katarzyna Brzoska, Anat Elhalal, Katy Ho, Libby Kinsey, Ross McWhirter, Madalina Nazare, Evelyn Ofori-Kuragu.