The Ethical Skills We Are Not Teaching: An Evaluation of University Level Courses on Artificial Intelligence, Ethics, and Society

A Report to the Social Sciences and Humanities Research Council
Knowledge Synthesis Grants Program

2021

Juan Luis Suárez, MBA, Ph.D., Ph.D. [Principal Investigator]

Professor, Faculty of Arts and Humanities

Director, CulturePlex

Western University

Daniel Varona, B.Sc.

Ph.D. Candidate, CulturePlex

Western University

The Ethical Skills We Are Not Teaching: An Evaluation Of University Level Courses On Artificial Intelligence, Ethics, And Society Is Co-Funded By The Social Sciences And Humanities Research Council And The Government Of Canada's future Skills Program.

Les Compétences Éthiques Que Nous N'enseignons Pas: Une Évaluation Des Cours Universitaires Sur L'intelligence Artificielle, L'éthique, Et La Société Est Cofinancé Par Le Conseil De Recherches En Sciences Humaines Et Le Programme Compétences Futures Du Gouvernement Du Canada.

CulturePlex
The Humanities of the Anthropocene

Western
Arts&Humanities

TABLE OF CONTENT

EXECUTIVE SUMMARY

Canadian universities are not teaching the necessary ethical skills for future workers in the Artificial Intelligence (AI) sector to safely, productively, and effectively engage with Automatic Decision-Making Systems (ADMS), i.e., AI and Machine Learning (ML). This training gap is consistent across all programs evaluated across 16 countries and risks the future of the burgeoning AI industry in Canada. This gap also exposes future workers to the unintended performance of several harms (bias, discrimination, unfairness, privacy breakings, etc.) from their companies and the humans who use or are the subject of their ADMS. The identified lack in training also hinders the efforts of our government and civil society to extend Canadian values of equity, diversity, and inclusion within the digital realm.

The report undertakes a review of 503 courses on non-functional issues of AI evaluated across 16 countries including Canada, to identify the skills used to train future workers. Our methodology takes a mixed methods approach with Natural Language Processing deployed to read a database of texts made up of course descriptions and syllabi, whereas close reading, bibliographical reviews, and analysis of policy documentation are utilized to set up the background of the evaluation and highlight the urgency of a coordinated effort to regulate this training across Canadian universities and colleges through Quality Assurance/Quality Improvement (QA/QI) mechanisms.

To support future workers as they grapple with the ethical complexity around AI issues (including the damage to basic human rights) and to provide practical skills that they can use in their working environments (industry, government, or NGOs), we propose to build the regulatory framework to develop these skills around the trustworthiness of ADMS. Trustworthiness has been gathering momentum over the last few years in industry forums and policy documents internationally as the superior concept to translate the complexities of AI into reflective, but actionable tools and skills that humans can critically use when building, using, or managing ADMS.

KEY MESSAGES

- According to our textual analysis, universities around the world (503 courses on non-functional issues of AI evaluated across 16 countries and 66 universities) are not teaching their students the ethical skills needed to prepare them to effectively and successfully to engage with ADMS i.e. AI and ML, either as developers of software, managers whose organizations sell products or services with ADMS components, or as users of those systems.
- In most courses (84.69%), the learning outcomes are poorly described and do not differentiate between the knowledge, values, and skills students will learn about and/or acquire. This lack of skills descriptions questions the ethical skills and principles students will bring into the job market and how they will deploy their learning when engaging with ADMS.
- The "notions of skills" uncovered by our analysis, found within the 503 course descriptions and syllabi, revealed that these notions cover a wide spectrum of terms and concepts (see below section on Most Representative Topics). Our analysis demonstrates minimal semantic similarity overlapping across course descriptions within this sample. One can conclude that universities have

not yet achieved a common ground of ethical skills for future workers in an AI-based economy that can serve as the standard for the education and training of our students.

  - The most relevant topic across the majority of courses in the sample is broadly framed by the "Ethical and Social Aspects" of ADMS technologies in settings which favor the use of "Critical Thinking" as the main learning method. This topic is followed by discussions centered in the role of "Data" as a main element influencing people's ethical and social compass. Examples in these courses are usually drawn from health care and robotics.
  - Most of the identified skills notions converged towards "Responsibility" as an object of discussion. "Responsibility" partially creates a subnetwork of topics and skills aimed to "Produce Solutions with Ethical and Social Impact" and for the "Design" of ethical systems. These course descriptions do not provide indications as to how future workers will acquire and deploy "Responsibility" in their engagement with ADMS.

- The complexity of ADMS, the multi-faceted philosophical and ethical dimensions of AI systems, and the risks and threats that ADMS pose for companies, civil society, and governments, indicate that discussions about the "Responsibility" of future workers and "Critical Thinking" are not solid enough strategies to guarantee the effectiveness of those workers in an AI-based economy, the safety of users, customers and citizens, or to foster the possibilities of success of Canadian companies in this very competitive space.
- A regulation of non-functional AI courses in universities and colleges through already established QA/QI mechanisms is urgently needed. This regulation would standardize the ethical skills future AI workers across industries, organizations, and governments will need to protect their organizations from unintended harm, uphold legal standards related to AI, promote Canadian social values of equity, diversity and inclusion, and stop the chain of bias and discrimination that affect equity-seeking populations and ADMS tend to perpetuate and amplify.
- We propose that these basic ethical skills are taught using the notion of "Trustworthy ADMS" when training future workers. "Trustworthy ADMS" are those which foster trust of AI users towards both products and development methods. An ADMS with quality integration of ethical elements such as privacy protection, robustness, or security is considered trustworthy. With respect to the software development method, trustworthy ADMS also results from the insertion and evaluation of ethical dispositions as part of the QA activities in the project's life cycle.
- We define the "ethical skills" of students and future workers using ADMS in relationship to the trustworthiness of the ADMS they engage with and, hence, as the set of learned abilities that will allow workers to perform the ethical actions required to build, safeguard, and protect the trustworthiness of those ADMS in the design of the product, the development of the software, and the management of the services they provide.
- Trustworthy ADMS should strengthen Canada's statistical, data science, artificial intelligence, and ML aided decision-making processes in every sector. It should also considerably improve the Canadian software industry's chances to lead the growing technology market, producing human-centered and environmentally sustainable ADMS products.

FULL REPORT

1. Background

1.1. Introduction

This section of the report focuses on the skills developers and users in work-related environments deploy to both engage and develop AI technologies which are trustworthy, unbiased, and non-discriminatory. We aim to identify skills that will allow those workers to ethically engage with AI systems. The ultimate goals of this exercise are: 1) to help provide future workers with the proper training and education to ethically engage with ADMS; 2) to contribute to a national AI ecosystem which upholds Canadian social and ethical values and contributes to an international consensus about values embedded in AI technologies and machine learning within the digital economy. We consider it crucial to identify these skills in both economic and social terms while building a digital society that is reliant on ADMS in a moment when "automation, artificial intelligence and other emerging technologies" are signaling large scale "changes in demand for skills in many sectors" and "social and emotional skills will be crucial to success" in an economy driven largely by digital and machine technologies (Antal et al., 2018; Government of Canada´s Future Skills Program, 2020).

AI is one of the most feared technologies emerging from digitization and it also comes with a set of new opportunities both economic and ethical in the second machine age including the emergence of a new, or transformed, work force that will have to deal with the challenges of automation and digitization (Brynjolfsson & McAfee, 2014). AI questions who we are as human beings across most dimensions of our personal and social existences, and it also puts a mirror in front of our ethical decision processes and the consequences they have for citizens, workers, and entrepreneurs. The perfect storm of AI (cheap parallel computation, development of Big Data techniques, and better algorithms) has led some experts to declare that AI will help us define humankind and is needed to tell us who we are as human beings (Kelly, 2016). This dialogue between our humanity and AI is happening in a slow and sometimes obscure way, at the level of digital ethics and, specifically, at the intersection of ethics and AI (Bucher, 2018; O'Neil, 2017). However, as necessary as the philosophical discourses on the perils and moral dilemmas of AI and ML are, it is the moment for our society to take a step further and translate these philosophical discourses into actionable, reflective, and built-in skills that our 21st century work force can use to navigate their relations with systems whose architecture (or parts of their architecture) are driven by AI, and specifically ML algorithms. Many of the jobs being created in the digital economy, but also most of other traditional jobs in the private, public, and non-for-profit sectors with any minimal intersection with digital and data systems will require some use of ML algorithms. Given the fast and wide adoption of these algorithms and ADMS in general across society and the economy, it is crucial that we parallel such expansion with the creation of a work force that is both literate and skillful on AI matters, and that can act on their ethical knowledge about artificial intelligence when developing, using, or managing such systems. Establishing such a trained work force will become critical to our success as a country and our stability as a society in a moment in which the COVID-19 pandemic has created a "magnification of existing inequities within the labour market" (Government of Canada´s Future Skills Program, 2020).

In recent years, most geopolitical powers have developed national strategies for AI that will guide the course of their economy, education, trade (Adès, Diaz & Russell, 2019) and international relations (Suárez, 2018). However, few governments have been as exhaustive (at least in the published documents) as the US government in describing the importance of having a workforce properly trained in the skills of AI, including ethics, for the future of the country. In 2016, the U.S. National Science and Technology Council (NSTC) published two important reports, Preparing for the Future of Artificial Intelligence and The National Artificial Intelligence Research and Development Plan, which delineated the contact points between government and AI in relation to public good and government effectiveness. These contact points included government regulation of AI; the connections between research and the creation of a large enough body of workers in the area; the economic impact of the automation of processes and jobs; the domains of equity, security and governance and their connections to justice and equality as pillars of democracy; and all aspects of security, defense, and geopolitics. The NSTC experts proposed to focus on seven priority lines: a) long term investment in research; b) development of effective methods of collaboration between humans and artificial intelligence; c) deepen our understanding and face the social, legal and ethical implications of artificial intelligence; d) guaranteeing the security of artificial intelligence systems; e) develop both contexts and datasets that are public and shared for training and testing; f) assess and evaluate artificial intelligence technologies; g) and improve our understanding about the work related needs and skills that an economy based on artificial intelligence will require. These reports also gave two recommendations to the American government: the creation of a framework to implement strategic lines a) to f), and the development and maintenance of a work force that can help implement g).

This report contributes to our understanding of the work-related needs for an economy based on AI. We assume that for this transformed workforce to be successful and effective, different levels of skill ethics will be incorporated (Grundke et al., 2018) in all seven strategic lines: a = general ethical skills; b = trust between humans and artificial intelligence; c = all around, deep ethical training; d = privacy, equity, fairness; e = equity, diversity, inclusion; f = deep ethics knowledge of algorithms.

To this end, we: 1) unpack and discuss the main ethical and social issues across the AI literature; 2) review the description and learning outcomes of numerous ethics and AI university courses to extract and identify the notions of skills used and the topics covered; and 3) analyze the notions of skills for ethical engagement with AI used to train students incorporating current knowledge in the domain.

1.2. Ethical and Social Issues in Automatic Decision-Making Systems

AI, through machine learning, seeks to mimic the natural learning processes existing across nature. It differs as automated learning is mainly based on a set of examples that algorithms use to learn from, instead of a set of rules. As with humans, ADMS can provide predictions and recommend decisions that are discriminatory to individuals or groups. As AI and other ADMS solutions are present in almost every aspect of the social, political, and economic fabric of today's economy and society, using ethics fundamentals to avoid discriminatory predictions and decisions is crucially important.

Verma and Rubin, and Mehrabi and others describe "discrimination" as the direct or indirect relationship between a protected attribute and the resulting prediction with a negative consequence for the decision object (Mehrabi et al., 2019; Verma & Rubin, 2018). They expand this definition by declaring that indirect discrimination manifest itself when implicit effects of protected attributes and other attributes that are not protected are considered. For example, the use of an individual's zip code in loan applications or in insurance policy premium calculations are examples showing how seemingly insensitive attributes can lead to a discriminatory decision. According to Zhang and others, residential areas have a representation of their inhabitants closely related to attributes such as race, income, etc. (Zhang, Wu & Wu, 2017).

However, the zip code (to continue with the same example) is not usually a protected attribute in the decision-making process because the law does not register it as a feature triggering discriminatory feature, like race or gender. In Jiahao and others, a set of attributes are identified which should be protected in an attempt to avoid discrimination in the aforementioned scenario and others, such as recruitment (Jiahao et al., 2019). Also, the statistical root of discrimination can be identified when the information learned about a group is used to judge an individual with similar characteristics. Hence, data and data collection procedures according to the scope of the intended decision or prediction is critical. The continued use of statistical methods in decision-making and/or the use of predictions might lead to a systematization of discrimination. It can therefore be understood that ML has scaled the impact of discrimination, "unintentionally institutionalized" these discriminatory methods through AI and other ADMS solutions and created a perpetual cycle where the object of discrimination itself becomes part of the knowledge base used in subsequent estimates. Another study justifies the use of these unintended discriminatory ADMS by providing two main reasons: 1) the model can provide a decision/prediction according to the need of the business; and 2) the lack of a less discriminatory alternative model (Schmidt, Siskin & Mansur, 2019). In fact, this promotes an attitude of resignation and the acceptance of discrimination and a subsequent bias as part of our economy and social interactions.

The specialized literature shows a tendency to hold ML algorithms accountable for the problem created by their inability to adequately deal with bias (Varona, Lizama-Mue & Suárez, 2020); however, the data used in training and the data collection methods are equally responsible for discriminatory predictions and recommendations. Therefore, discrimination has an origin and is also the cause of bias.

There are two main methodological trends in the studies aiming to optimize how ML identifies and eliminate bias. The first trend is aimed at algorithm calibration (Chouldechova, 2017; Feldman et.al, 2015; Fish, Kun & Lelkes, 2016; Hardt, Price & Srebro, 2016; Pedreshi, Ruggieri & Tur, 2008; Solon & Selbst, 2016; Zafar et al., 2015), while most recent trends (Holstein et al., 2019; Varona, 2020; Varona, Lizama-Mue & Suárez, 2020; Veale, Van Kleek & Binns, 2018) are trying to tackle the problem from early stages of AI algorithm/model design. Among the reports within the principled AI international framework is the UNI Global Union 2017 report "The Future World of Work" which describes bias as the action of using attributes like gender, race, sexual orientation, and others as elements of discrimination in a decision that is somehow harmful to humans (Fjeld et al., 2020). Then, the difference with "discrimination" is that "bias" represents the action while discrimination manifests itself in the result of using certain attributes in the decision-making process. Similarly, the G20 report highlights the existence of two types of sources for bias: the method, either in the design of the algorithm or in the way the data is collected; and in the distortion/corruption of the data used as the training basis for the model/algorithm (Abreiu et al., 2018).

The obligation of fairness defined by the Access Now Organization (2018) and The Public Voice Coalition (2018) suggests the existence of two benchmarks for the definition of bias in AI: 1) the statistical reference, expressed as the deviation of the prediction in contrast with the event´s actual occurrence; and 2) the social reference, from the evidence of statistical bias within the data representing a social bias. Also, it recognizes that decisions/predictions reflecting bias and discrimination should not be normatively unfair. It further clarifies that the single evaluation of the outcomes (algorithm calibration) is not enough to determine the fairness of the algorithm or model (Varona, Lizama-Mue & Suárez, 2020; Varona & Suárez, 2021). Consequently, the Access Now Organization and The Public Voice Coalition proposed the evaluation of pre-existing conditions in the data that can be further amplified by the ADMS before its design.

The House of Lords Select Committee on Artificial Intelligence and Martinho-Truswell and others have criticized the learning method development in ML, specifically how data is used during training (House of Lords Select Committee on Artificial Intelligence, 2018; Martinho-Truswell et al., 2018). The House of Lords Select Committee on Artificial Intelligence noted that if training datasets are unrepresentative, then the resulting identified patterns which systems are designed to spot will reflect those same patterns of prejudice, and consequently will produce unfair decisions/predictions. Martinho-Truswell and others highlighted that good quality data is essential for the widespread implementation of AI technologies, adding that if the data are wrong, poorly structured, or incomplete, then there is potential for ADMS to recommend unfair decisions. All four reports defined bias based on misleading decisions that are the product of such compromised datasets.

IBM has presented a set of unconscious bias definitions expressed in terms of their presence among the general population that AI designers need to be consciously aware of when developing ADMS solutions in order to produce a fair product (IBM, 2019). This adds the human factor to the algorithm-dataset dichotomy implying the so-called "soft skills" that need to be acquired by the ADMS developer and understood by all professionals making decisions based on these systems. It is our opinion that these skills are to be considered "hard" and formalized through a proper regulatory frame.

Several studies define fairness as the ability of ADMS to treat all similar individual or groups equally, and as the inability of ADMS to produce harm in any possible way (Mehrabi et al., 2019; Sahil & Rubin, 2018; T20; UNI Global Union, 2017; Vatican, 2020). The Indian National Strategy for AI report locates the issue of fairness at the forefront of discussion in academic, research, and policy fora, which merits both multidisciplinary dialogue and sustained research to come to an acceptable resolution (NITI Aayog, 2018). The report suggests the approach of identifying the in-built biases and assess their impact, to therefore find ways to reduce the bias until techniques to bring neutrality to data-feeding ADMS solutions, or to build ADMS solutions which ensure neutrality despite inherent biases are developed.

Mehrabi and others indicate it is crucial to understand the different kinds of discrimination that may occur given the numerous distinct available definitions of fairness (Mehrabi et al., 2019). As most experts are moving away from the reactive approach (Chouldechova, 2017; Hardt, Price & Srebro, 2016; Solon and Selbst, 2016) traditionally followed in ML to cope with bias and discrimination towards a more proactive style, the focus has moved from fairness (as a non-functional requirement) into trustworthy AI as a business model concept.

To achieve trustworthy AI, the High-Level Expert Group on Artificial Intelligence recommends enabling inclusion and diversity throughout the entire AI system's life cycle and involving all affected stakeholders throughout the process (High-Level Expert Group on Artificial Intelligence, 2019). Along with Abolfazlian, both studies describe three components trustworthy AI should comply with throughout the system's entire life cycle: 1) it should be lawful, complying with all applicable laws and regulations; 2) it should be ethical, ensuring adherence to ethical principles and values; and 3) it should be robust, both from a technical and social perspective, as even with good intentions, AI systems can cause unintentional harm (Abolfazlian, 2020). Similarly, Abhishek and others propose other three main components trustworthy AI should include: 1) ethics of algorithms (respect for human autonomy, prevention of harm, fairness, explicability); 2) ethics of data (human-centered, individual data control, transparency, accountability, equality); and 3) ethics of practice (responsibility, liability, codes, and regulations) (Abhishek et al., 2020). This is consistent with an attempt to harness unintended discrimination from law and policymaking, specifically from the International Law of Human Rights. The principled AI framework presented in Fjeld and others gathers a global effort to establish a set of policies and guidelines led by principles as a methodological reference when designing AI (Fjeld et al., 2020). Despite the progress that this mechanism might represent from the legal point of view, it is insufficient as a methodological mechanism manageable by ADMS designers given their background, and the language discrepancies among legal jargon and the software profession (Varona, 2020a,; Varona, 2020b; Varona & Suárez, 2021).

Predictive algorithms are used in crime prevention and justice management, emotions analysis, crowd management, classifiers, and selection processes (Hardt, Price & Srebro 2016; Varona, 2018; Zemel et al., 2013), identification of violent behavior, criminal potential (Sait Vural & Gök, 2017), suicidal tendencies (Ayat et al., 2013), and the use of automated estimation of people's sexual orientation in designing tailored marketing strategies (Walker, 2017). Given the implications of the sustained development of AI over the last few years and its widespread application in many social, legal and governmental domains, it is crucial that decision support systems are no longer conceived as a "set of transparent techniques and methods" that consume certain input parameters, to be later processed in arriving at certain estimations, but as complex systems, which are governed by an undecipherable network of rules usually called black boxes and used to obtain a given result (Varona, Lizama-Mué & Suárez, 2020). The principled AI international framework defined in Fjeld and others draws attention to a set of guiding principles towards using responsibility and transparency when creating and using AI (Fjeld et al., 2020).

Finally, in relation to privacy there are two main approaches to dealing with privacy in algorithmic contexts: algorithmic privacy and privacy by design. Algorithmic privacy is currently recognized as the way others have access and control our personal information (Jens-Erik, 2019) as it is assumed that the diffusion of algorithm-based products and services (Fast & Jago, 2020) may erode people's ability to care about, and protect, their privacy. It also describes a trend in dealing with data and algorithmic privacy resulting from a focus on a "benefits and not harms" moto that sees privacy related costs only after use has started and assumes that privacy loss is inevitable. Privacy by design (Cavoukian, 2011; Gurses, Troncoso & Diaz, 2011; Information and Privacy Commissioner/Ontario Canada & Registratiekamer The Netherlands, 1995) is a framework that calls for privacy to be considered throughout the whole engineering cycle so, that human values are well defined and considered during the whole process (Billard & Baptiste, 2019; Olszewska, 2019). A new proposal has emerged that intends to act as a bridge between both the local data privacy (Abadi et al., 2016; Albarghouthi & Hsu, 2018) and general data privacy (Dwork & Roth, 2014; Mazloom & Gordon, 2018, Johnson, Near & Song, 2018) models under an oblivious differential privacy (ODP) model that will help to design algorithms which enjoy the privacy guarantees of both local and global models where 1) data is collected, stored, and used in an encrypted form and are protected from the data collector; and 2) the data collector obtains information about the data only through the results of a DP-algorithm.

The complexity of the debates around the ethics of AI and the unintended consequences of the technologies involved call for a detailed and nuanced training for AI developers and users in job related environments that will guarantee their safety and help them uphold legal, social, and ethical values which need to be negotiated through the implementation and use of those systems in many contexts. It is crucial that future workers who will engage with ADMS in different capacities receive a formalized training using a set of agreed upon ethical skills which equips them with the expertise to properly design and develop those systems, evaluate the decisions of the systems in terms of their effect on equity deserving groups and their alignment with Canadian legislation and societal values, and prevent unintended harms on people and their organizations.

1.3. Use and Development of Ethical Automatic Decision-Making Systems in the Canadian Context

Canada has committed to facilitate the use and development of fairer and ethical ADMS, as shown by research in sectors like environment (Grasso et al., 2020), health care (McCradden et al., 2020), and administrative law (Scassa, 2020).

In the context of environment, Grasso and others demonstrate how the combination of algorithmic accountability frameworks and domain-specific codes of ethics help answer calls to uphold fairness and human values, specifically in domains that utilize machine learning algorithms (Grasso et al., 2020). In their study, they discuss their experience applying algorithmic accountability principles and frameworks to ecosystem forecasting, focusing on a case study to forecast shellfish toxicity in the Gulf of Maine. They adapted existing frameworks such as Datasheets for Datasets and ModelCards for Model Reporting, redirecting the focus of these methods towards personally identifiable private data to include public datasets, often used in ecosystem forecasting applications, to audit the case study. Their method showed that high-level algorithmic accountability frameworks and domain level codes of ethics complement each other when promoting more transparency, accountability, and fairness in ADMS, helping to avoid many of the unintended consequences that can result from deploying "black box" systems to solve complex problems.

Within the health care domain, McCradden and others suggest that taking a patient safety and QI approach to bias can support the quantification of bias-related effects on ML (McCradden et al., 2020). They believe it is necessary to use ethical principles to adequately quantify the impact of bias and reduce the potential of an ML tool to exacerbate inequalities; while arguing that patient safety and QI lenses support the quantification of relevant performance metrics, to minimize harm while promoting accountability, justice, and transparency. Consequently, they identified specific methods for operationalizing principles of nonmaleficence, relevance, accountability, transparency, and justice. They also suggested a set of practices to be adopted by health care institutions and regulators who embrace these ethical principles for the delivery of ML-assisted health care.

Similarly, according to Scassa, the adoption of automated decision-making techniques and technologies in government is a growing trend that raises concerns about fairness, transparency, and accountability (Scassa, 2020). This study also showcases the emergence of issues related to ensuring fairness and accountability in the governance of automated decision-making. This is especially important considering the prevalence of privacy, data protection, and transparency as matters of interest in most research studying government and administrative law. However, long-standing principles of administrative law have shaped the adoption of administrative decision-making processes and guided oversight of their fairness and accountability. Administrative law can provide an important lens used to assess the use of automated decision-making by governments. The paper explores the role and shape of administrative law principles in an era of automated decision-making and assesses the Directive on Automated Decision-Making (DADM) adopted by Canada's federal government in 2019, through an administrative law lens, to determine at what extent the Directive meets the principles of administrative fairness.

There are also several studies which have investigated the AI research and development field of study in the Canadian context. Varona explores the current approaches to address the fairness issue by means of bias and discrimination within the machine learning's fundamentals (Varona, Lizama-Mue & Suárez, 2020). The study underlines the inefficacy of algorithmic calibration and stresses the limitations of the philosophy of protecting attributes, among other elements. Kasirzadeh calls for caution with the use of counterfactuals when the facts to be considered are social categories such as race or gender (Kasirzadeh & Smart, 2021). The study reviews a broad body of papers from philosophy and social sciences on social ontology and the semantics of counterfactuals and concludes that this approach in ML fairness and social explainability can require an incoherent theory and definition of the social categories. Its findings suggest that often the social categories may not admit counterfactual manipulation, and hence may not appropriately satisfy the demands for evaluating the truth or falsity of counterfactuals. This is particularly important given the extensive use of counterfactuals in ML which can lead to misleading results when applied in high-stakes domains. The study states that even though counterfactuals play an essential part in some causal inferences, their use for questions of algorithmic fairness and social explanations can create more problems than they resolve. Accordingly, the authors propose a set of tenets for the use of counterfactuals for fairness and explanations in ML. Lastly, Govia communicates the potential spaces of mediation for anthropological practices identified by means of interviews and field observations performed when exploring the sociotechnical entanglement and ethical discussions around AI (Govia, 2020).

These studies share a common determination of finding mechanisms to achieve ethical performance and good practices related to their specific domain which reinforces the importance of determining the necessary skills that future practitioners must acquire to behave ethically and professionally in contexts involved with ADMS.

1.4. Policies and Institutions

Among Canadian policies and institutions addressing ethical issues in AI, the following stand out because of the impact this thought leadership has had internationally.

Amnesty International and Access Now have led The Toronto Declaration as a landmark statement in protecting human rights in the age of AI (Access Now Organization, 2018). The Declaration has been widely endorsed by the global human rights community. It calls on governments and companies to urgently protect human rights in the age of ML, AI, and advanced computing, with a focus on the right to equality and non-discrimination. It also proposes that human rights law and standards are put front and center in existing and emerging conversations and methods analyzing the impact of ML and related technologies.

Similarly, the University of Montreal published a Declaration for Responsible AI with the objective of developing an ethical framework for the development and deployment of AI, guiding the digital transition so society benefits from technological evolution, while also opening a national and international forum for discussion to collectively achieve equitable, inclusive, and ecologically sustainable AI (Université de Montreal, 2018).

There are other important AI research hubs like the Vector Institute, in Toronto; Mila, in Montreal; and AMII, in Edmonton that receive support from the Canadian Institute for Advanced Research (CIFAR) and its AI Chairs to help implement the Canadian AI-related policies and guidelines in practice, which have led the country to becoming a global leader in AI-related topics.

Our report aligns with existing consensus about the impacts of robotization, and the automation of the labor market identified by Cséfalvay that focuses "on strengthening the comparative advantages, the creativity, and the social intelligence of humans that robots will never be able to match" (Cséfalvay, 2019). We consider that in addition to the benefits that a proper training in AI ethics will have for a more efficient and equitable digital economy, it will also alleviate the bottlenecks of automation (tasks that are difficult to automate and require: social intelligence; cognitive intelligence; and perception and manipulation (Nedelkoska & Quintini, 2018)) by providing discursive and symbolic frameworks which will help workers to cognitively and socially bridge the gap between the digital and analogue aspects of our current digital condition (Suárez, 2020) and in turn reducing the levels of alienation and anxiety associated with that gap.

2. Analysis

2.1. Methodology

We gathered data from universities that distinguish themselves in computer sciences. To do so, we used the Guide2Research[1] Ranking for Computer Science & Electronics, where universities are sorted based on the sum of the h-index, and DBLP values of their currently affiliated scholars as gathered by May 10th, 2021 (2021).

The rank is based on a detailed examination of more than 6300 computer scientist profiles on Google Scholar and DBLP (Guide2Research, 2021a; Guide2Research, 2021b). The h-index threshold for being considered a leading scientist was set to 40, where most of their publications are in computer science and indexed in DBLP. In the construction of the Guide2Research, scientists' affiliations data was captured on their Google Scholar profile.

We gathered course descriptions, course goals, and learning outcomes from 66 universities in North

[1] The Guide2Research Ranking for Computer Science & Electronics is one of the leading portals for computer science research providing trusted data on scientific contributions since 2014, to support students with their career choices.

America, Europe, and Oceania ranked in top positions in the Guide2Research Ranking for Computer Science & Electronics. Course information was captured from the university portal, from the Computer Science, Software Engineering, Philosophy, Social Sciences, and Arts and Humanities faculties, departments and/or specialized colleges. Course information was gathered in English when available, and information in languages other than English was not captured for this evaluation. The inclusion criteria focused on gathering courses that were not only focused on the traditional functional approach but included soft elements such as privacy and other human-centered features in technology (from an ethical perspective), ethical impact of AI and intelligent technology, and social impact of ADMS solutions.

All these documents were stored in text files to be analyzed with several Natural Language Processing (NLP) tools. The study combined Natural Language Processing techniques like semantic similarity, n-gram extraction, and topic modeling combined with more traditional textual analysis techniques such as close reading, to identify notions of skills across the analyzed data. For this evaluation, we consider a "notion of skill" as the triangle formed by the several independent sentence parts in which a learning outcome or a goal, in a Course Description or a Syllabus, can be divided. These triangles (notions of skill) are formed through the co-occurrence of a verb, noun, and adjective in a same sentence, were the verb express an ability, the noun represents the skill's object, and the adjective shapes the scope/characteristic of the notion of skill.

The dataset was text-based, with the course description, stated goals and declared outcomes, which were filtered removing the stop words like prepositions, articles, pronouns, among others, so that repeated headings, footers, and margin notes resulted in a consolidated and semantically robust corpus. The data was processed using Python, a generic and modern computing language, widely used for text analytics (Oliphant, 2007; Python Software Foundation, 2021). Python's development environment is enriched with libraries like Gensim (Rehurek & Sojka, 2010), that was used for topic modelling; SciPy (Virtanen et al., 2020) tools including Pandas (McKinney, 2010), used for structuring the data; Matplotlib (Hunter, 2007) and Seaborn (Bisong, 2019), used for data visualization; IPython (Pérez & Granger, 2007) used for interactive computing and programming; and NLTK (Loper & Edward, 2009) used for word tokenization, entity recognition, and semantic similarity comparisons.

The data was processed using NLTK searching for n-grams ($1 \leq n \leq 3$), which is a technique that weights and ranks words combinations determining their value for the text based on their frequency. This analysis helped to understand from a macro perspective the content and scope of the courses.

The semantic similarity analysis was conducted using the sklearn v0.24.1 module, specifically the cosine similarity described by Pedragosa and others (Pedragosa et al., 2011), which instead of using the Euclidean distance between the two vectors representing word frequencies use the cosine of the angle formed by the two vectors. This is a practical technique to evaluate the closeness of different text corpora and discourses regarding their content when different language structures are employed, something that is particularly important when working with documents authored by people with different interests, backgrounds, languages, and geographies. No word embedding was considered in the semantic similarity analysis since we only took the word's stem into consideration regardless of its context.

We also performed topic modelling using the Gensim's implementation (Blei, Ng & Jordan, 2003) and it was applied following Latent Dirichlet Allocation (LDA) to detect topics among the courses' descriptions. The LDA is a generative probabilistic model in which each document is considered as a finite mix over an underlying set of topics. Each topic is represented as a set of words and their probability, which means that it is possible to rank topics on the corpus and the keywords in each topic. The technique of topic modelling helped to corroborate some inferences resulting from other tasks of the analysis regarding the context and scope of the courses.

Extra care was put into the topic modelling part of the analysis as, among the techniques used in the study, it is the one with an additional intrinsic uncertainty. One of the most important issues related to topic modelling with LDA is knowing the optimal number of topics (k) that should be examined. In consequence, different LDA models with variable values of k ($5 \leq k \leq 25$) were built, computed the coherence for each topic, and selected the model with the highest coherence value. The best results were found with eight topics and 38 keywords per topic, presented in the Results and Discussion section.

We also built a network based on the identified notions of skills and used Gephi v0.9.2 to evaluate different centrality measures like closeness centrality, to identify most recurrent abilities, scopes, and skills objects relevant across courses, to study the betweenness centrality measure, to identify notion of skills vertices with a bridge function among the multiple identified communities; and the eigen vector. This added a vicinity's relevant to the analysis, when determining nodes that are connected to other relevant ones in the network so further conclusions are drawn regarding the highlighted skills.

The network analysis described in the previous paragraph was complemented by means of close reading so we could contrast our findings which were based on notions of skills, with the skills stated in the courses' goals and outcomes in the pursue of additional specific and comprehensive understanding.

2.2. Results

2.2.1. Descriptive Analysis

The study includes 503 courses offered across 16 countries and 66 universities. The distribution of courses across countries seems to cluster these countries in different brackets (see coloration in Table 1). The USA is in a category by itself both in the number of universities collected in the sample and in the number of courses (361) offered in these institutions. A second bracket is made up by Canada, Switzerland, UK, and Singapore, with a range between 19 and 32 courses in the sample. It is important to note that the US, Canada, UK and Singapore (National University of Singapore) provide instruction mostly in English for the disciplines in this study. A third bracket includes countries whose universities use both English and the local languages, what might affect why their number of courses is lower. Finally, the bracket comprising countries with the lowest numbers of courses offered in this domain, includes English-speaking countries like Australia and India (for the purposes of scientific and technical university teaching) and others that mostly use their national languages).

Country	QTY	Country	QTY
USA	361	Belgium	5
Canada	32	Australia	3
Switzerland	24	Denmark	3
UK	20	France	2
Singapore	19	India	2
Hong Kong	12	Australia	1
Germany	10	Italy	1
Sweden	7	Netherlands	1
Total			**503**

Table 1 Distribution of courses by countries

Another element to note is that there seems to be no relation between the type of university system involved in the offering of these courses (public vs private; continental Europe vs UK), and the number and content of the courses their universities offer in the domain of AI, ethics, and society.

The columns G2RWR2, G2RNR3, refer to the Guide to Research University Ranking for computer sciences; World, and National rankings, respectively. The guide to research is an online community formed by graduate students and scholars providing verified data on scientific contributions since 2014 to help students in their study and career choices. These rankings are dedicated to distinguishing universities with computer science study programs based on the h-Index citation value and DBLP value of their scholar corpus, gathered from their Scholar and DBLP profiles by May 10th, 2021.

University	G2RWR	G2RNR	QTY
Carnegie Mellon University	1	1	42
University of Texas at Austin	15	14	38
University of Washington	11	10	24
University of California, Irvine	21	18	21
MIT	2	2	20
University of Waterloo	22	2	20
University of Pennsylvania	24	19	20
University of Southern California	10	9	17
Cornell University	19	16	15
Princeton University	20	17	14
ETH Zurich	16	2	13
University of Wisconsin-Madison	25	20	13
Georgia Institute of Technology	5	5	12
Johns Hopkins University	35	23	12
National University of Singapore	26	1	12
EPFL : École polytechnique fédérale de Lausanne(University of Lausanne)	7	1	11
University of Cambridge	43	5	11
Stony Brook University	62	35	10

Table 2 Universities offering 10 or more courses

The traditional university gap between teaching and research is reflected also on the data on Table 2, as there is no apparent direct relation between the university scholar's scientific production (in terms of number of published papers and citation index) and the number of courses a university offers in the domain of ethics, society, and AI.

The Carnegie Mellon University, and the University of Texas, Austin, separate themselves from the other universities leading the global landscape with 42 and 38 courses respectively, followed at a distance by the 24 courses offered at the University of California, Irvine, and the 20 courses found in MIT, Waterloo, and Penn.

91.1% of the 503 courses we analyzed were offered as part of 475 different study programs at undergraduate and graduate levels, while the remaining courses were offered as separated optional courses. That is, there is almost a 1 to 1 ratio between courses and programs, a spread that seems to confirm the exploratory character of this emerging domain that we have also detected through other analytical tools (see below). Of those differently named study programs, 7.95% of all the courses in our sample can be attributed to Engineering Ethics; Artificial Intelligence, and Ethics and Policy Issues in Computing comprise 6.13% of all courses each; Introduction to Software Engineering has 5.69% of the courses; Computer Science, 4.16%; Social Implications of Computer Technology, 3.72%; Computers, Ethics, and Public Policy, 3.50%; Computers and Society, and Computer Security offer 3.28% of courses each; Advanced Computer Security and Privacy, and Artificial Intelligence Methods for Social Good have 2.84% of courses each; Human-Computer Interaction offers 2.63% of the course, and Professional Ethics 2.41% of them.

In institutional terms, many different disciplinary centers offer programs and optional courses on ethics, society, and AI. The spread of disciplinary centers on campuses includes 52 schools, faculties, colleges, and departments. The most common are: Schools, Colleges, and Faculties of Engineering (12.3%); Schools, Colleges, and Faculties of Informatics (9.2%); Schools, Faculties, and Departments of Computer Sciences (8.71%); Schools, Faculties, and Colleges of Humanities and Social Sciences (5.64%), specifically through their Department of Philosophy; and School of Continuing Education (3.18%).

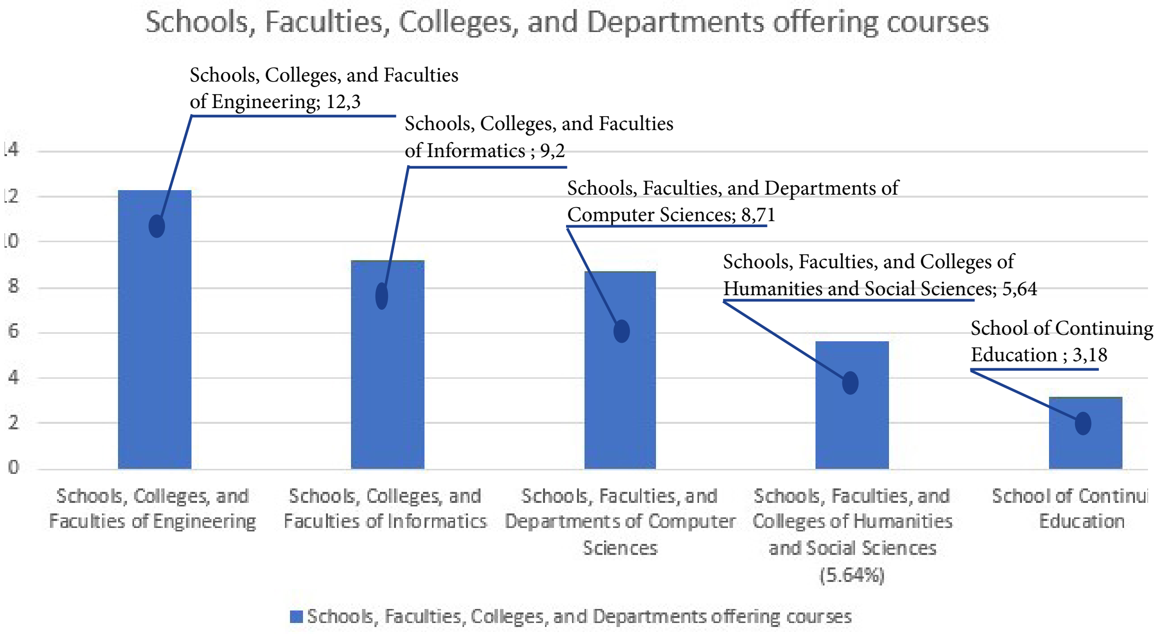


Figure 1 Distribution of course offering across the sample

A total of 478 (95%) courses in the sample provide course descriptions, while 74 (14.71%) courses provide a list of specified goals, and 76 (15.11%) provide a list of specified learning outcomes, aside from the course description. Amongst the sample, there are a scarce 46 (9.15%) courses providing course description, goals, and learning outcomes. The analyzed corpus consists of a text dataset formed by 31,430 words, from which 23,087 (73.46%) belonged to course descriptions, 3,627 (11.54%) enunciations of course goals, and 4,716 (15%) belong to the learning outcome descriptions. These low numbers are indicative of extremely grave pedagogical inconsistencies and absence of rigor, as the lack of structured information on the goals, outcomes and skills students will acquire reduce the students' ability to chart a course for their proper training in ethical skills and AI. It will also limit student abilities to apply these skills within their future career trajectories.

2.2.2. The Course Contexts and Domains

The context in which the courses are framed can be inferred using the n-grams extraction natural language processing technique as the saturation of high-frequency used terms provide the linguistic indexes that the author pragmatically uses to turn the attention of readers towards a sphere of reality. A brief portion of the isolated n-grams describing the courses' summaries can be seen in Table 3.

The table shows the unigrams reaching a relative frequency greater than 0.50 units, conditioning the cut at the top 16 n-grams . The relative frequency represents the ratio of an n-gram's absolute frequency with respect to the sum of the total frequency of all extracted n-grams. It allows further analysis of the n-grams as they absolute frequencies are normalized through their relative frequencies. The n-grams presented in Table 3 represent one third of the extracted n-grams. The extended analysis, which is not displayed herein, extends to the top 50 n-grams with relative frequencies as low as .24 units, for the case of unigrams; .04 units for the bigrams; and .02 units for the trigrams.

The n-grams analysis shows that the courses mainly focus on aspects related to general aspects like "data", "technology", "science", and "social" if we consider unigrams with relative frequencies of .98 and greater. That focus expands including a second block with "engineering", "computer", "systems", "issues", and "design" in a range of relative frequencies with an amplitude of 0.07 units. This second block differentiates from a third block by that same number of units, comprising terms like "ethical", "society", "policy", "security", "privacy", and "information". The extended analysis lists "ethics", "impact", and "challenges" in the 23rd, 36th, and 37th positions with relative frequencies of .44, .29, and .27 units, respectively.

The bigrams column shows that the evaluated courses are normally linked to the development (of software) of ADMS, with terms like "machine learning", "artificial intelligence", "data science", and "big data", occupying top positions in the table. This means that future engineers and coders are being taught about topics such as "public policy", "technology society", "social sciences", and "ethical issues" in addition to "computer sciences" and "computer engineering". Additionally, bigrams like "case studies", "real world", and "problem solving" may provide a hint about the practical hands-on approach the courses are designed to follow. The bigrams extended analysis places "software engineering" and "science engineering" in less favoured positions, 46th and 49th respectively, with relative frequencies as low as .04 units, which might represent some lag on an Engineering perspective with respect to other areas within Computer Science, and Humanities and Social Sciences in addressing the needed ethical character of automatic decision-making technologies, ethical development practices, and the negative social impact of this kind of technologies.

Unigrams	Abs. Freq	Rel. Freq	Bigrams	Abs. Freq	Rel. Freq	Trigrams	Abs. Freq	Rel. Freq
Data	294	1.27	machine learning	65	0.28	electrical computer engineering	13	0.06
Technology	257	1.25	computer science	55	0.24	science technology studies	10	0.04
Science	247	1.11	artificial intelligence	55	0.24	human centered design	10	0.04
Social	226	1.07	data science	44	0.19	human computer interaction	10	0.04
Engineering	180	0.98	security privacy	39	0.17	science technology society	9	0.04
Computer	177	0.78	case studies	38	0.16	obtain hands experience	7	0.03
Systems	175	0.77	public policy	34	0.15	humanities social sciences	7	0.03
Issues	168	0.76	decision making	29	0.13	science technology medicine	7	0.03
Design	165	0.73	intellectual property	29	0.13	participants obtain hands	6	0.03
Ethical	148	0.71	computer engineering	29	0.13	public policy issues	6	0.03
Society	143	0.64	technology society	23	0.10	machine learning data	6	0.03
Policy	141	0.62	social sciences	23	0.10	legal ethical issues	6	0.03
Security	135	0.61	ethical issues	20	0.09	ai robotic technologies	5	0.02
Privacy	131	0.58	real world	19	0.08	large scale data	5	0.02
Information	126	0.57	big data	19	0.08	information security privacy	5	0.02
Human	125	0.57	problem solving	16	0.07	centered design engineering	5	0.02

Table 3 Summary of n-gram extraction to the course summaries[2]

[2] We have considered "students" a stop word, and then eliminated it from the sample, for this analysis as most courses' description and syllabi use turns of phrase such as "Students will learn…", "Students will be familiarized with…", etc.

The sampled courses can be clustered in four main fields of study: Electrical and Computer Engineering; Science and Technology Studies; Science, Technology and Society; and Humanities and Social Sciences, according to the following trigrams "electrical computer engineering", "science technology studies", "science technology society", and "humanities social sciences". The most recurrent content across the surveyed courses can be described through the trigrams "human centered design", "centered design engineering", "human computer interaction", "information security privacy", "public policy issues", and "legal ethical issues". Trigrams like "obtain hands experience" and "participants obtain hands" help to justify the courses' practical hands-on approach, with emphasis on "machine learning data" and "large scale data"; while using mostly examples and case studies from "science technology medicine" and "ai robotic technologies".

2.2.3. Semantic Similarity Across Courses

As shown in Figure 2, 25% of the courses in the sample exhibits up to 32% semantic similarity. When increasing the number of analyzed courses to 50% of the courses in the sample, the value of the semantic similarity increases by ten percentile points, reaching a maximum of 42%. If instead of analyzing half the sampled courses we analyze 75% of them, then the semantic similarity values rise another nine percentile points up to a 51%.

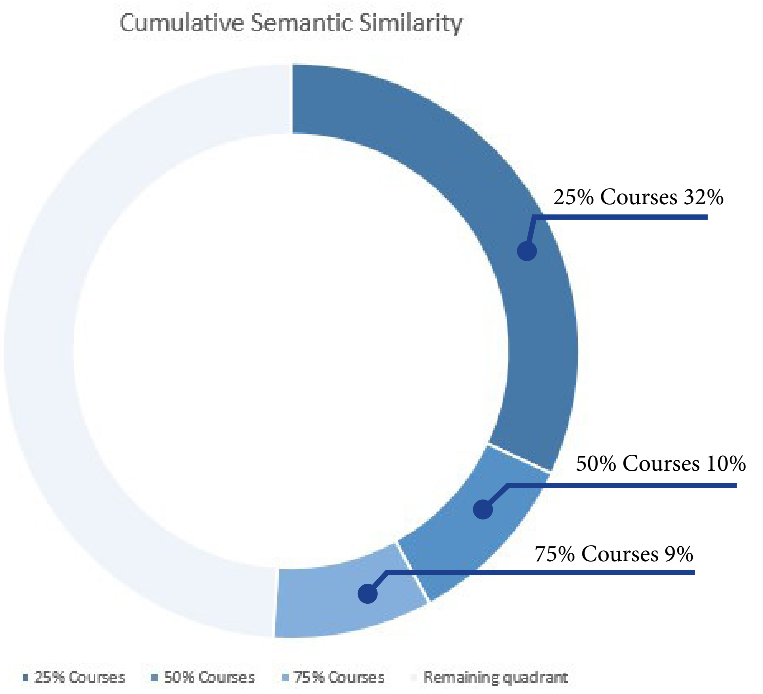


Figure 2 Semantic similarity distribution at 25%, 50%, and 75% courses in the sample

These values indicate a low semantic similarity across the course descriptions. It would be inappropriate to draw conclusions regarding the abilities and content of the courses solely based on the results of the semantic similarity metric, however, its low value might be an indicator of the multiple edges the topics linking technology and ethics can be addressed from. This low semantic similarity across courses may also be an indication that this is a burgeoning area of study, and the ways in which the content of these various courses grapples with these complex issues may illustrate the various forms in which the field is emerging.

Also, the low semantic similarity hints at the lack of consistency in the approach to creating learning outcomes that we observed and described above. At the same time, low semantic similarity across courses also points at the lack of consensus and common ground on the knowledge and skills that students and future workers of AI need to acquire.

2.2.4. Most Representative Topics Across Courses

Before identifying the most recurrent topics among the sample we first determined the optimal number of topics to be extracted using LDA's coherence measure[3]. The first significant jump in the coherence measure (0.3397) was achieved at eight topics with 30 words. From that point on, the coherence measure slightly raised between the values of 0.3554 and 0.3759, while the number of topics experience an exponential growth of 14, 20, 26, 32, and 38. These results inclined us towards the inclusion of the eight most relevant topics:

Topic 1.“technology”(0.035) + “society”(0.024) + “student”(0.024) + “science”(0.023) + “course”(0.019) + “broad”(0.016) + “social”(0.016) + “issue”(0.015) + “ethical”(0.014) + “be”'(0.011)

Topic 2.“datum”(0.052) + “healthcare”(0.041) + “ai”(0.036) + “use”(0.021) + “analytic”(0.020) + “learn”(0.020) + “student”(0.018) + “problem”(0.018) + “technique”(0.018) + “practical”(0.018)

Topic 3.“engineering”(0.028) + “research”(0.018) + “be”(0.018) + “software”(0.018) + “skill”(0.016) + “science”(0.015) + “able”(0.012) + “discipline”(0.011) + “societal”(0.010) + “system”(0.010)

Topic 4.“human”(0.056) + “design”(0.047) + “system”(0.040) + “computer”(0.031) + “interaction”(0.023) + “include”(0.022) + “study”(0.019) + “user”(0.018) + “topic”(0.017) + “social”(0.017)

Topic 5.“engineering”(0.029) + “paper”(0.020) + “function”(0.019) + “problem”(0.015) + “ability”(0.014) + “student”(0.014) + “provide”(0.012) + “make”(0.012) + “design”(0.012) + “use”(0.011)

Topic 6.“technology”(0.029) + “how”(0.021) + “cryptography”(0.020) + “encourage”(0.020) + “home”(0.019) + “mediate”(0.018) + “active”(0.018) + “social”(0.018) + “science”(0.016) + “course”(0.014)

Topic 7.“security”(0.076) + “privacy”(0.043) + “forensic”(0.030) + “usability”(0.024) + “policy”(0.023) + “system”(0.022) + “network”(0.021) + “include”(0.020) + “trust”(0.019) + “public”(0.018)

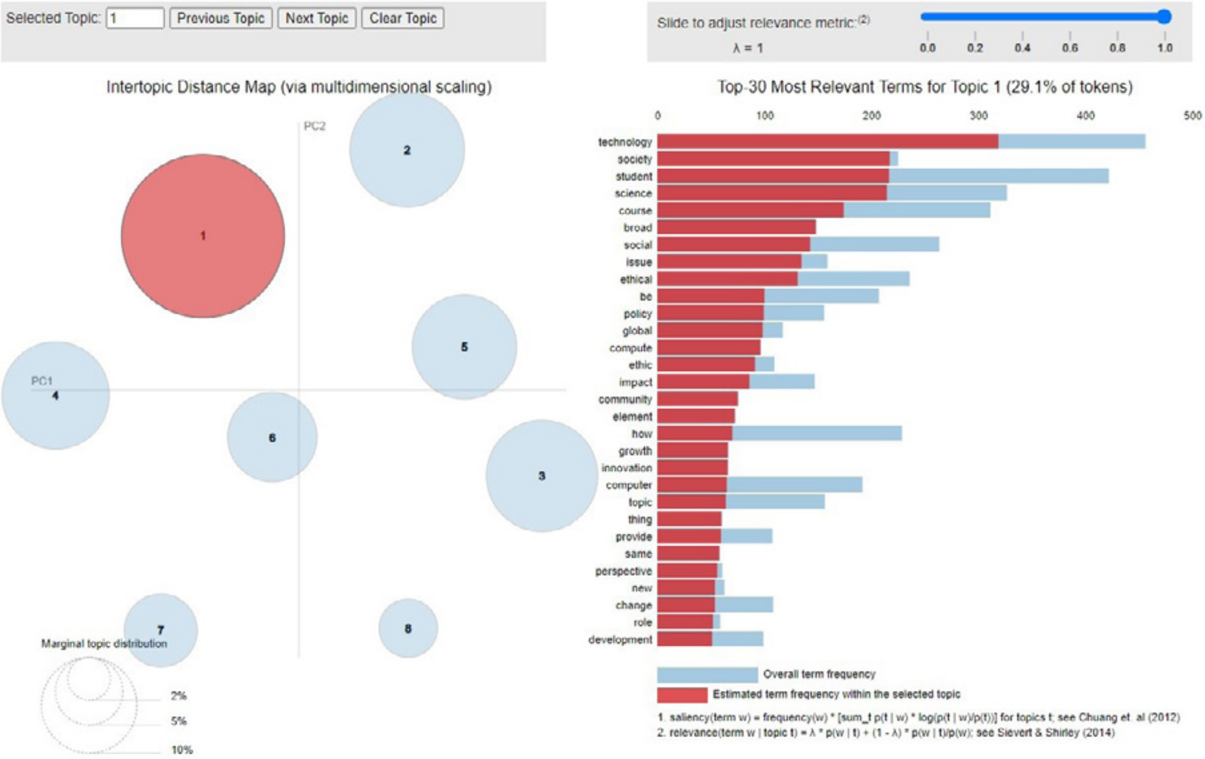
Topic 8.“datum”(0.064) + “data”(0.025) + “module”(0.024) + “analysis”(0.019) + “database”(0.015) + “large”(0.015) + “scale”(0.015) + “literature”(0.012) + “web”(0.010) + “method”(0.010)

[3] The Latent Dirichlet Allocation (LDA) coherence measure uses the frequencies of a set of terms co-occurring in semantically similar text blocks. This measure is useful to determine the minimum number of topics with higher coherence value

To provide a summarized view of the content of the topics, the eight listed topics included 10 terms each, out of the up to 30 terms that are relevant for each of them. The topics are sorted from highest to lowest according to their representativeness for the surveyed courses, as are the terms (tokens) listed within each topic. Figures 1 to 8 below show a comprehensive view of the content and relative size of the tokens making up each of the topics, along with a set of comparisons between topics in terms of the semantic intertopic distance properly adjusted by multidimensional scaling.

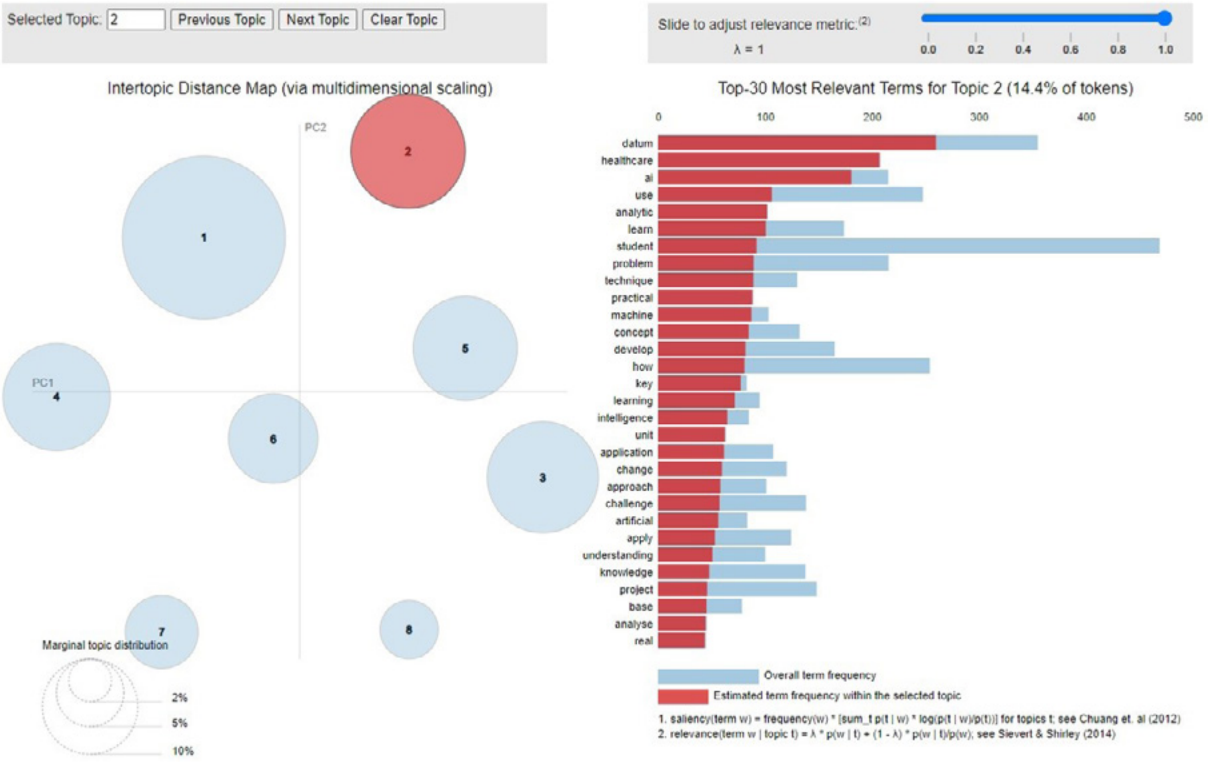
Some of the identified topics are particularly consistent with the n-gram analysis presented earlier. Topic 1, the most representative across the courses, highlights the scope of the courses in terms of technology, society, social, issue, and ethics. However, the content of Topic 1 also highlights the generality with which the courses approach ethical and social issues around AI. Topics 2 and 8 highlight the data driven character of the courses and the role that analytics (of big data) plays in the conceptualization of ethical issues in this domain. The remaining topics suggest there is a clear division between the objects of study (technology, health care, design, cryptography, or policy) targeted by the courses, and the fact that students will acquire certain abilities in the courses (skill, able, practical, learn, make, use, active), although the specific nature and name of those abilities seems to be missing from most of them.

Except for three of the tokens in Topic 7 (security, privacy, and trust) none of the other seven topics make any reference to any of the concrete ethical issues that make part of the philosophical and governance debates around AI and ADMS, the policy papers written around these issues across the world and analyzed in Section 1 of this report, or to the scholarly production around Canada's implementation of ADMS in government and public service systems.



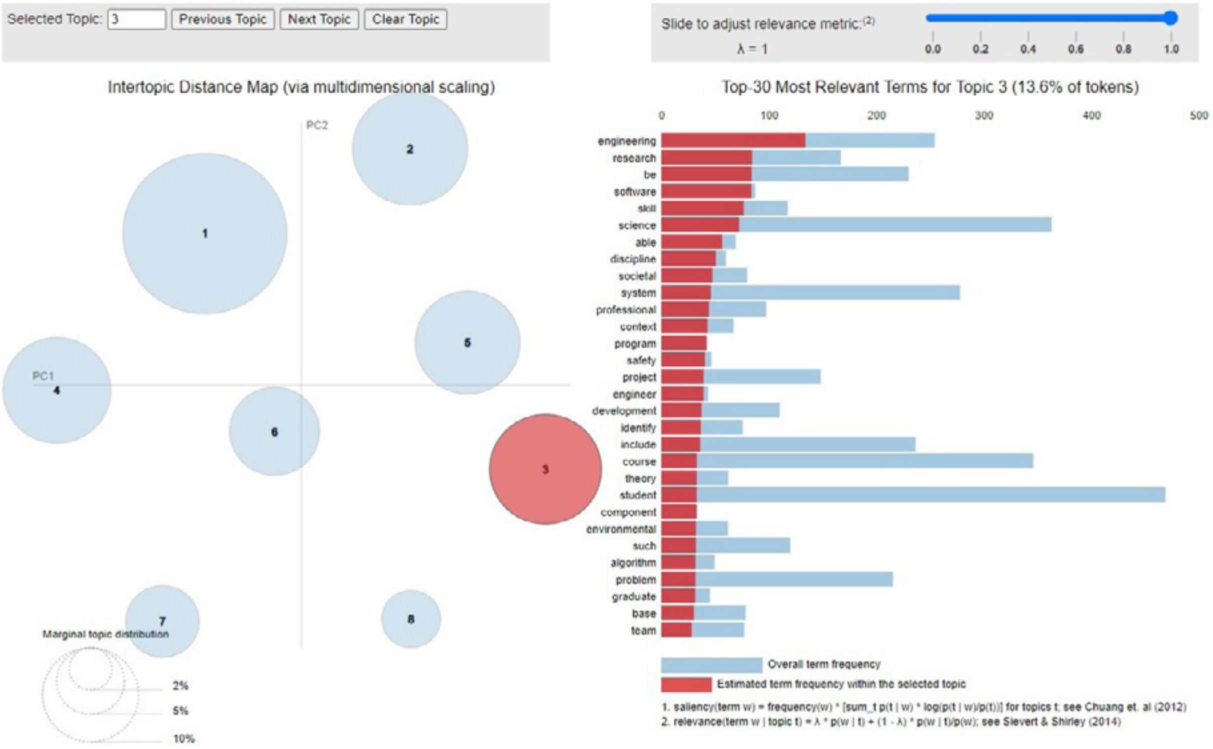
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3 Relevant terms distribution for topic one



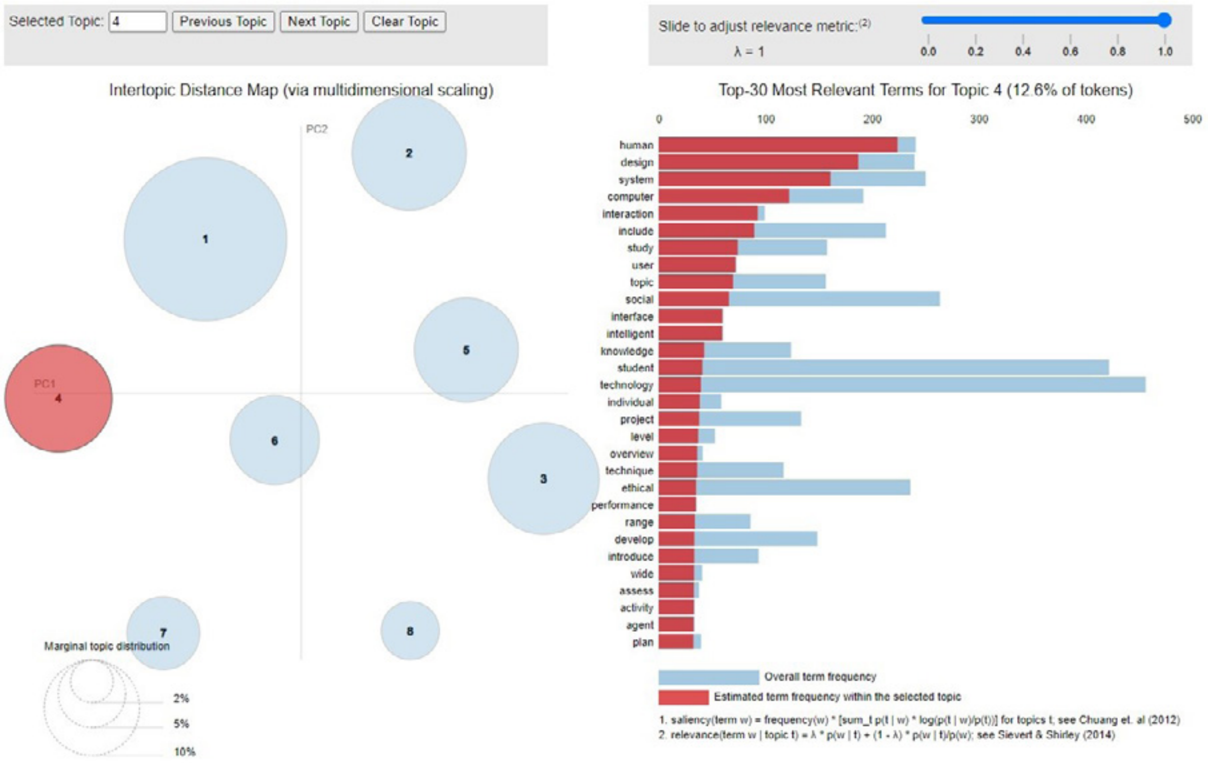
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 4 Relevant terms distribution for topic two



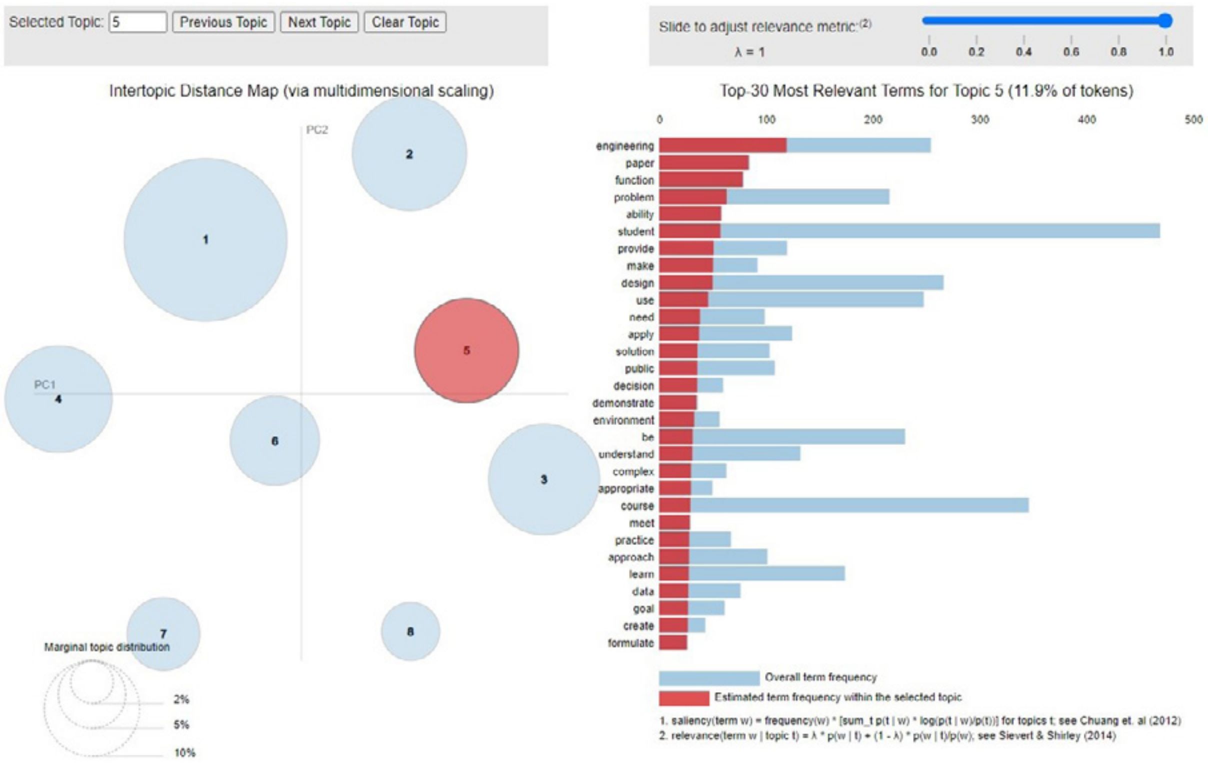
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 5 Relevant terms distribution for topic three



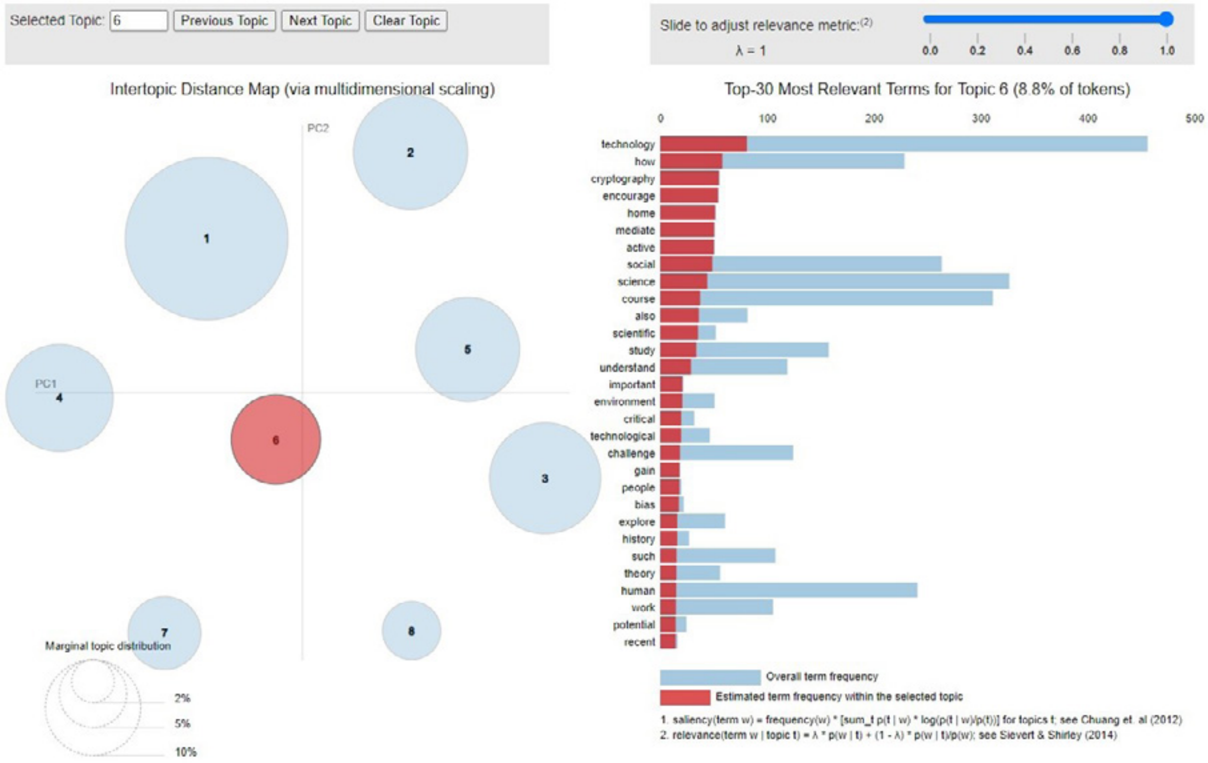
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 6 Relevant terms distribution for topic four



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 7 Relevant terms distribution for topic five

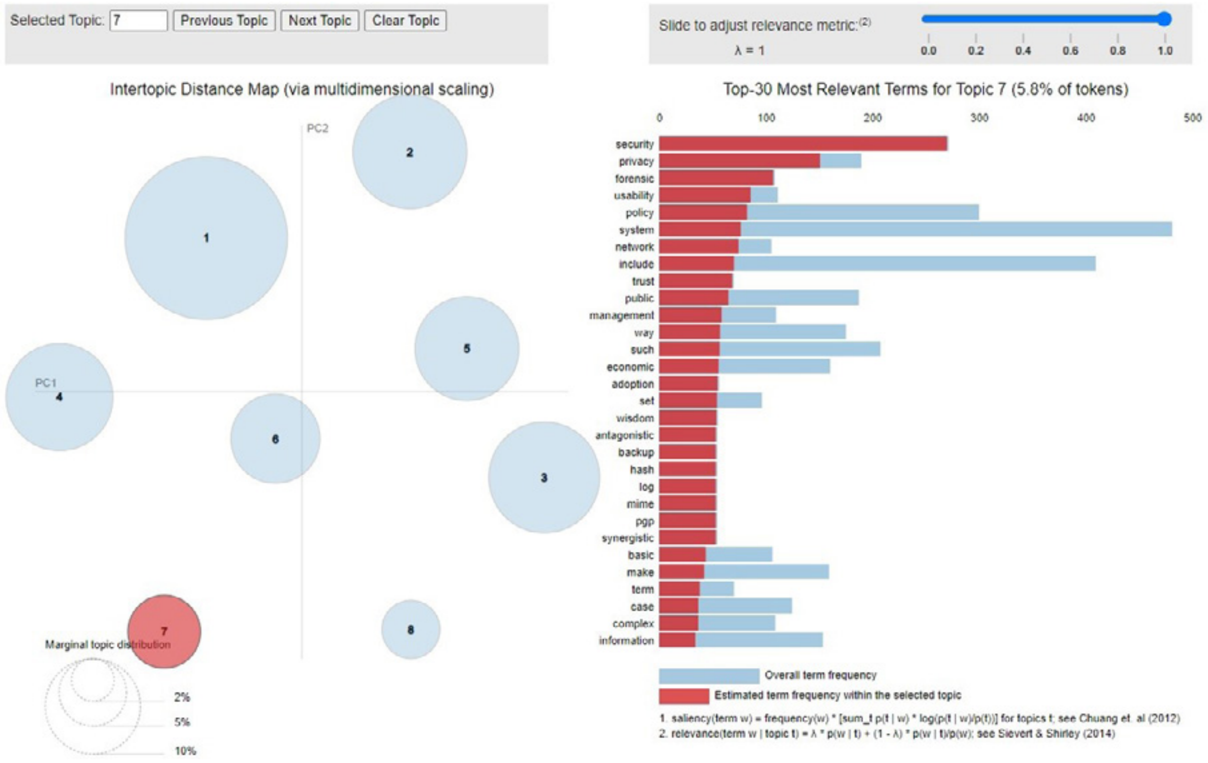


Selected Topic: 6

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 8 Relevant terms distribution for topic six

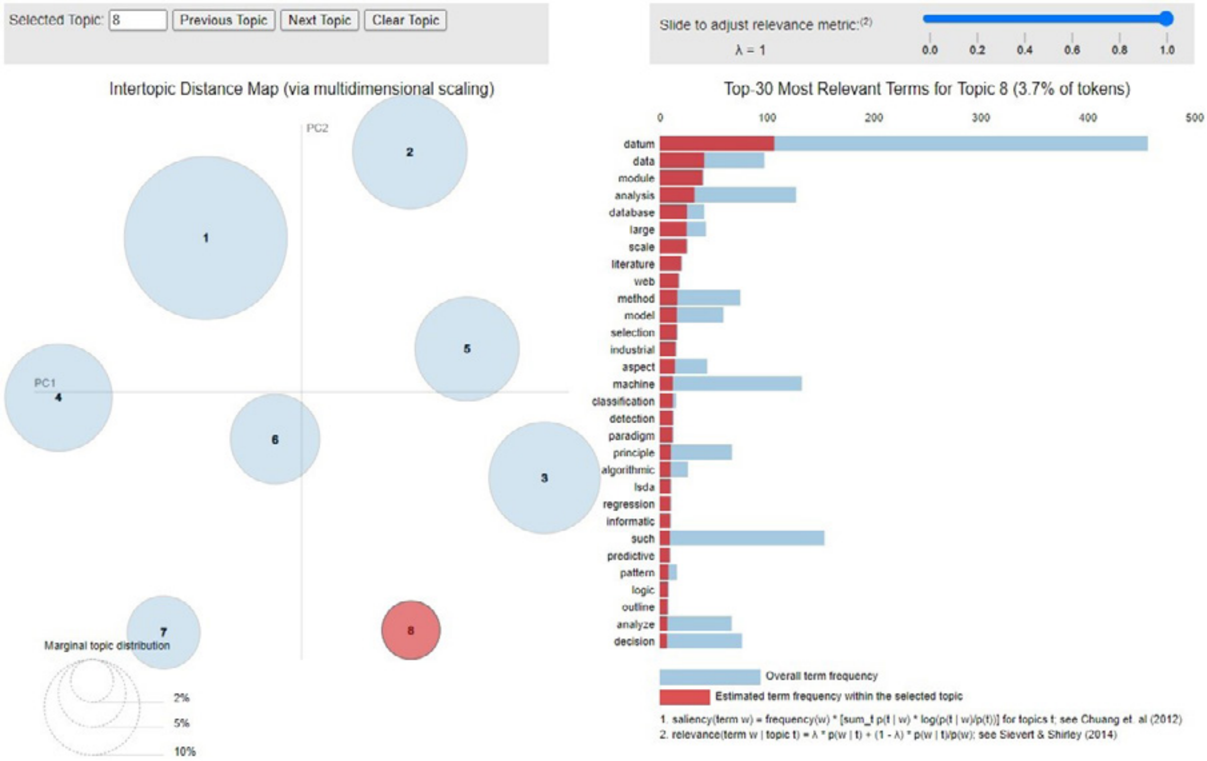


Selected Topic: 7

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 9 Relevant terms distribution for topic seven



Selected Topic: 8 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 10 Relevant terms distribution for topic eight

2.2.5. Course Goals and Outcomes.

To analyze the notion of skills described across the analyzed dataset, we narrowed down our focus from the course description to the declared goals and outcomes of the courses. A network was built based on nodes representing three kind of elements (verbs, adjectives, and nouns), where the edges connect nodes co-occurring within a same sentence. We assume that verb-type elements describe ability, while noun-type elements represent the object's abilities are focused on, and adjective-type elements define the scope of pairs formed by a given ability and a given object in the determination of notions of skills.

There is a total of 919 nodes and 15,346 edges in the network, establishing 258,501 triangles each representing a notion of skills by means of the co-occurrence of its vertices within a same sentence. Each of the vertices of a triangle represents one type of nodes, and no types are duplicated in a single triangle. In addition to providing a notion of skills across the settled goals and outcomes, the identified triangles help determine the relative relevance of the skills and allow for further comparisons.

There are ten communities identified in the network. Communities one and three gather around a third of the nodes in the network with 17.85% and 14.36%, respectively. Community three is also the community with more triangles, exhibiting 20.52% of the triangles in the network. This means a single community gathers one fifth of the total number of identified notions of skills. Along with communities two and seven, that include 7.83% and 9.25% of the total of nodes and are ranked 6th and 7th according to their size, the number of notions of skills identified in these three communities reach almost half the number of the notions drawn in the network (60% including community one which is ranked 4th in number of triangles). Communities two and seven, being smaller communities, exhibit a greater density of connections, hence they group more notions of skills than communities three and one.

Community two is the community with higher density value, and it is formed by notions of skills represented by triangles whose vertices include object-type elements like "students", "ability", "impact", "process" (in a noun function) "context", and "solutions"; scope-type elements such as "ethical", "social", "design" (in a descriptive function), "environmental", "global", "societal", "professional", and "relevant"; and verb-type elements as "apply", "design", "demonstrate", "engineer", and "meet; to provide examples of vertices linked to more than 1,000 triangles.

After evaluating the closeness centrality measure to explore the nodes that are relevant to the network based on how they bring the referred notions of skills together, we found the following nodes exhibited the highest centrality in the network: noun-type nodes "responsibilities", "systems", and "solutions", in that exact order; adjective-type nodes "general" and "social" in that same order; and the verb-type objects "test" and "produce". They all differ form the next node with higher closeness centrality by an entire quartile length, specifically 295,925 units. The elements "responsibilities", "test" and "produce" where represented by the nodes with the highest score. The verb-type nodes "test" and "produce" highlight the practical hands-on character of most identified notions of skills within the community.

The nodes representing the adjective-type elements "ethical" and "social", and the verb-type elements "design" and "engineer" are the top scoring using the betweenness centrality metric, which means they act as relevant bridges in the community two's frontier with other communities. Additionally, the Eigen metric pointed the elements "students", "ethical", "social", "ability", "design", "apply", "impact", and "demonstrate", in that order, represented by nodes connected to other relevant nodes within the community. The previous statements reinforce the initial assumption regarding the practical character of the identified notion of skills in the network. The multiple possible combinations of the identified notions of skills can illustrate skills that are relevant across the analyzed courses like "Produce solutions with ethical and social impact", "Test solutions social impact", and "Design" ethical systems, to provide some examples.

Like community two, community seven exhibits a practical hands-on character. Curiously, community seven is a community where the 1st and 2nd quartiles sorting the identified triangle's distribution include scarcely six vertices from a network of 85 nodes, denoting a degree of independence between the identified triangles. Among the most connected vertices can be found verb-type elements like "learn", "acquire", "issue", and "shape"; and the noun-type elements "technology", "course", "machine", "aspects", and "AI", connected as part of more than thousand notions of skills.

"Acquire" is represented by the node with the lowest value of the closeness centrality metric distancing from the next lower by 39,518 units. This may be the result of the style used to write the goals and outcomes linked in this community. When extending the analysis to the remaining nodes in the 1st quartile other abilities like "endow", "discern", and "drive" appear along with nouns like "behavior", "participant", "observer", "AI" and "course", leading us to believe the community gathers a set of skills that were redacted using a style that places the focus on the student rather than on the object of study. A close reading of the goals and outcomes supported this hypothesis when we found several items described using the "By the end of this/the course the students must be able to…", and "Students are expected to… " structure.

The study objects "technology" and "machines" distinguish themselves from the rest of vertices performing as bridges with other communities at the frontier of community seven. The ability "learn" is represented by the node connected with most relevant nodes within the network. It could be said community seven is characterized by the methodological approach of its stated goals and outcomes.

When analyzing community three, we observed that the nodes performing as vertices for 1,000 and more triangles represent mostly nouns and adjectives. Together with the two abilities in the group, this looks like a community that gathers a set of notions of skills exhibiting a general and theoretical character. The objects co-occurring with more scopes and abilities in a same sentence, within community three, are "knowledge", "science", "scientific", "society", "computer", "ethics", "mathematics", "program" and "area". The scope linked to more triangles are "basic", "necessary", and "development" – with an adjective function-; and the ability "study". These skills are included within the first and second quartile of triangles in the community.

Interestingly, other abilities like "debate", "reference", "engage", "prototyping", "prepare", and "synthetize" figure in the 4th quartile of identified skills within the community. The node representing the skill object "knowledge" is the closest node to all other nodes in the community exhibiting a closeness centrality value of 54, apart from the node with the second smallest closeness centrality measure by 41,075 units. Along with the node representing the skill's object "science", these two nodes exhibit the greater relative connectivity in the community's subnetwork, separating from the remaining nodes by a minimum of 7,987 units, which may indicate those two terms are not only common to most of the identified notion of skills within the community but serve as bridges to several notions in the delimitation of a skill. Lastly, the skill's object "knowledge" and the skill's scope "fundamental", according to the Eigen metric, are the nodes connected with other nodes with relevance to the network. When including other nodes in the 1st quartile of the of the Eigen measure within the community, other terms like "society", "basic", and "computer" appears. Consequently, it might be inferred that the abilities within community two can be linked to multiple possible combinations of terms addressed in the previous paragraph in the construction of skills, like in this example: "Study basic fundamental knowledge on society and computer."

Lastly, we found the ability-type elements "use", "implement", "develop", "include" and "describe"; scope-type element "appropriate"; and object-type elements "data", "level", and "skills", connected as part of more than thousand notions of skills in community one. The verb "orient" is represented by the node with least closeness centrality in the network, among other terms like "hardware", "webpage", "grasp", "good", "transferrable", "curate", "detail", "mandate", "notify", "critical", and "methodological" in the first quartile of the closeness centrality measure. There are no elements in the second quartile. Community one shares the ability "use", and the object "data" is its bridge with other communities. The nodes connected with the most relevant nodes within the network represent object-type elements "data" and "skills"; scope-type element "appropriate"; and ability-type elements "implement" and "develop".

3. Conclusions

This report shows that there is no set of common ethical skills being taught across universities to future workers engaging with different facets of AI systems. The detailed analysis of the course descriptions and goals of 503 courses in 66 universities shows that training on ethical and societal issues around ADMS is performed in multiple different ways, with little apparent overlap in terms of content, explicit goals, and skills acquired by students.

Most of the courses seem to focus on either the objects of study as they relate to existing university disciplines, or the fact that students will learn practical knowledge about certain topics. These courses tend not to clearly express the details (or even name) those skills and, except for a small number of detected topics, they do not point at the ethical skills that future workers will acquire in the courses.

While it is true that the philosophical and ethical debates around AI contain an inherent level of difficulty and complexity, the lack of precision (in the course descriptions and learning outcomes), along with gaps across courses (detected by low semantic similarity), and the use of objects of study and disciplines (as the pragmatic indexes of the language used in the courses' descriptions) signal a worrisome trend towards a poorly defined domain. This might be caused by the emerging nature of this new discipline(s), but that should not distract us from the urgent need to establish a regulation of training programs for future workers to be engaging with AI that truly train those future works in a minimum set of common ethical skills.

Although, the focus of the discussions in most of the evaluated courses center around the issue of responsibility in the use and development of AI systems which include ethical and social elements, the burden of the ethical quality of those systems cannot rest on the shoulders of the future workers of the sector only, unless these future workers are properly trained in a specific set of ethical skills that are common across the domain.

We propose that these common set of skills are designed through the notion of "Trustworthy ADMS" given the consensus that is being built around it and the methodological fact that to establish trustworthiness, a strong relationship must be established between the designers, the users, and the ADMS itself. "Trustworthy ADMS" are those that seek to foster the trust of AI users towards both products and development methods. With respect to the product, an ADM is trustworthy when trustworthiness is made a quality feature of the system through the integration of ethical elements such as privacy protection, robustness, or security. With respect to the software development method, the ADMS trustworthiness of a method results from the insertion and evaluation of ethical dispositions as part of the activities of quality assurance in the project's life cycle.

4. Areas for Further Research

To provide a more complete analysis of the current global context on the necessary skills to be taught in the field of ethical automatic decision-making, the analysis will need to expand to other languages than English, so other views can be incorporated to this study.

The current study is limited to the description and goals and outcomes declaration of the courses due to some access restrictions. Further research is needed where researchers can make use of the entire course syllabus in the construction of a skills map regarding ethical automatic decision-making around the globe.

Industry sponsored customized training and certification programs are increasingly more common, and these programs need to be considered so further analyses can be conducted while contrasting the academia's perspective and the industry needs regarding the issue of ethical automatic decision-making.

5. Knowledge Mobilization Plan

This project intends to deploy an integrated Knowledge Exchange plan (KEx) designed to produce outputs which allow various stakeholders to better understand the importance of integrating ethical skills for AI across all levels of instruction on programs devoted to machine learning, and for employers to consider these skills when describing requirements for job positions connected to the use of AI systems across organizations. We will follow a two-stage approach to knowledge mobilization. The first stage will focus on communicating and disseminating the present report, and the second stage will focus on using the report to develop a new research agenda focused on Ethical and Trustworthy ADMS.

With respect to the first stage, the report will be communicated and disseminated according to the SSHRC-Knowledge Synthesis guidelines and regulations. First, we will participate in an in-person or virtual knowledge mobilization forum six months after the grant has been awarded, to promote research findings with cross-sectoral stakeholders and knowledge users. We will share the synthesis report with all Canada's college and university level programs teaching AI and ethics. Second, we will engage with the Future Skills Centre to prepare and deliver two workshops with the Centre's partners during the six months following the finalization of the synthesis report. We expect these two workshops to be delivered online. Third, we will publish a minimum of two peer-reviewed articles on: 1) ethical skills for trustworthy artificial intelligence; 2) ethical and design considerations when engaging with ML at work. Fourth, we will also engage with different stakeholders through the publications of one volume of the CulturePlex Lab's Data Points devoted to the ethical skills and AI. The Data Points are data-based newsletters published throughout the year by CulturePlex lab and distributed among thought leaders in digital innovation in Canada, Latin America, and Europe. Lastly, in addition to existing courses in several departments, there are several data and AI initiatives at Western University. We will be collaborating with the Department of Research, Assessment, and Planning at Western's Student Experience to develop a unified set of skills and learning outcomes to be recommended for inclusion in all AI courses and programs at Western University.

The second stage of the Kex plan involves using the findings described in the report to develop new research projects, based on the opportunities and gaps identified. CulturePlex Research Associates and Collaborators will use this report as a discussion document for engagement with interested researchers and potential partners. The CulturePlex Director and the Lab's Research Network will join efforts to develop specific research proposals based on the report's results.

6. References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS '16) 308-318.

Abhishek, K., Tristan, B., Sasu, T., & Pan, S. (2020). Trustworthy AI in the age of pervasive computing and big data. Proceedings of IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops. March 23-27.

Abolfazlian, K. (2020). Trustworthy AI needs unbiased dictators!. Proceedings of Artificial Intelligence Applications and Innovations (AIAI 2020), May 2020. IFIP Advances in Information and Communication Technology, vol 584, 15-23. https://doi.org/10.1007/978-3-030-49186-4_2 2020.

Abreiu, R., Urvashi, A., Kate, C., Martin, R., & Anna, S.U. (2018) The future of work and education for the digital age: technological innovation and the future of work: A view from the South. Argentina: G20.

Access Now Organization. (2018). Human rights in the age of AI. AccessNowOrg.

Adès, J., Diaz, H., & Russell, N. (2019) Artificial intelligence and the global trade environment: Strategic foresight. The Conference Board of Canada.

Albarghouthi, A. & Hsu, J. (2018). Synthesizing coupling proofs of differential privacy. Proceedings of the ACM on Programming Languages, Vol. 2, No. POPL, Article 58.

Antal, K. & others. (2018). The next generation of emerging global challenges. Policy Horizons Canada/ Horizons de politiques Canada.

Ayat, S., Hojjat, F.A., Mehdi, A., Mahmood, A., Somayeh, A., & Zeynab, K. (2013). A comparison of artificial neural networks learning algorithms in predicting tendency for suicide. Neural Computing and Applications, 23, nº 5, 1381– 1386. 2013.

Billard, D. & Baptiste, B. (2019). Digital forensics and privacy-by-design: Example in a blockchain-based dynamic navigation system. In proceedings of Annual Privacy Forum APF '19. Privacy Technologies and Policy, 151-160. https://doi.org/10.1007/978-3-030-21752-5_10.

Bisong, E. (2019). Matplotlib and Seaborn. In Building Machine Learning and Deep Learning Models on Google Cloud Platform. doi:https://doi.org/10.1007/978-1-4842-4470-8_12.

Blei, D.M., Ng, A.Y. & Jordan, M.I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3(4-5), 993–1022.

Brynjolfsson, E. & McAfee, A. (2014). The second machine age: Work, progress, and prosperity in a time of brilliant technologies. Norton & Company.

Bucher, T. (2018). If...then: Algorithmic power and politics. Oxford University Press.

Cavoukian, A. (2011). The 7 foundational principles. Information and Privacy Commissioner of Ontario.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2): 153-163.

Cséfalvay, Z. (2019). What are the policy options? A systematic review of policy responses to the impacts of robotization and automation on the labour market. Technical Report of the JRC Working Papers on Corporate R&D and Innovation, No 02/2019.

Dwork, C. & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science, 9(3-4): 211-407.

Fast, N.J. & Jago, A.S. (2020). Privacy matters… or does It? Algorithms, rationalization, and the erosion of concern for privacy. Current opinion in psychology, 31, 44-48.

Feldman, M., Friedler, S.A., Moelle, J., Scheidegger, C. & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 259-268.

Fish, B., Kun, J. & Lelkes, Á.D. (2016). A confidence-based approach for balancing fairness and accuracy. Proceedings of the 2016 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 144-152.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center for Internet & Society.

Government of Canada´s Future Skills Program. (2020). Future Skills Centre / Centre des Competences futures, 2020. Retrieved July, 2021 from https://fsc-ccf.ca/.

Govia, L. (2020). Coproduction, ethics and artificial intelligence: A perspective from cultural anthropology. Journal of Digital Social Research, 2(3), 42–64. https://doi.org/10.33621/jdsr.v2i3.53.

Grasso, I., Russell, D., Matthews, A., Matthews, J. & Record, N.R. (2020). Applying algorithmic accountability frameworks with domain-specific codes of ethics: A case study in ecosystem forecasting for shellfish toxicity in the Gulf of Maine. In Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference FODS '20, 83–91 https://doi.org/10.1145/3412815.3416897.

Grundke, R., Marcolin, L., Nguyen, T.L.B. & Squicciarini, M. (2018). Which skills for the digital era? Returns to skills analysis. OECD Science, Technology and Industry Working Papers, 37.

Guide2Research. (2021). Our methodology. Retrieved May 22nd, 2021 from https://www.guide2research.com/our-methodology.

Guide2Research. (2021). World ranking of top computer scientists in 2021 (7th Edition). Retrieved May 21st, 2021 from https://www.guide2research.com/news-events/world-ranking-of-top-computer-scientists-2021.

Gurses, S., Troncoso, C. & Diaz, C. (2011). Engineering privacy by design. Computers, Privacy & Data Protection (CPDP'11), Jan 29, 14(3) 25.

Hardt, M., Price, E. & Srebro, N. (2016). Equality of opportunity in supervised learning. https://arxiv.org/abs/1610.02413.

High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI. European Commission.

Holstein, K., Wortman-Vaughan, J., Daumé, H., Dudik, M. & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need?. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.

House of Lords Select Committee on Artificial Intelligence. (2018). AI in the UK: ready, willing and able?. Authority of the House of Lords.

Hunter, J.D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9, 90-95. doi:DOI:10.1109/MCSE.2007.55.

IBM. (2019). IBM Everyday Ethics for AI. IBM.

Information and Privacy Commissioner/Ontario Canada & Registratiekamer The Netherlands. (1995). Privacy-enhancing technologies: The path to anonymity Volume 1. Information and Privacy Commissioner/ Ontario Canada & Registratiekamer.

Jens-Erik, M. (2019). Situating personal information: Privacy in the algorithmic age. In Human Rights in the Age of Platforms, 95-116.

Jiahao, C., Kallus, N., Mao, X., Svacha, G. & Udell, M. (2019). Fairness under unawareness: Assessing disparity when protected class is unobserved. In Proceedings of the Conference on Fairness, Accountability and Transparency ACM'19, 339– 348.

Johnson, N., Near, J.P. & Song, D. (2018). Towards practical differential privacy for SQL queries. In Proceedings of VLDB Endowment, 11(5) 526–539.

Kasirzadeh, A. & Smart, A. (2021). The use and misuse of counterfactuals in ethical machine learning. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency FACCT '21, 228–236. https://doi.org/10.1145/3442188.3445886.

Kelly, K. (2016). The inevitable: Understanding the 12 technological forces that will shape our future. Viking Press.

Loper, S.B. & Edward, E.K. (2009). Natural language processing with python. O'Reilly Media.

Martinho-Truswell, E., Miller, H., Asare, I.N., Petheram, A., Stirling, R., Gomez, C. & Martinez, C. (2018). Towards an AI strategy in mexico: Harnessing the AI revolution. The British Embassy in Mexico.

Mazloom, S. & Gordon, D.S. (2018). Secure computation with differentially private access patterns. In Proceedings of the ACM Conference on Computer and Communications Security (CCS´2018).

McCradden, M.D., Joshi, S., Anderson, J.A., Mazwi, M., Goldenberg, A. & Zlotnik Shaul, R. (2020). Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. Journal of the American Medical Informatics Association, 27(12) 2024–2027, https://doi.org/10.1093/jamia/ocaa085.

McKinney, W. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference (SPICY'10), 56-61.

Mehrabi, N., Morstatter, F., Nripsuta, S., Lerman, K. & Galstyan, A. (2019). A survey on bias and fairness in machine learning. Machine Learning.

Nedelkoska, L. & Glenda Q. (2018). Automation, skills use and training. OECD Social, Employment and Migration Working Papers, No. 202.

NITI Aayog. (2018). National strategy for artificial intelligence. India.

Oliphant, T.E. (2007). Python for scientific computing. Computing in Science & Engineering, 9(3), 10-20. 10.1109/MCSE.2007.58.

Olszewska, J.I. (2019). Designing transparent and autonomous intelligent vision systems. In Proceedings of the 11th International Conference on Agents and Artificial Intelligence (ICAART 2019), 850-856.

O'Neil, C. (2017). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books.

Pedregosa, F., Varoquaux, G., Gramfort, A., Vincent, M., Thirion, B., Grisel, O. & Duchesnay, E. (2011). Scikit-learn: Machine learning in {P}ython. Journal of Machine Learning Research, 12, 2825-2830.

Pedreshi, D., Ruggieri, S. & Tur, F. (2008). Discrimination-aware data mining. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM 560-568.

Pérez, F. & Granger, B.E. (2007). IPython: A system for interactive scientific computing. Computing in Science & Engineering, 9, 21-29. doi:DOI:10.1109/MCSE.2007.53.

Python Software Foundation. (2021). Python. (Python Org). Retrieved June 2021 from www.python.org.

Rehurek, R. & Sojka, P. (2010). Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.

Sahil, V. & Rubin, J. (2018). Fairness definitions explained. ACM/IEEE International Workshop on Software Fairness.

Sait Vural, M. & Gök, M. (2017). Criminal prediction using Naive Bayes theory. Neural Computing and Applications, 8(9) 2581-2592.

Scassa, T. (2021). Administrative law and the governance of automated decision-making: A critical look at canada's directive on automated decision-making. Forthcoming, University of British Columbia Law Review, 54:1 http://dx.doi.org/10.2139/ssrn.3722192.

Schmidt, N., Siskin, B. & Mansur, S. (2019). How data scientists help regulators and banks ensure fairness when implementing machine learning and artificial intelligence models. Conference on Fairness, Accountability, and Transparency. 2019.

Solon, B. & Selbst, A.D. (2016). Big data's disparate impact. CALIF. L. REV.104, 671-732.

Suárez, J.L. (2018). La nacionalización de la estrategia en torno a la inteligencia artificial Estado, política y futuro. Revista de occidente, no 446-447,5-18.

Suárez, J.L. (2020). The digital condition. Forthcoming.

T20. (2019). Task Force 7: The future of work and education for the digital age. Japan.

The Public Voice Coalition. (2018). Universal guidelines for AI. Belgium.

UNI Global Union and the future world of work. (2017). Top 10 principles for ethical artificial intelligence. UNI Global Union.

Université de Montréal. (2018). Montréal declaration for a responsible development of artificial intelligence. Université de Montréal.

University of Toronto. (2020). Canada's AI ecosystem government investment. University of Toronto.

Varona, D. (2020). AI systems are not racists just because. T-13 hours: Building Community Online in CSDH/SCHN2020.

Varona, D. (2020). Artificial intelligence design guiding principles: Review of "European ethical charter on the use of AI in judicial systems and their environment". Retrieved June 2021 from http://www.danielvarona.ca/2020/06/17/artificial-intelligence-design-guiding-principles-review-of-european-ethical-charter-on-the-use-of-ai-in-judicial-systems-and-their-environment/.

Varona, D. (2020). Artificial Intelligence design guiding principles: Review of "Recommendation of the council on Artificial Intelligence. Retrieved June 2021 from http://www.danielvarona.ca/2020/06/28/artificial-intelligence-design-guiding-principles-review-of-recommendation-of-the-council-on-artificial-intelligence/.

Varona, D. (2018). La responsabilidad ética del diseñador de sistemas en inteligencia artificial. Revista de occidente, nº 446-447, 104-114.

Varona, D. & Suárez J.L. (2021). Analysis of the principled AI framework's constraints in becoming a methodological reference for trustworthy AI design. Volume 2: Data Science, Statistical Modelling, and Machine Learning Methods, de HANDBOOK OF COMPUTATIONAL SOCIAL Vol II.

Varona, D., Lizama-Mue, Y. & Suárez, J.L. (2020). Machine learning's limitations in avoiding automation of bias. AI & SOCIETY, 36, 197-203. https://doi.org/10.1007/s00146-020-00996-y.

Vatican. (2020). Rome call for AI ethics. Italy.

Veale, M., Van Kleek, M. & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.

Verma, S. & Rubin, J. (2018). Fairness definitions explained. ACM/IEEE Proceedings of the International Workshop on Software Fairness (FairWare 2018).

Virtanen, P., Gommers, R., Oliphan, T.E., Haberland, M., Reddy, T., Curnapeau, D. & Van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods, 17, 261-272. doi:https://doi.org/10.1038/s41592-019-0686-2.

Walker, T. (2017). How much …? The rise of dynamic and personalized pricing. The Guardian.

Zafar, M.B., Valera, I., Gomez Rodriguez, M. & Gummadi, K.P. (2015). Fairness constraints: Mechanisms for fair classification. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS' 17). Volume 54.

Zemel, R., Wu, Y.L., Swersky, K., Pitassi, T. & Dwork, C. (2013). Learning fair representations. In Proceedings of the 30th International Conference on Machine Learning, Volume 28, 325-333.

Zhang, L., Wu, Y. & Wu, X. (2017). A causal framework for discovering and removing direct and indirect discrimination. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 3929-3935.

7. Appendix

7.1. Appendix I: List of Tables

7.2. Appendix II: List of Figures