ATHABASCA UNIVERSITY

OPERATIONALIZING A RESPONSIBLE AI SYSTEM IN THE PUBLIC SECTOR

BY

BRAD PAYNE

An essay submitted in partial fulfillment

Of the requirements for the degree of

MASTER OF SCIENCE in INFORMATION SYSTEMS

Victoria, British Columbia

January 9, 2024

@copyright Brad Payne, 2023

DEDICATION

Dedicated to those concerned and fascinated by the tension between harms and benefits, as well as to those who find solace in technical literacy while keeping humans at the forefront.

ABSTRACT

Responsible AI (RAI) describes a set of practices, processes and technologies that work towards improving trust when developing and deploying AI systems. Numerous studies examine ethical principles that underpin RAI systems, however relatively few case studies explore how practical it is to implement an RAI system. The first goal of this study was to evaluate the feasibility of integrating an RAI system with an existing recruitment system for digital talent roles in the public sector. The research determines a level of organizational readiness in the context of employing RAI tools, methods and processes that translate ethical principles into system requirements and design. The second goal was to evaluate barriers to adoption of RAI. A systematic review of the extant literature was conducted identifying 25 articles that were then synthesized by mapping the articles to the research questions. For the first goal, the case study relied on a participatory action research methodology and took an abductive approach to assess the feasibility of operationalizing RAI in the public sector. A responsible AI maturity model was used to compare a group's experiences with base practices and then determine levels of proficiency. For the second goal, a combination of theoretical frameworks including the Technology-Organization-Environment (TOE), Technology Acceptance Model (TAM) and Diffusion of Innovation (DOI) were applied to generate survey questions and evaluate barriers to RAI adoption. The findings support that RAI is not feasible to the extent that it is not currently being prioritized. Organizational readiness is low, with a government team demonstrating low proficiency in the necessary capabilities outlined by responsible AI maturity models. Barriers to adoption include factors in the 'Leadership Characteristics' and 'IT Expertise' categories. Recommendations are for the public sector

to consider prioritizing RAI, make shifts in organizational structure, incorporate staff training for AI, create a risk-based AI policy and adopt a modified team composition that adds RAI specific roles to traditional, cross-functional software development teams.

ACKNOWLEDGEMENTS

I would like to acknowledge Stephanie for her remarkable support and offer my gratitude to Koan and Lily who have had a dad who is “always” in front of a computer. I would also like to thank Steve C for advocating for my educational leave to work on this project, Jackie K for proofreading, Clint L and Kathreen R for writing program admission support letters many years ago and Dr. Stella George for her supervision and support.

Table of Contents

OPERATIONALIZING A RESPONSIBLE AI SYSTEM IN THE PUBLIC SECTOR.. i

DEDICATION ii

ABSTRACT i

ACKNOWLEDGEMENTS iii

LIST OF TABLES vi

LIST OF FIGURES vii

CHAPTER 1 1

INTRODUCTION 1

Regulatory Environment 3
Research Problem 7
Maturity Models 8
Meritocratic Hiring 9
Fairness Outcomes 11
Disclosure Risk 12
Research Questions 12
Evaluating Barriers to Adoption 13
Organization of the remaining chapters 14

CHAPTER II 16

REVIEW OF RELATED LITERATURE 16

General introduction — Operationalizing RAI 16
Literature Review 17
Literature related to RQ1: To what extent is it feasible to implement a responsible AI recruitment system in the public sector? 20
Literature related to RQ2: How can the public sector operationalize ethical principles when developing an AI recruitment system? 24
Statement of the Purpose 31

CHAPTER III 32

METHODOLOGY 32

Overview 32
Research Design 32
Strengths and Limitations 35

CHAPTER IV 36

ACTION RESEARCH PROCESS AND ANALYSIS 36

Cycle 1 — Problem Exploration: Requirements Elicitation 36
Cycle 1 — Actions 37
Cycle 1 — Analysis 41
Cycle 2 — Problem Exploration: Design 44
Cycle 2 — Actions 46
Cycle 2 — Analysis 48

CHAPTER V 50

RESULTS 50

Reflection: Cycle 1 — Questionnaire 50
Maturity Model Evaluation: Cycle 1 51

Reflection: Cycle 2 — Questionnaire 51
Maturity Model Evaluation: Cycle 2 55
Survey Results and Analysis — Adoption Barriers 56

CHAPTER VI 64

DISCUSSION 64

CHAPTER VII 70

CONCLUSION 70
Summary 70
Limitations 72
Conclusion 73

REFERENCES 74

APPENDICES 79

APPENDIX A 79
Questionnaire used for eliciting feedback on RAI tools and processes introduced 79

APPENDIX B 80
AI-MM for Trustworthy AI Software Development by Cho et al 80

APPENDIX C 80
Survey to examine adoption barriers for Responsible AI 80

APPENDIX D 84
User Flow Diagram comparison 84

APPENDIX E 85
Action Research Workshop 85

APPENDIX F 86
Taxonomy of currently existing forms of bias (Richardson and Gilbert) 86

APPENDIX G 87
System Architecture Design for Automated Redaction 87

LIST OF TABLES

		Page
1.	PICOC elements used in the search	18
2.	Search strings used for literature review	18
3.	Inclusion and Exclusion Criteria	19
4.	Results for Screening Papers	19
5.	Mapping Named Entities to Potential Disclosure Harms	38
6.	Fairness Software Toolkits	39
7.	Software requirements evaluation against an AI maturity model	50
8.	Software architecture evaluation against an AI maturity model	54

LIST OF FIGURES

Page

1. Architectural patterns for RAI by Lu et al. 45
2. Survey responses: Organizational role 56
3. Survey responses: Ethical AI adoption level 56
4. Survey responses: Usefulness of AI 58
5. Survey responses: Job ease 58
6. Survey responses: Leadership attitude 59
7. Survey responses: Leadership adoption 59
8. Survey responses: Government support 60
9. Survey responses: External pressure 60
10. Survey responses: IT expertise 61
11. Survey responses: IT size 61
12. Survey responses: Advantages of AI 62
13. Survey responses: AI compatibility 62
14. AI-MM for Trustworthy AI Software Development by Cho et al. 80
15. User Flow Diagram comparison 84
16. Taxonomy of currently existing forms of bias (Richardson and Gilbert) 86
17. System Architecture for integrating AI and Non-AI components in RedHat OpenShift 87

CHAPTER 1

INTRODUCTION

"Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals".[1] Responsible Artificial Intelligence (RAI) describes a set of practices, processes and technologies that work towards improving trust and accountability when developing and deploying AI systems [1]. The term RAI is widely conflated with 'Ethical AI' and 'Trustworthy AI' since they all emphasize ethical use of AI, though RAI also ensures auditability, accountability and liability [2]. The ubiquity and pervasiveness of AI technologies means its impact on our lives can generate increasingly significant concerns. The height of these concerns may be reflected in an open letter signed by prominent researchers who called for a moratorium on AI lab's experimentation[2]. The signatories call for an approach to AI research that is focused on making systems that are "…more accurate, safe, interpretable, transparent, robust, aligned, trustworthy and loyal." Asking ethical questions of the AI systems we use is imperative so long as there is a desire to benefit from these technologies, deploy them safely and have them align with human interests. Public sector organizations are not immune from the challenges, obligations and concerns around AI implementation.

[1] European Commission's definition of AI, https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines, accessed September 25, 2023
[2] Open letter calling for moratorium https://futureoflife.org/open-letter/pause-giant-ai-experiments/, accessed Sep 15, 2023

Currently, a set of digital principles guides AI work in the British Columbia (BC) public sector towards an ethical approach that includes oversight, documentation and consideration of applicable laws[3]. However, an 'ethical approach' is neither specifically defined nor contextualized with implementation details which reflects a broadly defined guideline with voluntary commitment that risks being largely ineffectual. Principle-based approaches have attracted criticism with some research concluding that principle-based AI ethics are both 'toothless' and reflect a dangerous misallocation of resources [3]. In response to these criticisms, the perceived utility of these principles may be improved by bridging the gap between theory and practice. Arguments have been made for employing standardized and proven engineering methodologies as well as leveraging certification and management strategies over principle-based approaches or principle-oriented checklists [1], [4], [5]. Further arguments propose that the key to operationalizing AI principles is understanding the distinction between intrinsic values and instrumental principles [6]. For example, autonomy is an intrinsic value because it is valuable in and of itself, whereas transparency or consent are instrumental principles because they are only valuable to the extent in which they protect or enhance autonomy.

This research is grounded in the Canadian public domain; knowledge of AI and designing ethical services are part of the current federal digital strategy[4]. A BC provincial public sector group had been tasked with collecting and manually redacting personally identifiable information (PII) in résumés prior to passing them onto a hiring manager. Manual redaction of hundreds of résumés was flagged as a significant, time consuming

[3] Draft principles for AI in BC public service, https://digital.gov.bc.ca/policies-standards/ai/, accessed Oct 17, 2023

[4] Canadian government's Digital Ambition 2022, https://www.canada.ca/en/government/system/digital-government/government-canada-digital-operations-strategic-plans/canada-digital-ambition.html, accessed Sep 25, 2023

task which relies on one or more individual's subjective assessment of redacted entities with no guarantee of consistency or accuracy. In response, the group expressed a desire to build a custom, online recruitment system and committed to integrating that system with an automated redaction component.

Public trust and acceptance of government services which employ RAI is not guaranteed and contingent on the ability to demonstrate sufficient competence, clarity of purpose and alignment with societal values. Determining a level of preparedness through an assessment of the public sector's ability to adopt and implement RAI presents a more fulsome picture of the missing capabilities necessary to effectively operationalize RAI. Discovering gaps in the public sector's ability to implement RAI speaks to strengths and weaknesses in areas such as human resources, team composition, organizational culture, technology, data, process and policy. Awareness of the evolving regulatory landscape around RAI positions the public sector in a state of readiness for forthcoming implications of the federal government's Artificial Intelligence and Data Act[5] (AIDA).

Regulatory Environment

At the time of writing the regulatory landscape for AI is emerging and remains fragmented with the European Union (EU) closest to being the first jurisdiction to implement laws that regulates AI[6]. Canadian laws are likely to be enacted soon after and will have implications for how the private sector implements AI including potential

[5] Current status of Canadian Bill C-27, https://www.parl.ca/LegisInfo/en/bill/44-1/c-27, accessed Oct 18, 2023

[6] Press release on progress of EU AI Act, https://www.europarl.europa.eu/news/en/press-room/20230609IPR96212/meps-ready-to-negotiate-first-ever-rules-for-safe-and-transparent-ai, accessed Oct 18, 2023

requirements for auditing in a regulatory context[7]. The Canadian government directive for Automated Decision Making[8], though limited to federal government implementations, is a risk-based approach that aims to identify "high-impact" AI applications. For instance, under this directive the use of AI in the hiring process is likely to be categorized as a "high-impact" activity due to the potential for recruitment decisions having significant consequences for the livelihoods and future careers of individuals. As a point of alignment, the EU AI Act (section 36)[9] identifies systems intended to be used for recruitment as 'high-risk' for which legal obligations and compliance are proportional to the risk. The Canadian directive also promotes ethical principles such as transparency, explainability, fairness, accountability, privacy and security and human-control which are common themes running through a variety of bills, guidelines and codes of practice.

International

In May of 2023, the intergovernmental political forum, the Group of Seven (G7) which includes Canada, France, the United Kingdom, Italy, Japan, Germany and the United States published the Hiroshima Communique[10]. The publication reflects work related to incubating AI policy and developing governance around the digital economy that is in line with democratic values including *"...*fairness, accountability, transparency, safety, protection from online harassment, hate and abuse and respect for privacy and human rights, fundamental freedoms and the protection of personal data." In January of 2023, the United States' National Institute of Standards and Technology (NIST) published an AI

[7] Audit language for AI in Bill C-27, https://www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading, accessed Oct 18, 2023
[8] Canadian Directive for Automated Decision Making, https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592, accessed Oct 18, 2023
[9] EU AI Act, https://artificialintelligenceact.eu/the-act/, accessed on September 22, 2023
[10] Hiroshima Communique, https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/20/g7-hiroshima-leaders-communique/, accessed Oct 18, 2023

Risk Management Framework [7] which is a voluntary framework that "…help[s] foster the responsible design, development, deployment and use of AI systems over time." The U.S. government introduced a bill, the "*Algorithmic Accountability Act of 2022*", that would direct the Federal Trade Commission (FTC) "…to require impact assessments of automated decision making systems and augmented critical decision processes…"[11]. At the time of writing the European Union (EU) AI Act[12] is a legislative proposal subject to a vote before potentially becoming the world's inaugural AI law. It includes explicit bans on certain technologies, for instance biometrics, and in some cases requires organizations to identify when they offer a 'high risk' AI system to which stricter obligations will apply in risk management, technical documentation and data governance[13]. In France, the Commission Nationale de l'Informatique et des Libertés (CNIL)[14] represents a policy initiative currently offering resources to guide organizations employing AI towards compliance with the General Data Protection Regulation (GDPR). In 2020, Spain created an AI governmental body, the Secretary of State for Digitalisation and AI with a strategic focus that includes "…establish[ing] an ethical and regulatory framework that guarantees the protection of individual and collective rights, with social welfare and sustainability."[15] In contrast, the United Kingdom (UK) decided against empowering a new, single regulator and offers a white paper that outlines five principles that existing regulators

[11] US Algorithmic Accountability Act, https://www.congress.gov/bill/117th-congress/house-bill/6580/text, accessed Oct 18, 2023
[12] EU AI Act proposal for regulation, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206, accessed Oct 18, 2023
[13] EU AI Act news, https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence, accessed Oct 18, 2023
[14] AI resources in France, https://www.cnil.fr/en/artificial-intelligence-ai, accessed Oct 18, 2023
[15] National AI Strategy in Spain, https://oecd.ai/en/dashboards/policy-initiatives/http:%2F%2Faipo.oecd.org%2F2021-data-policyInitiatives-27013, accessed Oct 18, 2023

should consider when regulating AI in the industries they monitor[16]. In 2023, China introduced regulations, The Deep Synthesis Provisions[17] which acknowledges the potential harms that can be caused by deepfakes.

National (Canada)

Bill C-27 is at its second reading in the Canadian Parliament where the Artificial Intelligence and Data Act (AIDA) and governance of AI in the private sector is on a pathway to becoming law[18]. The current Canadian government directive for Automated Decision Making[19] defines impact assessment levels and requires mitigation of potential harm resulting from high impact activities though is limited in scope to federal government initiatives. Ambiguity around the definitions of "artificial intelligence" and "high impact" leaves questions open as to how the law applies and to what it applies to. The Standards Council of Canada has published the "*Ethical design and use of automated decision systems*" [20] which provides a process and framework that helps organizations implement ethical principles. According to Tidjon and Khomh's study, Canada ranks 7th out of 20 countries for an AI readiness score which considers governance, infrastructure and technology [8]. The authors state that some support for AI ethics can be found in the Personal Information Protection and Electronic Documents Act (PIPEDA), Financial Consumer Agency of Canada Act and the Privacy Act. These laws were created as early

[16] UK government policy paper, https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper, accessed Oct 18, 2023
[17] Cyberspace Administration of China, http://www.cac.gov.cn/2022-12/11/c_1672221949318230.htm, accessed Oct 18, 2023
[18] Canadian Bill C-27, https://www.parl.ca/LegisInfo/en/bill/44-1/c-27, accessed Oct 18, 2023
[19] Canadian Directive on Automated Decision Making, https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592, accessed Oct 18, 2023
[20] Ethical Design and Use of Automated Decisions Systems, https://dgc-cgn.org/standards/find-a-standard/standards-in-automated-decision-systems-ai/cisoc101/, accessed Oct 18, 2023

as 1985 and the early 2000s and are likely inadequate in addressing ethical issues like autonomy, bias, justice and societal harms directly related to AI.

Provincial (British Columbia)

Development of AI policy is in progress at the time of writing, and is likely to include an alignment guide and take a principle-based approach [21]. Other provincial initiatives include a report from the Ombudsperson of BC and Office of the Information and Privacy Commissioner of BC that recommends a principle-based approach as well as practical considerations such as AI capacity building, public engagement, scrutiny and oversight.

Research Problem

Concerns about the trustworthiness of AI and the capability of governments to regulate the technology have not deterred AI applications from being deployed in the provincial public sector. A brief survey of publicly available information demonstrates the use of AI in procurement, development and research including applications tasked with sentiment analysis[22], chat bots[23], image recognition[24], wildfire detection and prediction[25], education analytics[26], AI based surgical systems, network planning and a variety of Natural Language Processing (NLP) applications. This study works with a team in the public sector who were tasked implementing a meritocratic hiring process by integrating an automated redaction system that targets PII in résumés. Ethical concepts most relevant to

[21] Draft principles for responsible use of AI, https://digital.gov.bc.ca/policies-standards/ai/, Oct 18, 2023
[22] Procurement opportunity for sentiment analysis, https://marketplace.digital.gov.bc.ca/opportunities/code-with-us/319eabb2-4df2-471a-958c-3475ad3c806e, accessed Oct 18, 2023
[23] Automated BC Covid Assistant, https://www2.gov.bc.ca/gov/content/covid-19/info/response, accessed Oct 18, 2023
[24] Automated License Plate Recognition in BC, https://www2.gov.bc.ca/gov/content/justice/criminal-justice/policing-in-bc/road-safety-auto-crime/auto-licence-place-recognition, accessed Oct 18, 2023
[25] Wildfire Predictive Services code repository, https://github.com/bcgov/wps-api, accessed Oct 18, 2023
[26] OCIO Thought Paper for AI, https://www2.gov.bc.ca/assets/gov/british-columbians-our-governments/services-policies-for-government/information-management-technology/information-security/thought_paper_-_artificial_intelligence.pdf, accessed Oct 18, 2023

a recruitment system are privacy, fairness and explainability [5]. Expectations that are unique to public sector organizations include adherence to social benefit, appropriate accommodation for whom their services are available, incentivization that is absent of a profit motive and alignment with applicable laws, policies and standards. The feasibility of implementing an RAI system can be influenced by a combination of these expectations and a number of other related factors including the degree to which organizations have the required foundational AI capabilities and capabilities specific to RAI [9]. From this perspective, assessing the level of organizational readiness can be accomplished by employing a responsible AI maturity model which defines a set of dimensions and relevant proficiency levels to measure against. For organizations whose level of readiness is determined to be low, it may be appropriate to include an investigation into technology adoption barriers which can reveal specific challenges and opportunities for growth. While recruitment systems typically apply AI in many areas in the hiring process [10] the scope of this research project looks at the feasibility of implementing an automated redaction system for PII in résumés. Named Entity Recognition (NER) is a sub-task of Natural Language Processing (NLP) and is employed for text anonymization [11], [12].

Maturity Models

Organizations can self-assess their current AI capability situation by employing a maturity model. RAI Maturity Models (RAIMM) typically consist of a representation of a maturity level and an assessment method [13]. Sadiq et al. identify critical dimensions for an organization to consider including "…Data, Analytics, Technology and Tools, Intelligent Automation, Governance, People and Organization". Cho et al. introduce a maturity model for trustworthy AI which includes ethical considerations such as fairness

and safety as well as an evaluation method to establish model validity [14]. Industry has also made contributions to the available list of AI maturity models including Microsoft[27], Gartner [28] and IBM[29].

Meritocratic Hiring

A meritocratic hiring process aims to reduce allocation harms, whereby opportunities are not allocated to otherwise qualified candidates based on conscious or unconscious bias that may arise from knowledge of personal attributes including but not limited to name, age, location, organization, nationality and gender. De-identification in this context aims to reduce the likelihood that bias of the evaluator can be associated with the candidate not receiving an interview. Automating the redaction of PII in résumés can efficiently produce a de-identified version so that screening decisions can be based on whether the candidate meets the criteria for the job as opposed to being based on any personal and protected characteristics.

Named Entity Recognition (NER) is a sub-task of NLP that is typically characterized as a sequence labelling problem [11]. Assessing whether Natural Language Processing (NLP) can have an impact on social justice can be determined by looking at ethical themes such as exclusion (demographic bias), overgeneralization (modelling side-effect), overexposure, underexposure and dual-use problems [15]. Language can be a proxy for human behaviour such as political dissent and can be used to predict human characteristics like belonging to a specific group. Recommendations for working with bias in NLP systems include establishing a connection between social hierarchies and

[27] Microsoft, Understanding AI Maturity, https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4DIvg, accessed Oct 18, 2023
[28] Gartner, CIO's Guide to AI, https://www.gartner.com/smarterwithgartner/the-cios-guide-to-artificial-intelligence, accessed Oct 18, 2023
[29] IBM, AI maturity framework, https://www.ibm.com/downloads/cas/OB8M18WR, accessed Oct 18, 2023

language and being explicit about why system behaviours are harmful, specific to whom and in what ways [16]. Engaging with members of the community to discover lived experiences can also aid in re-imagining power dynamics.

When employed towards a de-identification task, named entities such as Person (PERSON) and Location (LOC) reflect common direct identifier targets. De-identification, however, is different than anonymization [17]. De-identification may not go far enough to protect individuals' private information. Private, sensitive and confidential information which employers do not have a right to consider in employment decisions are readily available in résumés and can be used to infer information like lifestyle activities with considerable accuracy [18]. The risk of inferring confidential attributes such as belief, sexual orientation, gender identity, ethnic origin, health status or political opinions is possible by way of 'proxy' variables. Methods for de-identification in résumés need to go further than detecting direct identifiers and measuring model performance in order to achieve a reasonable privacy protection goal [19].

The GDPR defines anonymous information as "…information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable"[30]. This means that text anonymization can only be achieved if it includes indirect identifiers as well. Text anonymization in résumés can support the principle of meritocracy when direct identifiers, confidential attributes and quasi-identifiers can be masked while balancing data loss.

[30] Recital 26, EU GDPR, https://www.privacy-regulation.eu/en/recital-26-GDPR.htm, accessed Oct 18, 2023

Fairness Outcomes

In recruitment, forms of injustice are often expressed through categories such as gender, race, national origin or age [20]. Allocation harms can stem from knowing political affiliations, asserting some notion of 'culture fit' or inferring protected categories such as sexual orientation, religious beliefs or disability [18]. One approach to mitigate against discrimination in the hiring process is the concept of group unaware fairness[31]; what's fair is to disregard the protected attributes, such as what can be inferred from a name (i.e. gender and/or sex, marital status, national or ethnic origin) and choose candidates based purely on their merits. For example, should a protected category (e.g. male/female as per the Canadian Human Rights Commission[32]) be unavailable to a potential employer, then that employment decision can no longer be influenced by that category and will be less vulnerable to legal exposure. A redacted résumé aligns with group unaware fairness and reflects efforts to reduce bias while enhancing meritocracy which are recommendations found in research for building ethical AI in recruitment [21]. The task for text anonymization of résumés in this context needs to find all sources in the résumé text where protected attributes can be inferred and then apply appropriate entity labels. Should a text anonymization system fail to identify and mask a direct or indirect identifier in a résumé, harm can result from knowing or inferring protected or confidential attributes which may contribute to conscious or unconscious bias in the hiring decision.

[31] Google, Playing with AI Fairness, https://pair-code.github.io/what-if-tool/ai-fairness.html, accessed Oct 18, 2023

[32] Canadian Human Rights Commission, What is Discrimination?, https://www.chrc-ccdp.gc.ca/en/about-human-rights/what-discrimination, accessed Oct 18, 2023

Disclosure Risk

Masking common, named entity labels that are available in pre-trained NER models also risks falling short of meeting legal requirements for anonymization in the GDPR[33], CPPA[34], HIPA, PIPEDA[35], and PHIAPA[36]. For these reasons, expanding entity coverage beyond what pre-trained models offer to also include labels for quasi-identifiers and confidential information are necessary. A technical challenge associated with broader entity coverage is the task of fine-tuning a pre-trained NER model to identify more than the base entities while demonstrating sufficient performance [22]. Choosing an appropriate, pre-trained base model and an annotated dataset for training involves model architecture considerations, hardware availability (i.e. GPU) and an assessment of sufficient model performance, training times and model sizes.

Research Questions

RQ1: To what extent is it feasible to implement a responsible AI recruitment system in the public sector?

RQ2: How can the public sector operationalize ethical principles when developing an AI recruitment system?

This study was originally designed to determine the feasibility of operationalizing an RAI system in the public sector while reflecting on the effectiveness of tools, methods and processes that collectively inform 'best practices' around RAI systems. At several times during the research project, the public sector group pivoted away from AI work to focus on other priorities and was unable to operationalize an RAI system. In response, the

[33] GDPR, https://eur-lex.europa.eu/eli/reg/2016/679/oj, accessed Oct 18, 2023
[34] CCPA and CPRA Topic Page, https://iapp.org/resources/topics/ccpa-and-cpra/, accessed Oct 18, 2023
[35] PIPEDA, https://canlii.ca/t/7vwjaccessed Oct 18, 2023
[36] PHIPA, https://www.ontario.ca/laws/statute/04p03, accessed Oct 18, 2023

research focus also needed to pivot to evaluating adoption barriers of RAI in the public sector.

RQ3: What are the adoption barriers for RAI in the public sector?

Evaluating Barriers to Adoption

Evaluating RAI adoption barriers helps identify priority areas that could enable the positive aspects of AI technologies while mitigating harms. This information can be obtained from employing one of many frameworks that assess a hierarchy of relevant RAI adoption barriers [23]. Merhi's research extends the Technology-Organization-Environment (TOE) framework and identifies 11 barriers to RAI. The TOE framework is a theoretical model which describes factors that influence technology adoption [24]. Merhi determines that technical barriers are the most significant followed by organizational and legal (environmental) issues being the least prominent barrier, though the author's categorization of technical barriers includes ethical principles, namely justice, fairness, transparency, explainability, privacy and security viewed through a data quality lens. The Technology Acceptance Model (TAM) is an individual-focused theoretical model of adoption which aims to predict user acceptance by looking at the attributes 'perceived ease of use' and 'perceived usefulness' [25]. TAM was derived from a Theory of Reasoned Action (TRA) which undertakes explaining user acceptance behaviour [26]. The Theory of Planned Behaviour (TPB), also derived from TRA, can be used as a theoretical underpinning to predict individual behaviour [27]. One of the most widely used theoretical models in literature is the Diffusion of Innovation Theory (DOI) [28]. DOI is used to explain how an idea or product gains momentum over time and spreads through a specific population. DOI defines five adoption categories which include

innovators, early adopters, early majority, late majority, and laggards. DOI proposes five main factors that influence adoption and each of these factors are at play within each of the five adoption categories. These factors are relative advantage, compatibility, complexity, trialability and observability. Although DOI, TAM, TRA and TPB have been widely used, one of the limitations of these models is that they take an individualistic approach and do not consider organizational factors. When evaluating organizational barriers to adoption, differences between public and private sector should be taken into consideration since public sector organizations do not operate in a competitive space where profit-driven incentives for efficiency and cost reduction are more influential. These differences require asking slightly nuanced questions or interpreting the responses to those questions around relative organizational advantage with a public sector perspective.

Organization of the remaining chapters

The rest of the article is organized as follows. After this introductory section Chapter II provides a comprehensive, systematic review of related literature which maps articles to relevant research questions. Chapter II also describes the research design modifications that were necessary to adapt to the group's decision to prioritize different work. Chapter III describes the action research process which covers two cycles of identifying a problem, implementing a plan and analyzing the effectiveness of that plan. Chapter V includes the results from a questionnaires delivered at the end of both cycles and the results of the barriers to adoption survey. As well, proficiency levels of the group are compared against a responsible AI maturity model to highlight gaps in base practices and required competencies. Chapter VI includes a discussion of the implications of the results

and offers recommendations. Chapter VII summarizes the findings, presents a brief discussion on limitations and provides a conclusion.

CHAPTER II

REVIEW OF RELATED LITERATURE

General introduction — Operationalizing RAI

AI principles, frameworks, guidelines, laws, and applied ethics have yet to generate a comparable amount of evidence towards the practicality of implementation [29]. Translating high-level ethical principles into concrete guidance for system-level architecture design, development and deployment of responsible AI systems reflects an under-explored research area [30]. The gap between the 'what' of ethical principles and the 'how' of implementation can be reduced by creating a taxonomy of AI tools and methods which are mapped to ethical principles across the software development lifecycle [29]. Morley et al. research identify patterns that emerge from curating available tools highlighting an uneven distribution of tools related to the principle of explicability, a focus on individual harms over collective harms and an overall lack of usability. Morley et al. conclude that future research could be directed towards creating tools that fill in the gaps and evaluating existing tools by creating working examples.

Ayling and Chapman create a typology of AI ethics and tools for operationalizing responsible AI by way of a qualitative content analysis of relevant documents [31]. The structure of the typology draws inspiration from other domains with a much longer history of impact assessment and audit literature than AI. The authors conclude that all but one of the tools are designed for internal self-assessment during the development phase of AI. Of the 11 stakeholders identified, for whom the 39 identified tools and guidelines could be used by, developers are the predominant stakeholder group. A

software engineering roadmap which includes perspectives on tools, governance, processes and systems-level considerations for responsible-AI-by-design is a valuable component when organizations are preparing for RAI implementation [32].

Katell et al. use a participatory action research design to report on the creation of the Algorithmic Equity Toolkit (AET) in the context of the city of Seattle passing the Seattle Surveillance Ordinance which mandated the disclosure of surveillance technologies [33]. One of the goals of the research is to create a framework (AET) for community participation in algorithmic assessments which could be generalized to be applicable to other municipalities who wish to subject technologies to government oversight. The authors propose grounding algorithmic research in the context of specific situations to bring social context to an otherwise abstract analysis. Similarly, this case study grounds implementation of RAI in a specific time and place and generates recommendations that could be applicable to other public sector entities hoping to operationalize RAI.

Literature Review

Step 1 — Conducting the search

The search string was constructed according to recommendations discussed in the technical report by Kitchenham et al. [37] — PICOC is an acronym for Population, Intervention, Comparison, Outcomes and Context. For this study, the *Population* is the public sector and more specifically it is a provincial public sector in Canada. The main issue, or *Intervention* being explored is the implementation of RAI. Since this literature review is not comparing the implementation with any control treatment, *Comparison* is

[37]Guidelines for SLR, https://www.researchgate.net/publication/258968007_Kitchenham_B_Guidelines_for_performing_Systematic_Literature_Reviews_in_software_engineering_EBSE_Technical_Report_EBSE-2007-01, accessed Oct 18, 2023

left blank. The *Outcome* relates to a comprehensive understanding of the feasibility of operationalizing which requires bridging the gap between principle and practice. The *Context* is the hiring and recruitment processes that occurs in the workplace. The PICOC elements are presented in Table I.

Table I - PICOC elements used in the search

PICOC	Elements
Population	Provincial public sector in British Columbia
Intervention	Responsible AI
Comparison	N/A
Outcomes	Feasibility of operationalizing which requires bridging the gap between principle and practice.
Context	Hiring and recruitment process

Though slight modifications were necessary to adapt to varying syntactic constraints in each of the journal databases, the following search strings were used to obtain the results. Science Direct for instance, required fewer Boolean connectors and disallowed wildcards. When an option was made available to select subject areas, “Computer Science” and/or “Artificial Intelligence” further narrowed the search. The search strings used in this study are presented in Table II.

Table II - Search strings used for literature review

Database	Search String
IEEE	(‘public sector' OR 'public service’ OR 'government*’ OR ‘democra*’) AND (‘responsible AI' OR ‘ethic* AI' OR 'responsible NLP' OR ‘responsible NER’) AND (‘feasibil*’ OR 'implement*' OR ‘operational*')
ACM	(‘public sector' OR 'public service’ OR 'government*’ OR ‘democra*’) AND (‘responsible AI' OR ‘ethic* AI' OR 'responsible NLP' OR ‘responsible NER’) AND (‘feasibil*’ OR 'implement*' OR ‘operational*') AND ('recruit*' OR 'hiring' OR 'human resources' OR ‘HR’)
Science Direct	('public sector' OR ‘government’) AND (‘responsible AI’ OR ‘ethical AI’) AND ('feasibility' OR 'implementation') AND (‘recruitment’ OR 'hiring' OR ‘HRM’)

Step 2 — Study Selection and Data Sources

After the initial results were obtained and the inclusion and exclusion criteria (shown in Table III) were applied, it reduced the total number of studies significantly. In situations where applying the date range and peer review inclusion criteria to the search results still generated thousands of entries, the papers were sorted by relevance and filtered by reading the titles and abstracts of the top 40 results. After the titles of those secondary results were scanned for relevance, and if further clarification was needed, then the abstract was read as well.

Table III - Inclusion and exclusion criteria

Inclusion Criteria	Exclusion Criteria
Studies published after 2018	Studies that do not fulfill every inclusion criterion
Peer reviewed (conference, journal or workshop)	Full text access not available
The study answers one of the research questions	
Studies use the English language	

Step 3 — Analysis

Only after determining that both the abstract and title were relevant to the task at hand was the paper read in full to see if it answered at least one of the research questions. If it did not answer one of the research questions or was deemed to be of poor quality due to small sampling size, insufficient description of methodology or questionable validity then it was eliminated. The results at each step of the process is captured in Table IV.

Table IV - Results for screening papers

Database	Search Results	Inclusion/Exclusion	Quality Assessment	Final Primary Papers
IEEE	42	41	11	10
ACM	92,025	6,610	6	6
Science Direct	5,263	2,658	9	9

Step 4 — Overview of primary studies

Twenty five primary studies from three databases, or digital libraries form the basis of this systematic review. Ten were selected from IEEE, six from ACM and nine from Science Direct.

Literature related to RQ1: To what extent is it feasible to implement a responsible AI recruitment system in the public sector?

Assessing the feasibility in an organizational context can be facilitated by identifying enablers and barriers and determining which factors make implementation easy or difficult. Ravoka et at. identify organizational practices that enable the implementation of responsible AI and consider broad themes such as corporate responsibility, culture and organizational change [34]. Results of the authors' semi-structured interviews are organized by employing a consistent format to the responses from four key organizational questions. The structure of each of the responses maps out a route between existing work practices, emerging practices and aspirational work practices. The authors make note of a commonly low organizational state of readiness which means that individual effort is the primary engine behind advocating for organizational change and articulating a clear view of future state. In most cases, organizational resources, structures, roles and processes were not in place to support the implementation of RAI. Common themes from respondents highlight the importance of being able to veto an AI system, building channels of communication, identifying external and internal motivations for corporate change and acknowledging the difficulty of conceptualizing this type of organizational transition as linear movements. Sequencing organizational actions is challenging because of their interdependence and the fluid nature of organizational contexts. Their findings

highlight the precariousness of prevalent RAI practices which currently reflect a reliance on formidable individual effort instead of formalized and well-resourced organizational structures and processes.

Seidelin et al. conduct an audit of a public service algorithm that received legal challenges and media attention after a job seeking application was suspected of profiling Danish citizens which amounted to discrimination [35]. As an example of how governments can use algorithms to support decision-making which can then have significant negative impact on people's lives, the research speaks to a level of scrutiny that may be necessary to ensure that access to public services is free from biased outcomes. The authors note that even the software development process can be a source of divergence from the original intent to the actual implementation due to the "…rotating cast of characters.." who may work on it over time. The authors point to recent studies that stress the importance of questioning AI algorithms in the public sector to better understand their underlying assumptions. They frame their own auditing efforts as an extension of the principle of explainability. The authors describe using a cooperative audit which removes any adversarial notions around investigating the algorithmic causes of harm.

Gutierrez investigates motivations that could improve the effectiveness of implementing AI policies and influence behaviours of a target population towards a desired outcome [36]. In contrast to the time that it would take to manage AI technology with regulation, 'soft' law, which the author classifies as guidelines, standards, principles or internal policy can be a flexible and efficient alternative. However, without incentives or enforcement mechanisms, governing AI with 'soft' law will remain voluntary,

potentially vague and serve little purpose. Motivations for compliance that policymakers can consider include communicating and clarifying the enforcement of the impending regulation, appealing to an organization's self-interest and encouraging associations to provide a soft law alternative that can be uniformly implemented. The absence of both regulations and 'soft' law in the public sector presents challenges to the practicality of implementing an RAI system.

The research in [9] looks at developing a framework of required capabilities cultivated from a literature review of AI maturity models which can be used by organizations to determine their level of readiness for implementing responsible AI. The authors group these capabilities into two categories, foundational and responsible AI. According to the authors, foundational capabilities reflect necessary technical components which include data quality, data availability, financial, organizational strategies, innovation, change and human resources management, interdepartmental coordination and performance of AI models. RAI capabilities reflect social components and include employee ethical awareness, continuous impact analysis, security and privacy, fairness evaluation, understandability of AI models and accountability. RAI capabilities helps achieve humanistic objectives whereas foundational capabilities helps achieve instrumental objectives.

Gonzalez et al. research applicant reactions to the introduction of AI/ML-based recruiting processes [37]. The authors look at the differences in reactions to human-based, AI-ML based and augmented hiring practices. The results of their study determined that reactions were more favourable to recruitment systems with either a human decision maker or an augmented system which the authors define as a combination of both a

human decision maker and AI. Further analysis revealed a relationship between favourable reactions and a candidate's familiarity with AI which has implications for how and for whom organizations should incorporate AI into their recruitment system. Given that job applicants generally reacted unfavourably towards AI the researchers analyze various factors, such as perceptions of fairness, respect and control that could influence the reactions to AI in the recruitment process.

The extent to which citizens' trust in government and policy support impacts AI projects is investigated in [38]. The results of two online surveys provide evidence to support the notion that factors other than designing an actual trustworthy system are more crucial to the perceived trustworthiness of a system. How citizens respond to ethical AI information from government is the focus of the research. The authors' findings look to existing negative attitudes towards AI, privacy concerns and low trust in government as predictors for support of government AI projects. These attitudes have an impact on the effectiveness of communication including transparency documentation that describes what ethical AI measures have been taken.

Factors, affordances and constraints related to AI implementation in government are identified by the authors in [39] and evaluated using the TOE and Technology Affordances and Constraints Theory (TACT) frameworks. While these are common for evaluating standard technologies, Maragno et al. develop a conceptual framework by combining TOE and TACT that allows them to distinguish new affordances and challenges specific to AI. Technological barriers identified by the results of the research include difficulties in implementing explainable AI, availability of infrastructure, acquiring high quality interoperable data and difficulties collecting new data.

Organizational constraints include employee's mistrust of AI, lack of awareness and relevant human resources and a general reluctance to change. Environmental factors include citizen's mistrust of AI, administrative silos, cooperation with suppliers and the complexity and dynamism of the regulatory framework. The authors present three propositions for public sector organizations to consider for implementing AI systems while avoiding negative consequences for society: (1) Understanding AI technology's functionalities is a requirement for addressing explainability issues. Investments in data governance and infrastructure will afford data-driven activities. (2) Implementation requires an exchange between human and technological agents. While the creation of new jobs and augmentation of existing activities is likely, organizational mistrust born from a lack of awareness can hinder AI implementation. (3) A regulatory framework is a requirement that will play a role in the complex management of collaborations and engagement with multiple stakeholder groups.

Literature related to RQ2: How can the public sector operationalize ethical principles when developing an AI recruitment system?

Canca's research into operationalizing AI ethical principles presents the argument for distinguishing between core principles, which encapsulate intrinsic values, versus instrumental principles that are valuable only to the extent to which they protect or uphold intrinsic values [6]. The advantage of this taxonomy is that the myriad of AI principles can be grouped and categorized according to their function of upholding one of three intrinsic values, autonomy, justice and beneficence. One example is how transparency, an instrumental principle along with explainability, human control, privacy and consent are valuable in the context of how they can be used to uphold the intrinsic value of autonomy.

Making the distinction between instrumental and intrinsic values in the context of operationalization is important when confronted with making complex ethical decisions. For instance, when an application is transparent through its explanation of the reasoning behind the output of an AI decision and they are given a choice to 'opt out', a user can exercise their autonomy by choosing to use the application or not. The author notes that a list of instrumental principles is a good start for ethical AI, but it is neither sufficient nor a complete system for ethical analysis. Ideally, recognizing the difference between instrumental and intrinsic principles would help experts apply ethical theories based on an organization's inherent values and offer more comprehensive guidance for ethical analysis. Building on the idea that instrumental principles are interchangeable and dependant on the details of each project, one of the paper's contributions is a tool that helps developers determine relevant ethical concerns and satisfy core principles by implementing or interchanging different combinations of instrumental principles.

Saxena et al. provide public sector recommendations that afford insight into how to ethical principles can be operationalized when deploying an AI system [40]. The authors recognize unique challenges in the public sector such as inconsistently defined outcomes and poorly collected aggregate administrative data. They also identify opportunities and present design guidelines for algorithmic decision-making systems as well as detail implications for developing them that are specific to the public sector. Their framework of high-stakes algorithmic decision-making is employed to analyze four implementations of algorithms used in a Child-Welfare System. The framework applies a socio-technical perspective that serves as a theoretical lens that considers the interactions between three interdependencies: Human discretion, bureaucratic processes and algorithmic decision-

making. The authors' framework helps design new algorithms by assessing the strengths and weaknesses of existing algorithms employed for high-stakes decision-making in the public sector. Some of their design guidelines include supporting the practitioner's decision making process, applying oversight, consulting key stakeholders, making algorithmic outputs multidimensional, creating safeguards to check for gaming behaviours and avoiding mandated decision outcomes.

Four techniques are discussed in [41] for achieving the goals of what the authors refer to as Socially Responsible AI (SRA) algorithms. The techniques are adversarial machine learning, interpretability and explainability, causal learning and uncertainty quantification. Adversarial machine learning relates to improving the security posture and robustness of SRA. Human-level intelligence, which is an overarching goal of SRA, requires causal reasoning in order to achieve that goal and the inevitability of models to encounter new data samples that are different than those used in training creates uncertainty, which can be measured and communicated as a form of transparency. While the focus of the research is primarily on algorithms, it does so in the context of the relationship between AI systems and humans while speaking to implementation details that are relevant to bridging the gap between principles and practice. For example, causal learning methods such as counterfactual fairness and causal interpretability are presented as a means to address fairness related issues.

The research activities employed by Ananjeva et al. included semi-structured interviews, participant observations, working alongside the team and conducting workshops when interventions were deemed necessary to solve problems [42].

Tidjon and Khomh contextualize gaps in the practical guidance of ethical AI principles by comparing levels of readiness across 29 countries in 6 continents [8]. From the authors' analysis of the literature on practical guidance the following were found to be lacking: tools that cover all ethical AI principles, effective standards and step-by-step technical instructions, practical training courses on AI ethics, corporate governance of AI and coverage of ethics in artificial general intelligence. The authors furthered their analysis by presenting root causes for those gaps which include human bias, lack of metrics for AI principles, complexity of the interconnection of AI, poor diversity within the AI community, issues in how AI research is funded, and the time, effort and money that implementing AI requires when bringing it into production. By identifying gaps, root causes and considering solutions for principle-to-practice gap mitigation the researchers are able to make recommendations. The authors recommend diversity in the teams that design, develop and guide ethical implementations of AI as well as more metrics that can evaluate whether AI systems are ethically-aligned. Human-centered research, more public-private partnerships for AI ethics research and alignment of projects with EthicsOps rounds out all five recommendations.

Gutierrez describes the importance that motivation plays in successfully implementing ethical principles [36]. The author distinguishes between 'soft law' and 'hard law' incentives; repercussions being the most obvious motivation for complying with hard law. While soft law, which includes "…principles, guidelines, codes, certifications and standards..." is presented as a quick and responsive alternative to 'hard' regulations, the author points out that they are often voluntary and lack enforcement mechanisms. The author presents arguments that look at the incentives to have stakeholders abide by soft

laws which includes framing compliance around self-interest, or as a caution, warning or nudging to get stakeholders to act in ways that benefit public objectives and as a mechanism to communicate anticipating or pre-empting hard law.

Peters et al. employ the *Responsible Design Process* and the *Spheres of Technology Experience* which are two frameworks for translating ethical principles into actionable strategies in a case study that evaluates the outcomes of ethical analysis in the context of mental health technology [43]. The contribution of their research is a list of ten recommendations that address the difficulties and challenges of implementation some of which includes the following: technology supporting user autonomy, transparent use of data and secure storage, considering the impact of risks for all stakeholders and being well informed of accessibility guidelines. Though the case study is specific to mental health applications where biomedical ethics already plays a significant role in the development of technology, the authors note that AI brings with it a number of new ethical concerns such as transparency and human autonomy.

The authors in [44] introduce basic concepts and central themes found in AI ethics, including human well-being, safety, privacy, transparency, fairness and accountability. Kazim and Koshiyama outline how relevant approaches map to each of the themes. One example includes mapping both ‘human-in-the-loop’ and algorithmic impact assessments to the principle of accountability. Since the focus of the paper is to provide a discussion, the principles and relevant approaches are presented as high-level concepts in the sense that they refer to a general practice, like data minimization or process, such as ethical-by-design.

The authors in [2] provide a multidisciplinary analysis of trustworthy AI by answering the what, why and how questions for seven technical requirements. The practical applicability of RAI is contextualized by a multitude of considerations including necessary definitions, a philosophical approach, widely recognized AI principles, and a risk-based viewpoint to regulations. An overview of technical possibilities is presented as responses to the 'how can this requirement be met in practice' question. For instance, the technical possibilities for satisfying the accountability requirement include employing algorithms for causal inference and reasoning, redress mechanisms, accountability frameworks and using auditing procedures for validating models. For example, the authors include the Muir Trust Scale which measures predictability, reliability, competence and trust. The other requirements identified by the authors are human agency and oversight, technical robustness and safety, privacy and data governance, transparency, accountability, societal and environmental wellbeing and finally diversity, non-discrimination and fairness.

A link between proposed legislation and a requirement to assess the social, economic and ethical impacts for automated decision systems is explored in the research by Oduru et al. [45] The authors chose three U.S. bills and the EU AI Act to examine the new assessment requirements that would be imposed on developers should they be passed into law. Compliance work that includes more rigorous accountability documentation as well as practical methods for carrying out that work in support of public interest are identified as areas needing attention. The authors recognize that operationalizing these new requirements can have wide impacts to existing organizational processes, including

keeping an inventory of algorithms and machine learning models it is using while also being able to provide transparency around the data used and how decisions are made.

The authors in [46] analyze the effectiveness of designing a job matching algorithm with a fair-by-design approach. Under the umbrella of RAI in human resource management (HRM), safeguarding algorithms are introduced alongside the automatic recruitment system to prevent bias and discriminatory outcomes. The authors offer some insights into the challenges associated with achieving fairness, namely that a definition of fairness is multifaceted, contextual and open to subjective perception and importantly, that declarative methods used in recruitment including surveys, questionnaires and tests have a history of reliability issues when it comes to predicting future job performance. Employing an RAI solution in an already flawed HRM system reveals some of the limitations that can be expected when applying a technical solution in a complex philosophical and sociological space. Despite these limitations the authors test the outcomes of their model based on the Equality of Opportunity fairness definition. The researcher's safeguarding methods include testing the matching algorithm for bias across protected and sensitive attributes of job applicants.

Defining transparency and explainability requirements in practice are the focus of research efforts in [47]. Balasubramaniam et al. offer a model, a template and lessons learned when applying the explainability template which offers a systematic way to define explainability as a non-functional requirement for AI systems. The explainability model that the authors present asks practitioners to answer "1) to whom to explain, 2) what to explain, 3) in what kind of situation to explain, and 4) who explains." [47]

Statement of the Purpose

This research examines the degree to which integrating an RAI system with a recruitment system can be implemented in the public sector and evaluates RAI adoption barriers. This case study involves the experiences of a cross-functional working group while they attempt to operationalize ethical principles found in RAI. Given the contextual nature of ethics and legal systems, a case study grounds assessments of feasibility, organizational readiness and practical guidelines in a specific time and place. The working group is tasked with integrating an automated redaction system that targets personally identifiable information (PII) in résumés with existing recruitment software so that process efficiencies can be gained while minimizing the impact of conscious, or unconscious bias during the hiring process.

CHAPTER III

METHODOLOGY

Overview

An action research (AR) methodology placed a researcher alongside a public sector working group who sought to integrate automated redaction of PII in résumés with an existing recruitment system. The cyclical process of action research methodology included finding a problem, exploring problem solutions through interventions (action plans) and evaluating the effectiveness of the intervention with a reflective activity. These AR cycles were carried out in synchronicity with the working group's agile sprint ceremonies. A series of workshops at the beginning of every cycle provided opportunities for the team to explore existing problems that are unique to different stages of the software development lifecycle [48]. A research paper [32] was used to reference new processes, tools, architectural and system-level considerations that the team could action as they navigated the solution to the identified problem. At the end of each cycle, reflective exercises were conducted to assess the experiences and overall impact using a questionnaire based on Costello et al. which is included for reference in Appendix A [49].

Research Design

The original research design scheduled five voluntary workshops over the course of four months. Each of the workshops introduced RAI tools and processes relevant to a phase in the software development lifecycle (SDL) that the team was focusing on. These phases include requirements, design, implementation, verification/validation and operation. The effectiveness and relevance of the tools was captured by collecting feedback from the group through a questionnaire. At the end of each action research cycle

observations of group experiences were evaluated against the relevant process area in an AI maturity model (AIMM). Cho et al.'s maturity model for Trustworthy AI is included for reference in Appendix B [14]. The model defines one or more base practices for each of the process areas and observations of the team provided evidence of these base practices being present as well as an evaluation of the capability dimension according to the five levels in the Capability Maturity Model (CMM) [50]. The CMM is a well-known metric to assess software process improvement; the levels are initial (ad hoc, L1), repeatable (policies established, L2), defined (documented, L3), managed (quantitative goals are set, L4) and optimizing (focused on process improvement, L5) [51]. An incremental numeric value is used (L1-L5) to reflect the primary researcher's assessment of the capability level for each of the base practices. If evidence of the base practice was not observable it is indicated with an 'x' and the capability level is equal to a zero value.

Pivot in Research Design

Rationale for the pivot: The working group was able to participate in two of the five action research cycles (requirements and design) over three and half months, though ended up pivoting away from AI work multiple times to focus on other priorities. When the team did prioritize AI work, barriers were encountered that contributed to a failure to integrate RAI with the existing recruitment system.

How data collection changed: The experiences and feedback of the group during those two action research cycles through observations and questionnaires informs a partial assessment of the feasibility of operationalizing an RAI system. As the working group pivoted away from their initial commitment to RAI work, the research focus also pivoted towards discovering barriers to adopting RAI. After the group confirmed they would no

longer be working on an AI implementation, they were asked to complete a survey that focused on barriers to adoption. The survey is included for reference in Appendix C and covers five adoption categories commonly found in frameworks for analyzing technology adoption barriers including (1) individual characteristics such as perceived usefulness, (2) leadership characteristics such as attitude and innovation, (3) environmental characteristics such as external pressures and government regulation, (4) organizational characteristics such as IT expertise and size and finally (5) innovation characteristics such as relative advantage and compatibility [24].

Reflection on research question impact: Pivoting away from AI work during the research period presented a significant limitation to the intent of the original research design. However, what that choice says about feasibility is quite strong. The group's pivot provides support for answering RQ1 in the following way: Implementation of an RAI system in the public sector is feasible to the extent that it can be prioritized. However, the experiences of the group does not begin and end at prioritization and a more holistic analysis can still be made from some of the real challenges faced during implementation. The impact of the pivot on RQ2 is seemingly less significant as the question scope is broader; a theoretical approach to operationalizing ethical principles can augment the analysis of empirical evidence from the two action research cycles which focused on the theory-to-practice gap. To address the limitations that the pivot imposed on the first two research questions, a third research question (RQ3) about general adoption barriers provided insights into what barriers prevented the prioritization and adoption of RAI.

Strengths and Limitations

Limitations of the original research design includes assuming the team was both committed to prioritizing RAI work and had the necessary proficiency in foundational skills to carry out the work. The original research design also assumes that the work will be carried out sequentially through each of the five software development phases. The research deadline required a cadence that may not have provided the flexibility for discovery, research spikes, or prioritizing other work. In an agile context the software development lifecycle (SDL) is considered cyclical, with each of the SDL phases being visited multiple times to build out new features. A linear SDL is reminiscent of the waterfall methodology, which has long since been considered a current trend in software development [52]. Overlaying a linear, fixed cadence research design on top of a cyclical software development process can create tension between a flexible process and a rigid structure.

CHAPTER IV

ACTION RESEARCH PROCESS AND ANALYSIS

Cycle 1 — Problem Exploration: Requirements Elicitation

For the first cycle, the group explored the problem "*How can high-level ethical principles translate into practical features".* A new tool, ECCOLA (ethical user stories), was presented prior to asking the group to generate user stories relevant to their work [48]. Software requirements is one of the process dimensions of an AI maturity model [14] and is a skill area in the Software Engineering Body of Knowledge (SWEBOK) [53]. Proficiency levels for software requirements activities are defined in the Software Engineering Competency Model (SWECOM) [54] as a range with low to high levels presented sequentially as 1) following, 2) assisting, 3) performing, 4) leading and 5) creating. The group was asked to assisting with preparation of requirements which is a software requirements activity that requires a low level of proficiency [54] and should not have presented any major obstacles for a group with emerging competencies. In total seventeen user stories were created, then grouped thematically and prioritized. The remaining prioritized user stories reflected the work that could have been actioned in the upcoming sprint cycle.

Another example of how ethical principles translate into practical features was presented to the group as a diagram comparing two user flows for uploading a résumé. The comparison highlighted how user flow changes when an ethical principle, in this case transparency, is treated as a requirement; the intent was to further establish the connection between an ethical principle and its translation as a practical feature. Overall, the problem exploration workshop introduced a practical tool (ECCOLA), presented a relevant

business example (user flow) and offered alternative options for determining requirements when ethical principles are in scope. The materials presented to the group are available in Appendix D and E.

Cycle 1 — Actions

Though ethical user stories were created in the workshop using ECCOLA they were broadly defined and in need of refinement. The most frequent ethical principles that all users stories reflected were transparency, fairness and accuracy. These concepts needed clarification, as well as discovering which named entities mapped to specific harms and identifying measurements of fairness. The team's understanding of specific ethical principles would directly influence the features that could be built, so drilling down into the details of what the user stories elicited was necessary. To refine the requirements elicited in the user stories and push the discussion forward the following questions were posed to the team in Cycle 1:

a) **What does a fair hiring process mean to the group and how will the success of fairness be measured?**
b) **What other entities besides names, locations and organizations do we want to consider?**
c) **Can we enumerate the types of bias that we're attempting to mitigate against? Are there any other potential sources of bias or that a hiring manager could use as a proxy to infer personal or protected characteristics?**
d) **Is fairness defined by specific (measurable) outcomes that would demonstrate fairness success?**

a) Fairness Refinement Requirements

The group was presented with different ways that fair could be interpreted with respect to recruitment. As an example, the following takes into consideration one protected category (Male/Female) from the Canadian Human Rights Commission[38].

- *Group Unaware fairness*: What's fair is to disregard/redact the protected attributes (such as what can be inferred from a Name) and choose candidates purely on their merits.
- *Group Threshold fairness*: What's fair is to look at the historical biases reflected in the data between females/males and modify the algorithm to afford a more lenient predictive model for females due to the likelihood that their work histories may be interrupted by multi-month gaps, in the case of parental leave, for instance.
- *Demographic parity*: If 30% of the job applicants for a job are female, then 30% of the interview pool ought to be female.
- *Equal Opportunity*: The number of candidates invited to be part of the interview pool are 50% male and 50% female, regardless of whether or not there was a 50/50 split on applicants for the job.

b) Named Entity Mapping Requirements

The team was asked about which named entities it wanted to redact. To help make a connection, they were contextualized according to the type of harms they could potentially be linked to. With the redaction system, the aim was to mitigate the risk that hiring managers could apply their own bias (consciously or unconsciously) by limiting

[38] CHRC, What is discrimination?, https://www.chrc-ccdp.gc.ca/en/about-human-rights/what-discrimination, accessed Oct 18, 2023

direct identifiers and quasi-identifiers where they can be used to infer attributes. Table V maps named entities to potential harms and was included in the communication to the team.

Table V: Mapping Named Entities to Potential Disclosure Harms

Label	**Link to Potential harms**
PERSON	*Know*: Name, title, username/alias, initials, nickname *Infer*: Gender and/or sex, marital status, national or ethnic origin, online profile
GPE	*Know*: Place of birth, cities of employment or schooling *Infer*: National or ethnic origin, socioeconomic status
LOC	*Know:* Areas, recognizable locations, place of residence, education, birth *Infer:* National or ethnic origin, socioeconomic status
FAC	*Know:* Schools attended, places of work *Infer:* National or ethnic origin, socioeconomic status
ORG	*Know:* Name of school, place of work, volunteer activities, place of worship *Infer:* Religion, political affiliation, national, racial or ethnic origin
DATE	*Know:* Date of birth, cumulative years of employment. *Infer:* Age of candidate
DEM	*Know:* Native language, descent, heritage, ethnicity, job titles, ranks, education, physical descriptions, diagnosis, birthmarks, ages *Infer:* National or ethnic origin, race, disability, age, economic and health status
CODE	*Know:* Phone, passport number, area code, country code *Infer:* Region of residence, national origin
NORP	*Know:* Nationalities, religious or political groups *Infer:* Religion, political affiliation, national, racial or ethnic origin
ADDRESS	*Know:* Place of residence *Infer:* Socioeconomic status

c) Bias Enumeration Requirements

Minimizing bias is one of the goals in RAI though a narrow definition based on statistical accuracy can miss the larger context in which structural inequalities persist [55]. Removing statistical bias in a dataset is not the only outcome that needs to be considered; by conceptualizing the problem that bias presents in the context of social

inequality and expanding the definition of bias, RAI systems can potentially have a larger effect towards shifting existing power structures. The team was asked about the types of bias it was attempting to mitigate against. For instance, were they concerned with ageism, ableism, or discrimination by affiliation? Many résumés include a 'Personal Interests' section, where candidates state what they like to do in their spare time which could include belonging to a particular group, whereby one can infer either human behaviour, group affiliation or personal characteristics. To accompany the communication which was intended to further refinement the fairness requirements, a copy of Richardson and Gilbert's diagram for a taxonomy of currently existing forms of bias [56] accompanied the communication which is available in Appendix F.

d) Fairness Measurement Requirements

The communication included a curated list of fairness toolkits preceded by a caveat that reducing fairness to a quantifiable value may be overly reductionistic in producing different types of fairness measurements based on fairness indicators (i.e. statistical parity, etc.). While open source toolkits are plentiful, there are important differences in their capabilities and interpretations of the concept of fairness which highlights both their limitations and opportunities for improvement as well as the significance of allowing a context-specific use case to drive toolkit use [57]. The list of toolkits sent to the group is presented in Table VI.

Table VI. Fairness Software Toolkits

Name	Organization	Language
Fairness Toolkit[39]	Google	Python (TF), can evaluate non-TF models with Apache Beam SDK[40]
What-If Tool[41]	Google	Python (TensorFlow)
Aequitas[42]	University of Chicago, Data Science for Social Good	Python package
AI Fairness 360[43]	IBM	Python and R
LiFT[44]	LinkedIn	Scala/Spark
ML Fairness Gym[45]	Google	Python
fairlearn[46]	Fairlearn organization (community driven), origins in Microsoft research	Python package
Audit-AI[47]	PyMetrics	Python package
FairML[48]	Julius Adebayo	Python and R package
InterpretML[49]	Microsoft Research	Python
LIME[50]	Marco Ribeiro	Python package
SHAP[51]	Scott Lundberg, University of Washington, Microsoft Research	Python package

Cycle 1 — Analysis

Action: Ethical User Story Requirements

The first workshop provided an opportunity for the group to learn how to use ECCOLA to generate user stories with ethical principles in mind, yet a choice was made not to employ ECCOLA during the action phase, at least not to generate user stories. Instead, the group chose to extract the ethical principles found in ECCOLA to re-define what the

[39] Google, TensorFlow Fairness Indicators, https://www.tensorflow.org/tfx/guide/fairness_indicators, accessed Oct 18, 2023
[40] Google, TensorFlow Fairness Indicators with non-TF models, https://www.tensorflow.org/tfx/guide/fairness_indicators#using_fairness_indicators_with_non-tensorflow_models, accessed Oct 18, 2023
[41] Google, What-If Tool, https://pair-code.github.io/what-if-tool/, accessed Oct 18, 2023
[42] Aequitas, https://github.com/dssg/aequitas, accessed Oct 18, 2023
[43] IBM, AI Fairness 360, https://github.com/Trusted-AI/AIF360, accessed Oct 18, 2023
[44] LinkedIn, LiFT, https://github.com/linkedin/LiFT, accessed Oct 18, 2023
[45] Google, ML Fairness Gym, https://github.com/google/ml-fairness-gym, accessed Oct 18, 2023
[46] Fairlearn, https://github.com/fairlearn/fairlearn/, accessed Oct 18, 2023
[47] PyMetrics, Audit-AI, https://github.com/pymetrics/audit-ai, accessed Oct 18, 2023
[48] FairML, https://github.com/adebayoj/fairml, accessed Oct 18, 2023
[49] Microsoft, InterpretML, https://github.com/interpretml/interpret, accessed Oct 18, 2023
[50] LIME, https://github.com/marcotcr/lime, accessed Oct 18, 2023
[51] SHAP, https://github.com/slundberg/shap, accessed Oct 18, 2023

principles meant to them in a recruitment context which would then inform a set of team-based, internal guidelines. The group utilized a tool designed for bridging the gap between principles and practical application in an effort to elaborate on existing ethical principle definitions, however, this misaligns resources towards an unintended objective. In addition, the work was not completed.

The group's decision to not use ECCOLA was likely due to their lower prioritization of AI work, rather than an indication of the tool's usefulness for integrating ethics into the software development process. However, deciding to use the principles from ECCOLA to draft team guidelines reflects a curious deviance from using the tool for what it was intended. It also represents a redundant exercise; internal team guidelines would eventually be superseded by forthcoming government policy. The proficiency level required to carry out the problem solution was entry level, bringing into question whether team competencies were sufficient or if the direction was misguided. Alternatively, the introduction of ethical principles as requirements may have presented a complex task which introduced more uncertainty than the group knew what to do with. However, given that a standards-based product specification document was rejected and no other requirements elicitation process was employed opens up questions about how receptive the group was to new tools and how new processes may be perceived as a threat to their preferred and established work methods.

Action: Fairness Refinement Requirements

After the communication was sent out on fairness, there was some consensus from the group that they were pursuing an idea of fairness that aligned with the concept of ‘group unaware fairness’. Tension was observed when discussions brought to the surface

different perspectives on the overall objective of the redaction project. Achieving more diversity in hiring was presented as a primary goal and was perceived to be at odds with the intention of meritocracy for which 'group unaware fairness' is designed to support. Though tensions were observed in discussions about overall team goals these types of discussions reflect a healthy and necessary step in navigating team alignment with complex issues.

Action: Named Entity Mapping Requirements

The group was unable to establish requirements for which named entities in a résumé it wanted to target. Initially there was some confidence expressed in knowing that names, organizations and place of origin could be targeted for redaction with most pre-trained models, however, when other entities were introduced and subsequently mapped to harms the group became less clear on what was desirable. Adding to the group's indecision was the realization of the trade-off between information loss and harm mitigation when selecting more entities for redaction. For example, to mitigate against the risk of ageism, DATE entities may be targeted so that date of birth and/or years of experience cannot be used to either calculate age or infer age based on years of experience. However, job requirements might state that a minimum number of years of experience is necessary. Hence, the trade-off of information loss over harm mitigation.

The requirements for named entities that the group wanted to target but failed to identify, had many downstream technical implications which proved to be a significant implementation barrier. Given that these requirements were never established, none of the subsequent work could be completed. Should the named entity requirements not match what was available from any available pre-trained models then fine-tuning would be

necessary. Subsequent downstream impacts from that decision point includes considerations of which annotated dataset could be used to fine-tune the model and whether the infrastructure available would be sufficient to support the model training process.

Action: Bias Enumeration Requirements

The group was unable to enumerate the types of bias it was wanting to mitigate against, resulting in a lack of subsequent discussions; this result may be attributed to their prioritization of other work. However, some of the feedback indicates members feeling overwhelmed by the impact that these decisions may have on prospective candidates. That feeling coupled with some of the previously identified lack of clarity around diversity versus meritocracy goals might have compounded feelings of inertia.

Action: Fairness Measurement Requirements

After the introduction of fairness toolkits, there was no response or discussion that followed from the team indicating which tool they wanted to use, if any, or which measurement was important to the project. The group's decision to commit to work other than the AI integration project offers one possible explanation for why a response to the fairness measurement requirement was never produced.

Cycle 2 — Problem Exploration: Design

The problem exploration for Cycle 2 was introduced as "*how can we incorporate responsible-AI-by-design into AI systems through system-level architectural style, patterns and techniques?*". Software architecture design is one of the process dimensions of an AI maturity model [14] and is a skill area in the Software Engineering Body of Knowledge (SWEBOK) [53]. To demonstrate the integration of an AI redaction system

with existing recruitment software in the private cloud infrastructure familiar to the group, a reference architecture design was presented and is available in APPENDIX G. The architecture demonstrates how the integration of an annotation editor could be accomplished while reflecting the ethical principle of human-control in a system design. Since most of the team was unfamiliar with what 'human-in-the-loop' could mean in the context of verifying and correcting named entities for redaction, as well as being new to what role annotated datasets play in fine-tuning a pre-trained NER model, a demonstration of an opensource annotation editor[52] provided a real-world example of an otherwise abstract notion. After the introduction of a system design diagram and demonstration of an annotation editor, alternative responsible-AI-by-design architectural patterns in [32] were presented to the group. Emphasis was put on the connection between how different ethical principles effect system architecture. Figure 1 displays the architectural patterns for RAI by Lu et al.

Pattern name	AI mode switcher	Ethical sandbox	Multi-model predictor	Federated learner	Ethical black box	Cross-system auditor	Co-versioning registry	Continuous validator
Objective	Trustworthiness	Trustworthiness	Trustworthiness	Trustworthiness	Trust	Trust	Trust	Trustworthiness
Target users	Architects	Archtiects	Architects	Architects	Operators	Operators	Data scientists	Operators
Impacted stakeholders	Employees, management, consumers, regulators	Employees, management, consumers, regulators	Employees, management, consumers, regulators	Employees, management, consumers, regulators	Employees, management, consumers, regulators	Employees, management, consumers, regulators	Employees, management, consumers, regulators	Employees, management, consumers, regulators
Lifecycle stages	Design	Design	Design	Design	Operation	Operation	Operation	Operation
Privacy	N/A	Yes	N/A	Yes	Yes	Yes	N/A	Yes
Accountability	N/A	N/A	N/A	N/A	Yes	Yes	Yes	Yes
Safety&security	Yes	Yes	Yes	Yes	Yes	Yes	N/A	Yes
Trans. & expla.	N/A	N/A	N/A	N/A	Yes	Yes	Yes	Yes
Fairness	N/A	Yes	Yes	N/A	Yes	Yes	N/A	Yes
Human control	Yes	N/A	N/A	N/A	Yes	Yes	N/A	Yes
Human values	Yes	Yes	N/A	N/A	Yes	Yes	N/A	Yes
Mechanism	Invocation, dismissal, kill switch, override, fallback	Ethical margin, watchdog	N/A	N/A	Immutable log	Immutable log	N/A	Rebuild alert
Applicability	High	High	High	Medium	High	Low	High	High
Benefits	Increased safety, contestability and autonomy	Isolated environment, increased security and safety	Enabled cross-validation and fault-tolerance	Increased privacy and reliability	Enabled accountability and regulatory conformance	Enabled global-view accountability	Enabled different levels of co-versioning	Enabled run-time alert and explicit mitigation
Drawbacks	Limited by quality of the non-AI component, limited in real-time	Difficult to use for hard-to-quantify ethical risks, performance penalty	Increased development effort, require specialised skills	Increased development effort, potential sampling bias, performance penalty	Performance penalty, Increased development effort.	Potential privacy breach risk, performance penalty	Increased storage cost	Difficult to use for hard-to-quantify ethical risks
Known uses	Tesla autopilot [98], Waymo [105], Baidu autonomous mini-bus [7]	Safe-visor [65], Fastcase AI Sandbox [44], AI Sandbox [87]	TFUtils [81], Kubeflow [62], AWS SageMaker [5]	TensorFlow Federated [97], LEAF [14], FATE [45]	Idependent audits for safety [43], ethical black box [107], distributed blakbox [42]	Global-view accountability [76], blackbox for XAI [1]	DVC [23], Mlflow Model Registry [21], Replicate.ai [86]	Neptune [80], AWS SageMaker Model Monitor [6], Qualdo [88]

[52] React NLP Annotate, https://github.com/UniversalDataTool/react-nlp-annotate, accessed Oct 18, 2023

Figure 1. Architectural patterns for RAI by Lu et al. [32]

Similar to software requirements activities, proficiency levels for software design are defined in the Software Engineering Competency Model (SWECOM) [54] as a range with low to high levels. The group was asked to assist in software design reviews which is consistent with a software competency that requires entry level proficiency and should have avoided any barriers for participation due to varying levels of competency in the group [54]. Since the proposed architecture design (Appendix G) had to make assumptions about what the requirements actually were, the task for the group was to come up with which ethical principles would be prioritized in the actual building of the automated redaction system. Though the presentation was primarily focused on technical architecture, the workshop communicated how responsible-AI-by-design applies to different non-technical disciplines as well. For example, if human control was determined to be a high priority ethical principle how would the user interface incorporate an 'opt out' option for potential candidates who choose not to have their résumé data used for fine-tuning a model? The team was encouraged to take action during the sprint that also included familiar tools and processes for service design, UX/UI design and content design.

Cycle 2 — Actions

At the beginning of the problem solution phase a decision was made by the group to pivot away from AI work and prioritize all work related to the launching of the team's recruitment site. As such, any AI related work was removed from the sprint backlog and the actions for Cycle 2 were put on hold. After the website launch, the group picked up where it left off with AI work. The design work created after the launch of the recruitment website included:

- designing a user workflow for a candidate uploading their résumé,
- validating assumptions made with the initial system architecture design, Presidio deployment as an alternative prototype
- content design for opting out or providing consent prior to uploading their résumé
- content design for candidates to learn more about how AI would be used in the recruitment process.
- Redact PII in PDFs

Redact PII in PDFs

The work ticket titled “redact PII in PDFs” was created by the team yet the scope of the work as it was articulated initially was outside of design considerations. While it described a feature request it failed to address essential requirements or feature specifications making implementation open to interpretation. To bring design back in to focus so that solutions could be applied to the problem of incorporating responsible-AI-by-design, the primary researcher provided a product specification template using content from sample templates in the ISO/IEC/IEEE International Standard [58].

Presidio Deployment

To demonstrate an alternative system design that did not involve fine-tuning a pre-trained model yet could facilitate a human-in-the-loop workflow for redaction, an instance Presidio[53] was modified[54], configured and deployed to a development namespace in the team’s OpenShift environment. Presidio is an open source Microsoft product

[53] Presidio – Data Protection and De-identification SDK, https://github.com/microsoft/presidio, accessed Oct 18, 2023
[54] Modified Helm charts for Presidio deployment to OCP, https://github.com/bdolor/charts, accessed Oct 18, 2023

designed for privacy protection that comes with a Helm chart for deploying to Kubernetes clusters. Once deployed, REST APIs sit in front of a pre-trained NER model that can identify and anonymize PII in text.

Cycle 2 — Analysis

The second workshop provided an opportunity to establish connections between responsible-AI-by-design and different types of work. Multiple demonstrations of how ethical principles effect system architecture were presented to the group, including a prototype of an annotation editor and a proposed system architecture design. Deployment of an instance of Presidio acted as a demonstration of an alternative system architecture which briefly engaged another member of the technical team, yet other than the primary researcher, team members did not complete design work during Cycle 2. The team pivoted away from RAI related work twice during Cycle 2 which reduced the amount of time available, making a successful implementation prior to the research deadline unrealistic.

Action: Redact PII in PDFs

Though the product specification document could have been used to design specifications of the feature “redact PII in PDFs” it was ultimately rejected with a claim that it did not align with the way the team worked though no alternative design tool or process was offered. Design specifications remained undetermined. A definition of done, which is a list of agreed upon criteria to guide the work simply stated ‘input a résumé, spit out a redacted résumé’ which offered no further clarity. Similar to what was discovered in Cycle 1 further refinement was necessary including but not limited to knowing which entities were to be redacted. Despite these unknowns a demonstration of redacting PII in

PDFs was presented to the group on two separate occasions. A sample Python script in the second demonstration was modified to iterate through 100 sample résumés and handle multi-page PDFs. A possible explanation for a lack of engagement from the team, despite significant individual effort is that the level of engagement required more capability and commitment than the team was willing to give.

Action: Presidio Deployment

The prototype of Presidio gave the team an opportunity to become familiar with a working example of a web based, text anonymization application. Integration points were made available with the application's REST APIs which included out-of-the-box entity coverage and performance metrics for processed text. Documentation for the application is thorough[55]. As well, a quick deployment could de-mystify preconceived notions that AI was out of reach and did not necessarily require any specific capabilities that the team did not already possess. Though the deployment engaged members of the technical team beyond the primary researcher, it failed to elicit much response from other team members.

[55] Presidio Documentation, https://microsoft.github.io/presidio/, accessed Oct 18, 2023

CHAPTER V

RESULTS

Reflection: Cycle 1 — Questionnaire

A facilitated workshop followed the action implementation for Cycle 1 where feedback was collected from the questionnaire in Appendix A. A thematic qualitative analysis [59] of the feedback identified the following themes from participants for Cycle 1.

Theme 1: (barrier) Engagement: A number of respondents cited forgetting about the RAI work, or plainly stated that they did not make time for the work, or were busy with other priorities. Some responses claimed they did not know how to contribute or expressed confusion around what the main action was. Negative sentiments included a proclamation that the primary researcher should "take their independent researcher hat off" and push the group to "pay more attention", followed by a claim that the lack of engagement was the responsibility of the primary researcher. A small number of statements reflected a desire for more time.

Theme 2 (barrier) Distrust of AI: One comment asserted their "distrust of technology as a solution to relationship problems" while others expressed intimidation with the AI problem space. Not knowing very much about RAI was expressed as a concern and especially in the context of how decisions may impact people's lives. Some unease was revealed with comments indicating how far behind government is on this topic and that industry may end up defining ethical use of AI. One response reflected a desire to see AI used in only low risk applications in order to avoid negative, unintended consequences.

Maturity Model Evaluation: Cycle 1

For each base practice that was present an evaluation of capability was determined, based on the five levels in the Capability Maturity Model (CMM) which includes initial (ad hoc), repeatable (policies established), defined (documented), managed (quantitative goals are set) and optimizing (focused on process improvement) [50]. Evaluating the team's overall capability for software requirements analysis against a Trusthworthy AIMM generates a low maturity score, with numerous incomplete capabilities, as presented in Table VII.

Table VII Software requirements evaluation against an AI maturity model [14]

Process Area	Process	Base Practice	Capability Dimension	Evidence
Plan and Design	Software Requirements Analysis (SRA)	Define development goals	L1	Capability for setting development goals was a characteristic of an individual and not the team
		Define function and performance	x	
		Function/performance evaluation metric	x	
		Functional and behavioral evaluation metric	x	
		Fairness risk factor analysis	L1	Fairness discussions regarding diversity and meritocracy reflects ad hoc fairness analysis
		Fairness evaluation metric	x	
		Stakeholder review—fairness	x	
		Safety risk factor analysis	L1	Human-in-the-loop presented as a safety check for model error
		Safety evaluation metric	x	
		Safety countermeasures	x	
		Stakeholder review—safety	x	

Reflection: Cycle 2 — Questionnaire

A facilitated workshop for Cycle 2 collected feedback from the questionnaire in Appendix A. A thematic qualitative analysis [59] of the feedback identifies the following themes from participants for Cycle 2.

Theme 1 — (barrier) Organizational Constraints, Roles and Responsibilities. The AI project drew the attention of provincial privacy officers who questioned the permissibility of working on developing an AI system without establishing that privacy violations had occurred and without an official AI policy that would grant them the authority to do so. The absence of an organizational policy related to AI places the burden of AI governance on existing policies and the subsequent enforcement of those policies left to their relevant organizational roles. One possible explanation is the extent to which hype around the recently released ChatGPT[56] adds to a climate of heightened concern with respect to AI while potentially conflating an already unclear definition of AI with generative AI. The existence of numerous AI projects already in progress or deployed within the provincial public sector makes attempts to reign in new AI projects, without any established AI policies, appear as a reactive and ad hoc response. In Canada the Privacy Act, Personal Information Protection and Electronic Documents Act (PIPEDA) and Financial Agency of Canada Act support some AI ethical principles [8]; in British Columbia the Freedom of Information and Protection of Privacy Act (FOIPPA) could be used as a legal mechanism to enforce relevant laws that align with ethical principles, though approaching AI governance with privacy considerations alone is insufficient to address the range of potential issues that arise from AI such as a lack of autonomy.

Other concerns included questions about the ability of the Hiring Manager to simply refuse redacted résumés putting the efforts of the team on uncertain ground. The realization that a hiring manager could exercise their autonomy and simply refuse a redacted résumé underscores how an ethical principle can permeate to different aspects of

[56] ChatGPT, https://chat.openai.com/, accessed Oct 18, 2023

design, in this case to different users. Up to this point the assumption was that the principle of autonomy would apply to the prospective candidate, however, perhaps because other users such as the hiring manager were not considered, it was simply assumed they would want a redacted résumé. If a system is designed without verifying whether an AI solution is something that users want, it risks building a system few people will choose to use.

Theme 2 — (barriers) Ethical Concerns. Feedback for Cycle 2 included a question about whether the team could make any claim to a fair hiring process even without AI augmenting decisions. This is an indication that the process of deconstructing the recruitment process revealed larger systemic questions about the overall practice of hiring. Recognizing that manual hiring procedures suffer from a history of questionable integrity, some research supports the argument that focusing on using RAI for efficiency and objectivity puts an emphasis on solving the wrong problem [60].

Some concern around ‘impact’ was expressed in the problem exploration workshop including a desire to measure the success or failure of automated redaction in reducing bias or improving diversity of successful candidates. Wanting to measure the reliability and validity of the redaction system reflected a desire to see fairness expressed as a quantifiable value. One explanation could be cultural as the optics around outcomes tends to be valued; claiming a reduction in bias by a certain percentage carries a different weight than an announcement of a modified process to reduce the potential impact of bias.

Confusion around project goals (diversity versus meritocracy) could explain some of the differences in levels engagement which brings attention to the importance of

establishing and clear goals or carving out time to focus specifically on ethical discussions. During the workshop, some members made note of how the team was acting as if the goals were going to be given to them instead of being designed by them. Specific ethical concerns included a desire to ensure that human supervision would always be part of the system design and a concern about how the use of AI would be perceived by users of the system while obtaining consent from prospective candidates.

Theme 3 — (barriers) Implementation Constraints. The design cycle seemed to elicit some questions about the ability of the team to implement the solution, citing concerns around the maturity of the product team and a failure to incorporate AI into the initial product vision. Establishing that requirements are a necessary pre-requisite for implementation was echoed in the feedback from the group. Some of this could be attributed to the 'newness' of the product team which was a sentiment expressed by the participants. Shortcomings in the required skills of the group were acknowledged citing concerns about who on the team would be able to maintain an RAI system after implementation. Granular concerns about implementation details included which specific file formats for résumés would be accepted, which entities could be redacted and how data is collected and stored. In contrast to the overall lack of engagement, a desire to understand the intricacies of model building and fine-tuning for accuracy was expressed by some members of the team.

The decision to pivot away from AI work demonstrates a choice to designate AI as less important and was a significant barrier to implementation. Task prioritization is considered a team capability measure [61] and the absence of any prioritization criteria is an indication of low proficiency in that skill area; prioritization decisions were subjective

and made without transparency. Focusing on other non-AI related tasks pushes the categorization of AI work from a “do it now” prioritization level to a “do it later” level. However, since the research project had a known scope and end date, prioritizing other tasks moved AI work to a “never do it” prioritization level. This prioritization issue is commonly referred to as an issue starvation problem, whereby a practitioner’s interpretation of business value varies and is not validated against a more formal, open prioritization approach that considers a priority score, aging score and aging weight [62].

Theme 4 — (enablers and barriers) Engagement. Responses reflect engagement levels at opposite ends of the spectrum, ranging from declarations of a successful deployment of Presidio on OpenShift infrastructure to expressing confusion about what participants were expected to do other than attend workshops. Other examples demonstrating disparities between engagement levels of the participants include statements that describe being tasked with unrelated (to AI) work and being concerned about attaching an AI project to cultural and philosophical issues that may change over time.

Maturity Model Evaluation: Cycle 2

Evaluating the team’s overall ability for software architecture design against the process area in an AIMM generates a low score with many incomplete competencies as presented in Table VIII.

Table VIII Software architecture evaluation against an AI maturity model

Process Area	Process	Base Practice	Capability Dimension	Evidence
Plan and Design	Software Architecture Design (SAD)	Model design	x	User-centric design is an established, organizational practice
		User-conscious design	L1	
		Design considering maintenance	x	
		Design considering fairness maintenance	x	Unresolved discussions on fairness definitions impacts how maintenance can be built into the design
		Design considering safety maintenance	x	Group discussions about safety did not occur preventing design considerations

Survey Results and Analysis — Adoption Barriers

After the group pivoted away from AI work the second time, they were asked to complete a survey that focused on barriers to adoption.

Demographic Information

Figure 2. Survey responses: Organizational role

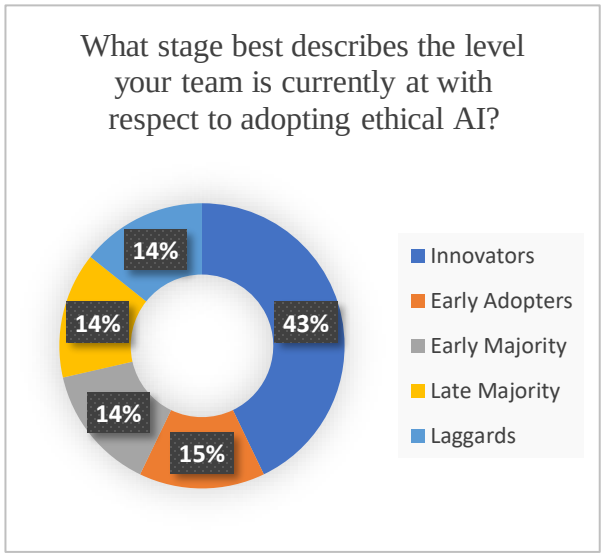

Figure 3. Survey responses: Ethical AI adoption level

Demographic Information: The roles of respondents reflect a balance between technical, administrative and leadership roles. In response to the question "what stage best describes the level your team is currently at with respect to adopting ethical AI" a relationship can be seen between how people responded and what their role is. While leadership roles self-assessed at a higher level of adoption (Innovators) than technical or administrative roles they simultaneously cite practical constraints such as 'lack of funding' and 'official direction', demonstrating a juxtaposition between the idea of AI and the perceived impracticality of implementation. One leadership comment placed emphasis on how RAI 'could' be used while implying that it wouldn't be used outside of an experimental research context. Acknowledging an absence of elements necessary for implementation while maintaining a confident assertion of being at the highest level of ethical AI adoption reflects an overly optimistic view and a departure from an interpretation of adoption that is anything but theoretical. At best, it could reflect a general enthusiasm towards the idea of ethical AI. The lowest level of self-assessment (Laggards) comes from a technical role, which seems to suggest that knowledge of practical implementation details has some effect on the confidence levels associated with assessing adoption.

User Acceptance Factors

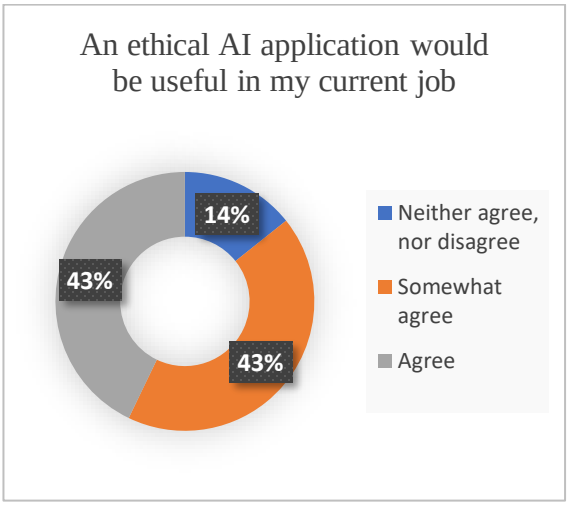

Figure 4. Survey responses: Usefulness of AI

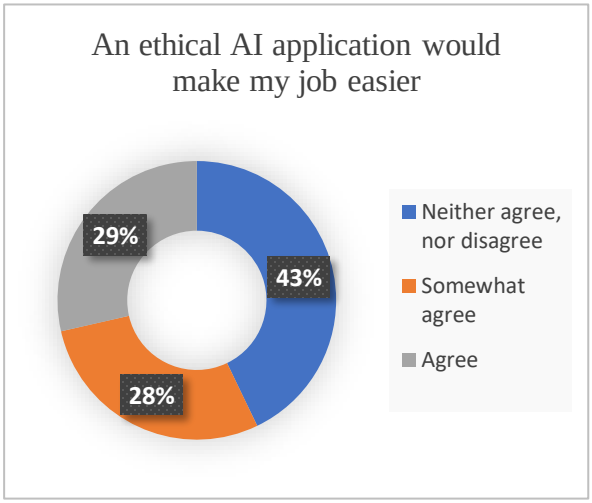

Figure 5. Survey responses: Job ease

User Acceptance Factors: With respect to perceived utility or the usefulness that RAI would have on respondent's jobs there was a general, positive sentiment across all roles. The most negative sentiment reflected a neutral response whereas the majority "agreed" or "somewhat agreed" that an AI application would be useful in their current job. With respect to the perception that RAI would make their job easier nearly half of the respondents provided a neutral response whereas the other half either "agreed" or "somewhat agreed".

Leadership Characteristics

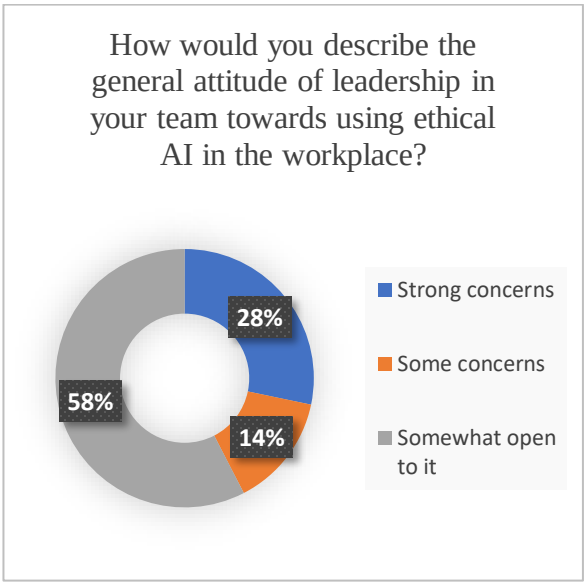

Figure 6. Survey responses: Leadership attitude

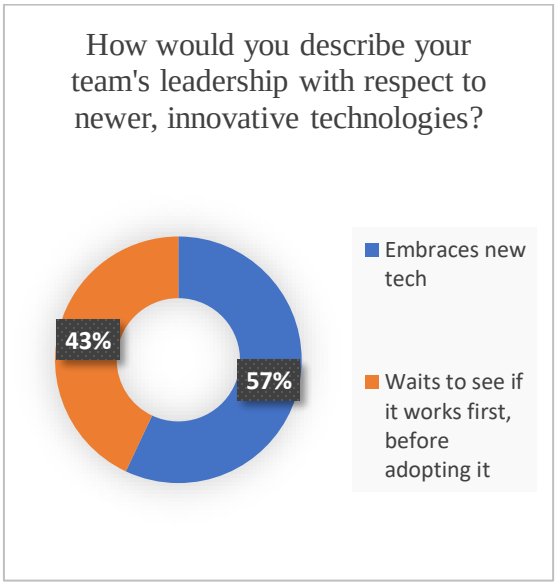

Figure 7. Survey responses: Leadership adoption

Leadership Characteristics: A distinction is evident between the responses from leadership roles and all other roles. Leadership roles responded that there are "concerns" or "strong concerns" with respect to the attitude towards AI whereas all other roles agreed that leadership is "somewhat open" to it. The disparity in perspectives may be attributed to a strategy of optics, where the appearance of being open to AI is advantageous and a distraction that masks addressing real concerns, or a result of a multi-level hierarchical organization where 'leadership' is a relative term depending on who is being reported to. It could be the case that higher levels of 'leadership' have sent a different message than the leadership roles on this team. A more optimistic view of leadership's willingness to embrace new technology also came from non-leadership roles, whereas a more pragmatic 'wait and see if it works' reflected their own perspective.

Environmental Characteristics

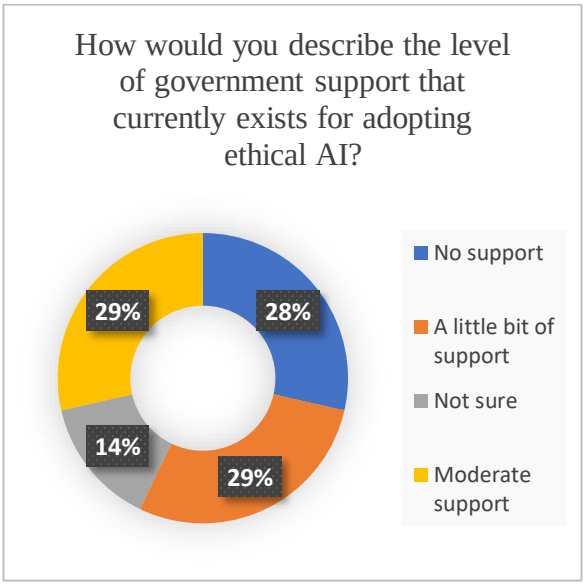

Figure 8. Survey responses: Government support

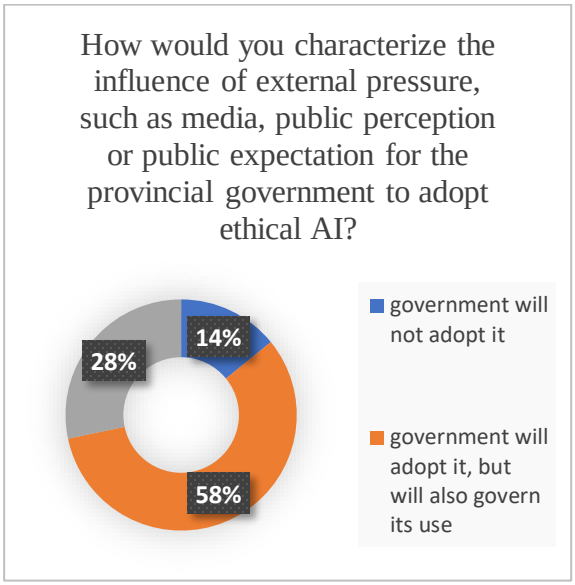

Figure 9. Survey responses: External pressure

Environmental Characteristics: All responses ruled out that government support for adopting RAI was strong. Instead, a fairly equal distribution of responses ranged from "no support" to "little support", "don't know" to "moderate support". The absence of a trend in any particular direction suggests that explicit support for AI has not been made apparent leaving the level of support open to subjective interpretation. Conversely, the majority of respondents felt that public perception was that government would "adopt RAI but also govern its use". Only one respondent thought that government would "not adopt RAI" while the others' claimed they did not know.

IT Expertise

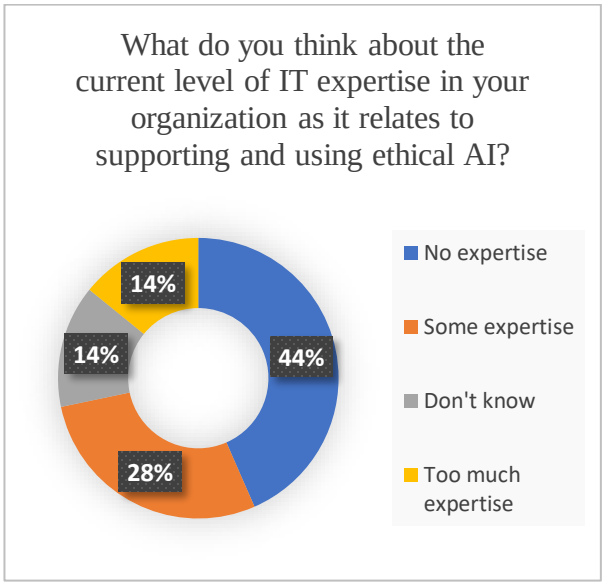

Figure 10. Survey responses: IT expertise

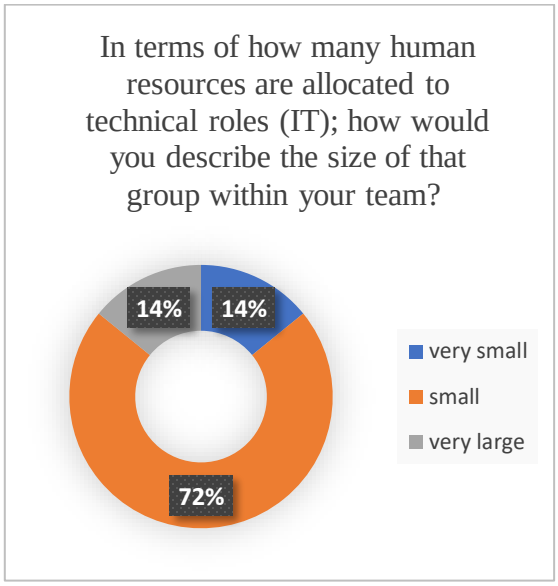

Figure 11. Survey responses: IT size

IT Expertise: Most respondents agreed that the level of RAI expertise in the organization was either "some" or "none". Interestingly one response selected "too much expertise" but it's not clear if this was simply the reaction of a non-technical team member to feeling overwhelmed by the technical aspects of the research project. The vast majority of respondents agreed that the size of technical roles relative to the team was "small" or "very small".

Relative Advantage

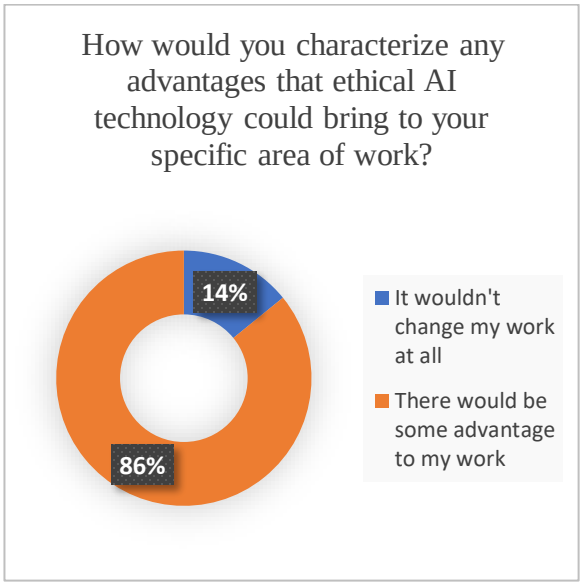

Figure 12. Survey responses: Advantages of AI

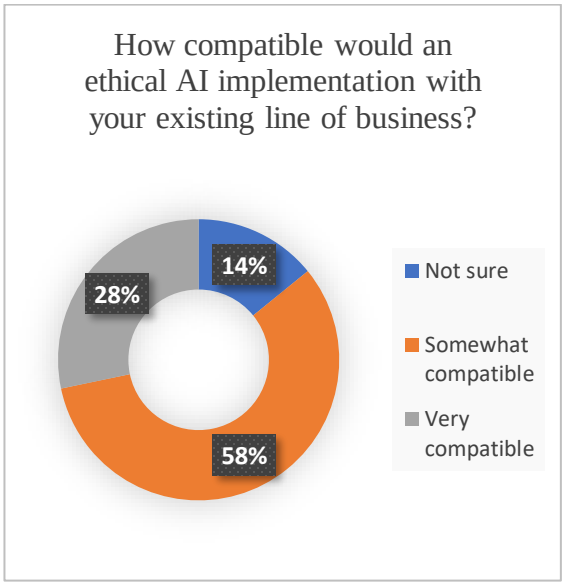

Figure 13. Survey responses: AI compatibility

Relative Advantage: All but one response agreed that RAI would bring some advantage to their work, further commenting that it would help with scaling up services and afford workflow efficiencies. In terms of perceptions on how compatible RAI would be with their line of work, most responded “somewhat compatible” or “very compatible”.

Comments

Other Comments: Responses in the open space for additional comments offered regrets, advice and explanations for the implementation failure. Since the AI project was halted prematurely, some team members expressed a wish to have been able to offer more, citing other work taking precedence, being new to government or simply not understanding RAI well enough or the impact that it could have. Other comments build on the notion that they simply did not know how to proceed and required someone in a dedicated RAI leadership role for guidance and facilitation. One comment takes on an exclusionary tone stating that the main problem can be attributed to how the research sessions did not

"...cater to the personality of the team", or consider team characteristics. In recruitment, terms like 'culture fit' have historically been used as a euphemism to justify exclusion [63] which reflects a departure from expectation especially given the group's effort to acknowledge and reduce potential forms of bias. 'Deficit narratives' reduce people to what they lack by evaluating them against a conformance standard [60]; the notion that there is requirement for a person to 'fit into' or 'cater to' a predominant 'culture', 'team personality', or collection of team characteristics is consistent with masking individual prejudice. The comment invites some examination into what degree bias played a role in the reception of research project and could offer some insight into the reluctance of some members to fully participate.

CHAPTER VI

DISCUSSION

The examination of a technical system viewed through an ethical lens led to an unpacking of manual hiring procedures and an important discovery that they suffer from a history of questionable integrity [60]. Though efficiency gains were the primary business goal, the implementation of RAI towards only that goal would have focused on solving the wrong problem. While it is true that redacting PII in résumés would have a positive impact on a fairer hiring process, it would not fix an already biased hiring procedure.

The implications of the public sector not prioritizing RAI and demonstrating a low level of organizational readiness means that government is not yet prepared, equipped or willing to direct resources towards adopting RAI. One possible explanation, if not merely an indication of catching up to a rapidly evolving technology, is that the costs of delay have been chosen over the costs of implementation. Costs associated with delay can be understood in monetary terms, human rights violations, citizen impact and potential loss of life, whereas costs associated with adoption would require a shift in organizational practices and funding. Factors affecting prioritization include whether a project or initiative is required by policy or law, and at the time of writing there is neither. Prioritization is also affected by whether it represents a whole-of-government priority which suggests that RAI is less of a provincial priority than affordable housing, health care, safe communities, reconciliation, climate change or the economy. Given the established priorities, unless RAI can be utilized to advance one or more of those initiatives, then it is unlikely to attain government priority status without significant individual effort, or an increase in public demand. Negative media attention as is

described in [35] may also have the potential to change the level of prioritization. For now, improving RAI prioritization could look at highlighting the likelihood of lasting benefits, or how it supports existing digital strategies as is outlined in the Canadian federal government's Digital Ambition. Both of these could frame the adoption of RAI as part of a 'good news' narrative.

Recommendation 1: Prioritize RAI

Research findings indicate that public expectations are consistent with government not only adopting RAI but governing its use. While this external factor is not the only reason for prioritizing the adoption of RAI over keeping it at a low priority state, public expectations should have an impact on government objectives. The choice to prioritize or not could be contextualized as an evaluation of the costs of dealing with the fallout of a news story about an AI application found to be causing harm to citizens [35] versus a positive news story about the adoption of RAI to mitigate harms. Given the relationship between favourable reactions to public sector use of AI and familiarity with AI, different communication strategies could be directed towards specific audiences [37] including special considerations of those whose trust in government is low [38] and for whom general knowledge of AI is emerging. Specific to the use of "high-risk" implementations of AI, such as recruitment, it will be important to consider communication strategies that establish how AI augments decisions rather than makes decisions on its own. Other factors that will impact RAI prioritization are whether it is required by law, and if it can be found to align with current programs or priorities.

Recommendation 2: Risk-based Policy

Both the Canadian and the European Union's legal initiatives acknowledge mitigating potential harms while encouraging responsible innovation. A risk-based approach makes distinctions between different types of AI applications with graduated compliance requirements that impose stricter obligations for higher risk systems. As an example, transparency and explainability requirements would require more robust technical documentation, model explainability, and data governance documentation in a system that informs decisions about recruitment than one that informs decisions about user perceptions using sentiment analysis. If a high-risk system has the potential to cause significant damage even prior to being utilized in the first place, a risk-based policy would require the ability to audit or require a formal review. Auditability of an AI enabled recruitment system could also be considered an extension of explainability [35]. In accordance with the EU AI Act, a risk-based policy may benefit from including explicit bans on certain technologies that pose unacceptable levels of risk. These would include biometric identification systems, biometric categorization systems which use sensitive characteristics, predictive policing systems, emotion recognition systems and facial recognition databases. A risk-based approach would also require the identification of different levels of 'high risk' AI systems and a very clear definition of what AI is, what it is not and what type of risks would categorize an application as 'high-risk' versus low-risk [2]. Determining organization-specific motivations should be considered prior to providing incentives for policy compliance [36].

Recommendation 3: Organizational Structure Shift

A shift in organizational practices may be necessary to support and carry out RAI implementation. Research results indicate a low level of organizational readiness which means that individual effort is likely to be the primary engine behind advocating for organizational change and articulating a clear view of future state [34]. Beyond the precariousness of individual effort, organizational accountability would drive the establishment of clear roles and responsibilities including a governance structure which places an AI review board or ethics advisory group at the top of an organization led by a chief AI officer to provide guidance and evaluate AI projects that have been proposed [64]. Adequate organizational structure plays a role in successfully operationalizing RAI. Functions of this group could include auditing and risk assessments, keeping an inventory of AI initiatives as well as staff training that emphasizes an ethical trajectory for software development. These organizational groups would ensure ethical requirements are embedded into the software development process including accountability, explainability, auditability, robustness and the absence of bias. A sequential series of steps that organizations can consider for operationalizing RAI: “Empower responsible AI leadership, develop principles, policies, and training, establish human and AI governance, conduct responsible AI reviews, integrate tools and methods, and build and test a response plan.” [41] Other proposals include setting up a governance structure that determines whether an AI system is appropriate in the first place, which necessitates making available a veto power and includes whistleblower protections [34].

Recommendation 4: Staff Training

Education initiatives related to the basics of AI, ethics and more advanced developer-centric training will address some of the obstacles encountered in the research project. Data engineering, data science and machine learning operations (MLOps) proficiencies would strengthen an organization's technical capacity to deliver RAI projects. General distrust of AI, concerns about its impact and not knowing enough about it can be addressed through internal, introductory courses of AI sponsored by a digital academy that is tasked with supporting professional development and improving digital literacy skills. Ethics in technology courses could provide a philosophical foundation for navigating the complexities of terms such as fairness. The deconstruction of the recruitment process revealed larger systemic questions about the overall practice of hiring which means that ethics in technology does not require RAI to be relevant, though implies that implementing RAI can be a catalyst for discovery. AI ethics training would be strengthened by making a distinction between instrumental principles and inherent values [6]. Identifying intrinsic values, such as autonomy, justice and beneficence could help contextualize which combination of instrumental principles are best suited for upholding values when confronted with making complex ethical decisions. A framework for high-stakes decision making could complement formal training in ethical decision making [40].

Auditability should be considered as an extension of explainability and human oversight, which is also in alignment with auditability requirements in Bill C-27 and the EU's AI Act. Training could include mechanisms that are required to facilitate an AI system's auditability.

Recommendation 5: Team Composition

Current team composition is not sufficiently resourced for implementing RAI solutions. Maturity models can be used by teams to determine their level of readiness for implementing responsible AI [9] which requires technical and social skills that are unique and not typical of traditional software development teams. Along with roles associated with a traditional software development team in an agile context, AI technologies require roles that include Data Engineers, Data Scientists and Machine Learning Operations (MLOps). Responsible AI may also require expertise from legal professionals, as well as ethicists or sociologists. Similar to the role that a privacy officer plays when complaints arise around privacy issues, a dedicated AI officer role should be identified when end-users are inconvenienced, or AI systems act in unforeseen or unpredictable ways. Developers will likely be required to document more and be required to explain technical concepts in an accessible and understandable way [45]. Risk assessments will need to be conducted by people with technical proficiency including auditing skills, applying tools and processes as well as knowledge of socio-technical harms, legal and ethical constraints.

CHAPTER VII

CONCLUSION

Summary

The feasibility of RAI implementation in the public sector can best be described as the result of a combination of contextual factors, however, the inability of the group to operationalize an RAI project provides evidence of a strong relationship between feasibility and the extent that RAI can be prioritized. Other contributing factors include the capability level of the team, level of AI awareness, an attitude of distrust towards AI, uncertain leadership buy-in and cultural acceptance. The practicality of translating ethical requirements into practical features, facilitated by the introduction of RAI tools and processed was limited by the commitment and proficiency level of the team. Significant barriers to adoption need to be addressed prior to undergoing a full feasibility assessment.

Findings

The main findings of the action research in Cycle 1 are that an absence of AI policy or law contributes to a feeling of uncertainty which has a negative impact on implementation. Other factors negatively affecting implementation include having to navigate complex ethical discussions, limited team capabilities, and feeling overwhelmed by high impact decisions. For success to be measured and to avoid reducing fairness and meritocracy to numerical values, performance indicators may need to incorporate non-quantifiable attributes. Reflective activities for Cycle 1 revealed a lack of team engagement, low team capability levels and a general distrust of AI that diminished the ability of the team to implement RAI.

The main findings of Cycle 2 are the relationship between low prioritization and a reasonable expectation of implementation failure. Reflective activities for Cycle 2 uncovered organizational constraints related to the limitations of existing roles to cover new AI responsibilities, unresolved ethical considerations, implementation constraints related to team maturity, and low engagement issues that revealed a divided focus and poor team cohesion. Maturity model analysis for the planning and design process area highlights many missing capabilities and the presence of low level proficiency levels with respect to best practices in RAI.

The main findings of the adoption barriers survey are that leadership roles tend to portray an overly optimistic view of the level of adoption of RAI while simultaneously expressing "strong concerns" and aligning with a "wait and see if it works" attitude. Comparatively, more technical roles tend towards assessing an organizational level of RAI adoption that is less optimistic due to a more realistic assessment of practical constraints. Technical roles portray an optimistic view of leadership's willingness to embrace new technology that is seemingly unaware of their expressed reservations. Internal perception about government support for RAI is that it is neither strong nor clearly stated, yet public perception expects government to not only adopt RAI but govern its use. Technical expertise specific to AI is relatively low and technical teams are small; both are barriers to RAI adoption. Conversely, the advantages of RAI, its potential benefits, utility and compatibility with existing lines of business are factors that would appear to enable the adoption of RAI.

Limitations

Since there is no guarantee that Bill C-27 in Canada will become law, creating and funding programs on the basis of an anticipated yet still undetermined future poses a degree of risk. Arguably, the likelihood of living in a world where AI proliferates unregulated is less likely than living in one where governments respond to mounting concerns. Similarly, the likelihood of Canada's Bill C-27 interoperability with the scaffolding of the EU AI Act should be weighed against the likelihood that it will be radically different than the EU AI Act, which is on a path to become the first AI-specific law. Nevertheless, this research assumes the likelihood is high that the provincial public sector will be obliged to implement and enforce AI related laws that look similar to what Bill C-27 is now.

The advantage of a case study is that it grounds the research in a specific time and place, creating a frame of reference for many of the otherwise broadly defined legal and ethical concepts that can be understood in a real-life context. One limitation of this case study is how far the findings can be generalized to other public sector teams and to other sectors with different legal jurisdictions. As much as possible, the information provided in the Methodology and Appendices sections could facilitate replication of the study.

Though an assessment of organizational maturity could have been more informed had the group moved the implementation further along, shifting to an evaluation of technological, organizational and environmental barriers helps inform a strategy for what the organization could and should do to prepare for RAI adoption.

Conclusion

At the time of the study, RAI is not being prioritized in the provincial public sector. Organizational readiness is low, with one of the teams demonstrating low proficiency in the necessary capabilities outlined by a responsible AI maturity model. In the provincial public sector government-wide support for RAI is also low; operationalizing RAI was not feasible to the extent that it was not prioritized. Barriers to adoption are identified in broad categories labelled 'Leadership Characteristics' and 'IT Expertise' which indicates a lack of clear direction about the public sector's approach to RAI as well as suggesting that changes to recruitment and training are necessary to address technology gaps. The expectation that government will not only adopt AI but govern its use points to a degree of public pressure towards overcoming barriers to adoption. The deconstruction of recruitment workflows revealed larger systemic questions about fairness in the hiring process that would have otherwise continued to go unnoticed. Other considerations that could have a positive impact on implementation are the creation of risk-based AI policy, shifts in organizational structure, AI and ethics training and a modified team composition that adds RAI specific roles to traditional, cross-functional software development teams.

REFERENCES

[1] B. Shneiderman, “Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems,” *ACM Trans. Interact. Intell. Syst.*, vol. 10, no. 4, pp. 1–31, Dec. 2020, doi: 10.1145/3419764.

[2] N. Díaz-Rodríguez, J. Del Ser, M. Coeckelbergh, M. López De Prado, E. Herrera-Viedma, and F. Herrera, “Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation,” *Inf. Fusion*, vol. 99, p. 101896, Nov. 2023, doi: 10.1016/j.inffus.2023.101896.

[3] L. Munn, “The uselessness of AI ethics,” *AI Ethics*, vol. 3, no. 3, pp. 869–877, Aug. 2023, doi: 10.1007/s43681-022-00209-w.

[4] H. Bleher and M. Braun, “Reflections on Putting AI Ethics into Practice: How Three AI Ethics Approaches Conceptualize Theory and Practice,” *Sci. Eng. Ethics*, vol. 29, no. 3, p. 21, Jun. 2023, doi: 10.1007/s11948-023-00443-3.

[5] J. C. Ibáñez and M. V. Olmeda, “Operationalising AI ethics: how are companies bridging the gap between practice and principles? An exploratory study,” *AI Soc.*, vol. 37, no. 4, pp. 1663–1687, Dec. 2022, doi: 10.1007/s00146-021-01267-0.

[6] C. Canca, “Operationalizing AI ethics principles,” *Commun. ACM*, vol. 63, no. 12, pp. 18–21, Nov. 2020, doi: 10.1145/3430368.

[7] E. Tabassi, “AI Risk Management Framework: AI RMF (1.0),” National Institute of Standards and Technology, Gaithersburg, MD, error: NIST AI 100-1, 2023. doi: 10.6028/NIST.AI.100-1.

[8] L. N. Tidjon and F. Khomh, “The Different Faces of AI Ethics Across the World: A Principle-To-Practice Gap Analysis,” *IEEE Trans. Artif. Intell.*, pp. 1–20, 2022, doi: 10.1109/TAI.2022.3225132.

[9] P. Akbarighatar, “A sociotechnical perspective for responsible AI maturity models: Findings from a mixed-method literature review,” 2023.

[10] “8 Digital technologies and artificial intelligence (AI): implications for using intuition and analytics in personnel selection,” in *Analytics and Intuition in the Process of Selecting Talent*, De Gruyter, 2022, pp. 357–393. doi: 10.1515/9783110980967-008.

[11] A. Akbik, D. Blythe, and R. Vollgraf, “Contextual String Embeddings for Sequence Labeling,” p. 12.

[12] M. Maree, A. B. Kmail, and M. Belkhatir, “Analysis and shortcomings of e-recruitment systems: Towards a semantics-based approach addressing knowledge incompleteness and limited domain coverage,” *J. Inf. Sci.*, vol. 45, no. 6, pp. 713–735, Dec. 2019, doi: 10.1177/0165551518811449.

[13] R. B. Sadiq, N. Safie, A. H. Abd Rahman, and S. Goudarzi, “Artificial intelligence maturity model: a systematic literature review,” *PeerJ Comput. Sci.*, vol. 7, p. e661, Aug. 2021, doi: 10.7717/peerj-cs.661.

[14] S. Cho, I. Kim, J. Kim, H. Woo, and W. Shin, “A Maturity Model for Trustworthy AI Software Development,” *Appl. Sci.*, vol. 13, no. 8, p. 4771, Apr. 2023, doi: 10.3390/app13084771.

[15] D. Hovy and S. L. Spruit, “The Social Impact of Natural Language Processing,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Berlin, Germany: Association for Computational Linguistics, 2016, pp. 591–598. doi: 10.18653/v1/P16-2096.

[16] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, “Language (Technology) is Power: A Critical Survey of ‘Bias’ in NLP,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, 2020, pp. 5454–5476. doi: 10.18653/v1/2020.acl-main.485.

[17] O. Ferrndez, B. R. South, S. Shen, F. J. Friedlin, M. H. Samore, and S. M. Meystre, “Evaluating current automatic de-identification methods with Veteran’s health administration clinical documents,” *BMC Med. Res. Methodol.*, vol. 12, no. 1, pp. 109–124, Jan. 2012, doi: 10.1186/1471-2288-12-109.

[18] B. Dattner, T. Chamorro-Premuzic, R. Buchband, and L. Schettler, “The Legal and Ethical Implications of Using AI in Hiring,” 2019.

[19] F. Hassan, D. Sanchez, J. Soria-Comas, and J. Domingo-Ferrer, “Automatic Anonymization of Textual Documents: Detecting Sensitive Information via Word Embeddings,” in *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, Rotorua, New Zealand: IEEE, Aug. 2019, pp. 358–365. doi: 10.1109/TrustCom/BigDataSE.2019.00055.

[20] J.-M. John-Mathews, D. Cardon, and C. Balagué, “From Reality to World. A Critical Perspective on AI Fairness,” *J. Bus. Ethics*, vol. 178, no. 4, pp. 945–959, Jul. 2022, doi: 10.1007/s10551-022-05055-8.

[21] T. Chamorro-Premuzic, F. Polli, and B. Dattner, “Building Ethical AI for Talent Management,” p. 5, 2019.

[22] I. Pilán, P. Lison, L. Øvrelid, A. Papadopoulou, D. Sánchez, and M. Batet, “The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization,” *Comput. Linguist.*, vol. 48, no. 4, pp. 1053–1101, Dec. 2022, doi: 10.1162/coli_a_00458.

[23] M. I. Merhi, “An Assessment of the Barriers Impacting Responsible Artificial Intelligence,” *Inf. Syst. Front.*, vol. 25, no. 3, pp. 1147–1160, Jun. 2023, doi: 10.1007/s10796-022-10276-3.

[24] M. A. Hameed, S. Counsell, and S. Swift, “A conceptual model for the process of IT innovation adoption in organizations,” *J. Eng. Technol. Manag.*, vol. 29, no. 3, pp. 358–390, Jul. 2012, doi: 10.1016/j.jengtecman.2012.03.007.

[25] F. D. Davis, “Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology,” *MIS Q.*, vol. 13, no. 3, p. 319, Sep. 1989, doi: 10.2307/249008.

[26] M. Fishbein and I. Ajzen, “Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research.” Addison-Wesley, 1975.

[27] I. Ajzen, “The theory of planned behavior,” *Organ. Behav. Hum. Decis. Process.*, vol. 50, no. 2, pp. 179–211, 1991, doi: https://doi.org/10.1016/0749-5978(91)90020-T.

[28] "Diffusion of Innovation Theory." Accessed: Nov. 18, 2022. [Online]. Available: https://sphweb.bumc.bu.edu/otlt/MPH-Modules/SB/BehavioralChangeTheories/BehavioralChangeTheories4.html
[29] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices," *Sci. Eng. Ethics*, vol. 26, no. 4, pp. 2141–2168, Aug. 2020, doi: 10.1007/s11948-019-00165-5.
[30] Q. Lu, L. Zhu, X. Xu, J. Whittle, D. Douglas, and C. Sanderson, "Software engineering for Responsible AI: An empirical study and operationalised patterns," in *2022 IEEE/ACM 44th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, May 2022, pp. 241–242. doi: 10.1145/3510457.3513063.
[31] J. Ayling and A. Chapman, "Putting AI ethics to work: are the tools fit for purpose?," *AI Ethics*, vol. 2, no. 3, pp. 405–429, Aug. 2022, doi: 10.1007/s43681-021-00084-x.
[32] Q. Lu, L. Zhu, X. Xu, J. Whittle, and Z. Xing, "Towards a Roadmap on Software Engineering for Responsible AI," in *2022 IEEE/ACM 1st International Conference on AI Engineering – Software Engineering for AI (CAIN)*, May 2022, pp. 101–112. doi: 10.1145/3522664.3528607.
[33] M. Katell *et al.*, "Toward situated interventions for algorithmic equity: lessons from the field," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona Spain: ACM, Jan. 2020, pp. 45–55. doi: 10.1145/3351095.3372874.
[34] B. Rakova, J. Yang, H. Cramer, and R. Chowdhury, "Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW1, pp. 1–23, Apr. 2021, doi: 10.1145/3449081.
[35] C. Seidelin, T. Moreau, I. Shklovski, and N. Holten Møller, "Auditing Risk Prediction of Long-Term Unemployment," *Proc. ACM Hum.-Comput. Interact.*, vol. 6, no. GROUP, pp. 1–12, Jan. 2022, doi: 10.1145/3492827.
[36] C. I. Gutierrez, "Uncovering Incentives for Implementing AI Governance Programs: Evidence From the Field," *IEEE Trans. Artif. Intell.*, vol. 4, no. 4, pp. 792–798, Aug. 2023, doi: 10.1109/TAI.2022.3171748.
[37] M. F. Gonzalez *et al.*, "Allying with AI? Reactions toward human-based, AI/ML-based, and augmented hiring processes," *Comput. Hum. Behav.*, vol. 130, p. 107179, May 2022, doi: 10.1016/j.chb.2022.107179.
[38] B. Kleizen, W. Van Dooren, K. Verhoest, and E. Tan, "Do citizens trust trustworthy artificial intelligence? Experimental evidence on the limits of ethical AI measures in government," *Gov. Inf. Q.*, p. 101834, Jun. 2023, doi: 10.1016/j.giq.2023.101834.
[39] G. Maragno, L. Tangi, L. Gastaldi, and M. Benedetti, "Exploring the factors, affordances and constraints outlining the implementation of Artificial Intelligence in public sector organizations," *Int. J. Inf. Manag.*, vol. 73, p. 102686, Dec. 2023, doi: 10.1016/j.ijinfomgt.2023.102686.
[40] D. Saxena, K. Badillo-Urquiola, P. J. Wisniewski, and S. Guha, "A Framework of High-Stakes Algorithmic Decision-Making for the Public Sector Developed through a

Case Study of Child-Welfare," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW2, pp. 1–41, Oct. 2021, doi: 10.1145/3476089.
[41] L. Cheng, K. R. Varshney, and H. Liu, "Socially Responsible AI Algorithms: Issues, Purposes, and Challenges," *J. Artif. Intell. Res.*, vol. 71, pp. 1137–1181, Aug. 2021, doi: 10.1613/jair.1.12814.
[42] A. Ananjeva, J. S. Persson, and A. Bruun, "Integrating UX work with agile development through user stories: An action research study in a small software company," *J. Syst. Softw.*, vol. 170, p. 110785, Dec. 2020, doi: 10.1016/j.jss.2020.110785.
[43] D. Peters, K. Vold, D. Robinson, and R. A. Calvo, "Responsible AI—Two Frameworks for Ethical Design Practice," *IEEE Trans. Technol. Soc.*, vol. 1, no. 1, pp. 34–47, Mar. 2020, doi: 10.1109/TTS.2020.2974991.
[44] E. Kazim and A. S. Koshiyama, "A high-level overview of AI ethics," *Patterns*, vol. 2, no. 9, p. 100314, Sep. 2021, doi: 10.1016/j.patter.2021.100314.
[45] S. Oduro, E. Moss, and J. Metcalf, "Obligations to assess: Recent trends in AI accountability regulations," *Patterns*, vol. 3, no. 11, p. 100608, Nov. 2022, doi: 10.1016/j.patter.2022.100608.
[46] S. Delecraz, L. Eltarr, M. Becuwe, H. Bouxin, N. Boutin, and O. Oullier, "Responsible Artificial Intelligence in Human Resources Technology: An innovative inclusive and fair by design matching algorithm for job recruitment purposes," *J. Responsible Technol.*, vol. 11, p. 100041, Oct. 2022, doi: 10.1016/j.jrt.2022.100041.
[47] N. Balasubramaniam, M. Kauppinen, A. Rannisto, K. Hiekkanen, and S. Kujala, "Transparency and explainability of AI systems: From ethical guidelines to requirements," *Inf. Softw. Technol.*, vol. 159, p. 107197, Jul. 2023, doi: 10.1016/j.infsof.2023.107197.
[48] V. Vakkuri, K.-K. Kemell, and P. Abrahamsson, "ECCOLA - a Method for Implementing Ethically Aligned AI Systems," in *2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, Aug. 2020, pp. 195–204. doi: 10.1109/SEAA51224.2020.00043.
[49] G. J. Costello, K. Conboy, and B. Donnellan, "Reflections on 'reflection' in Action Research".
[50] M. C. Paulk, B. Curtis, M. B. Chrissis, and C. V. Weber, "Capability maturity model, version 1.1," *IEEE Softw.*, vol. 10, no. 4, pp. 18–27, Jul. 1993, doi: 10.1109/52.219617.
[51] S. Amershi *et al.*, "Software Engineering for Machine Learning: A Case Study," in *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, Montreal, QC, Canada: IEEE, May 2019, pp. 291–300. doi: 10.1109/ICSE-SEIP.2019.00042.
[52] O. Cico, L. Jaccheri, A. Nguyen-Duc, and H. Zhang, "Exploring the intersection between software industry and Software Engineering education - A systematic mapping of Software Engineering Trends," *J. Syst. Softw.*, vol. 172, p. 110736, Feb. 2021, doi: 10.1016/j.jss.2020.110736.
[53] P. Bourque, R. E. Fairley, and IEEE Computer Society, *Guide to the software engineering body of knowledge*. 2014.
[54] IEEE, "Software Engineering Competency Model," *IEEE*, p. 168, 2014.

[55] M. Zajko, “Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates,” *Sociol. Compass*, vol. 16, no. 3, Mar. 2022, doi: 10.1111/soc4.12962.
[56] B. Richardson and J. E. Gilbert, “A Framework for Fairness: A Systematic Review of Existing Fair AI Solutions.” arXiv, Dec. 10, 2021. Accessed: Nov. 13, 2022. [Online]. Available: http://arxiv.org/abs/2112.05700
[57] M. S. A. Lee and J. Singh, “The Landscape and Gaps in Open Source Fairness Toolkits,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, May 2021, pp. 1–13. doi: 10.1145/3411764.3445261.
[58] “ISO/IEC/IEEE International Standard - Systems and software engineering -- Life cycle processes -- Requirements engineering,” IEEE. doi: 10.1109/IEEESTD.2018.8559686.
[59] V. Braun and V. Clarke, “Using thematic analysis in psychology,” *Qual. Res. Psychol.*, vol. 3, no. 2, pp. 77–101, Jan. 2006, doi: 10.1191/1478088706qp063oa.
[60] S. Kassir, L. Baker, J. Dolphin, and F. Polli, “AI for hiring in context: a perspective on overcoming the unique challenges of employment research to mitigate disparate impact,” *AI Ethics*, vol. 3, no. 3, pp. 845–868, Aug. 2023, doi: 10.1007/s43681-022-00208-x.
[61] S. D. Vishnubhotla, E. Mendes, and L. Lundberg, “Understanding the perceived relevance of capability measures: A survey of Agile Software Development practitioners,” *J. Syst. Softw.*, vol. 180, p. 111013, Oct. 2021, doi: 10.1016/j.jss.2021.111013.
[62] F. A. Bukhsh, Z. A. Bukhsh, and M. Daneva, “A systematic literature review on requirement prioritization techniques and their empirical evaluation,” *Comput. Stand. Interfaces*, vol. 69, p. 103389, Mar. 2020, doi: 10.1016/j.csi.2019.103389.
[63] E. Drage and K. Mackereth, “Does AI Debias Recruitment? Race, Gender, and AI’s ‘Eradication of Difference,’” *Philos. Technol.*, vol. 35, no. 4, p. 89, Dec. 2022, doi: 10.1007/s13347-022-00543-1.
[64] P. B. De Laat, “Companies Committed to Responsible AI: From Principles towards Implementation and Regulation?,” *Philos. Technol.*, vol. 34, no. 4, pp. 1135–1193, Dec. 2021, doi: 10.1007/s13347-021-00474-3.

APPENDICES

APPENDIX A

Questionnaire used for eliciting feedback on RAI tools and processes introduced

Section A. The Action (s)
A1: What was the main Action (s) taken during the sprint cycle?
INTENDED:
UNINTENDED:
Any other reflections that you think appropriate to include here?
Section B. The Outcomes (results)
B1: Did I get the outcomes I wanted? (or more realistically, what were the outcomes I got and how well do these accord with those I sought?)
INTENDED:
UNINTENDED:
B2: To the extent that I got them, do I still want them (i.e. are they any use)? Why or why not?
B3: To the extent that I didn't get them, why not?
B4: Did I succeed in carrying out the planned actions? If not, what prevented or discouraged me?
Any other reflections that you think appropriate to include here?
Section C. Learning
C1: When faced with a similar situation in future: Would I try to pursue different outcomes based on what I have learned in this sprint cycle?
C2: Would I try different actions to pursue similar outcomes?
C3: What have I learned about myself, my skills, my attitudes etc.?
C4: What are the changes to my perceptions and knowledge about the topic of Responsible AI?
Any other reflections that you think appropriate to include here?
Section D. Future Work
D1: What are the implications for further actions in this situation? e.g. Do I have any practical suggestions for further changes or follow-on projects in the organization? Do I have any suggestions for changes or for new structures or systems?
D2: What is my feedback (positive and negative) for the research process?
D3: Do I see value in similar research collaboration in the future and do I have any suggestions?
Any other reflections that you think appropriate to include here?

APPENDIX B

AI-MM for Trustworthy AI Software Development by Cho et al

Process Area (5)	Process (13)	Base Practice (53)
Plan and Design (P&D)	Software Requirements Analysis (SRA)	[Common] Define development goals, Define function and performance, Function/performance evaluation metric, Functional and behavioral evaluation metric (4) [Fairness] Fairness risk factor analysis, Fairness evaluation metric, Stakeholder review—fairness (3) [Safety] Safety risk factor analysis, Safety evaluation metric, Safety countermeasures, Stakeholder review—safety (4)
	Software Architecture Design (SAD)	[Common] Model design, User-conscious design, Design considering maintenance (3) [Fairness] Design considering fairness maintenance (1) [Safety] Design considering safety maintenance (1)
Data Collection and Processing (Data C&P)	Data Collection (DCO)	[Common] Data information specification, Data acquisition plan, Securing data for verification (3) [Fairness] Data bias assessment (1)
	Data Cleaning (DCL)	[Common] Data representativeness review, Data error check (2) [Fairness] Data bias periodic check (1)
	Data Preprocessing (DPR)	[Common] Data preprocessing result record, Data selection criteria record (2) [Fairness] Data bias and sensitive information preprocessing, Data distribution verification (2) [Safety] Data attack check (1)
Model Development and Evaluation (Model D&E)	Training Process Management (TPM)	[Common] Training preparation record, Training history record (2) [Fairness] Bias review during model learning (1) [Safety] Open-source safety check (1)
	Performance Evaluation of AI model (PEA)	[Common] Performance evaluation record, Evaluation variation check, Model change impact analysis (3) [Fairness] Fairness maintenance check (1)
	Safety Evaluation of AI model (SEA)	[Safety] Safety check of the model source code, Model safety check from external attacks, Provide model reliability (3)
	Final AI model Management (FAM)	[Common] Model detailed specifications record, Detailed design of the model record (2) [Fairness] Model's fairness evaluation result record (1)
System Evaluation (SE)	System Safety Evaluation (SSE)	[Safety] Safety evaluation according to system configuration, System safety evaluation from external attacks (2)
	System Safety Preparedness (SSP)	[Safety] User error notice in advance, Provide exception handling policy, Consider human intervention (3)
Operation and Monitoring (O&M)	AI Infrastructure (AIN)	[Common] Infrastructure construction plan, Infrastructure utilization record (2) [Fairness] Fairness management infrastructure (1) [Safety] Safety management infrastructure (1)
	AI model Operation Management (AOM)	[Common] Define quality control criteria after model deployment (1) [Fairness] Monitoring model quality (1)

Figure 14. AI-MM for Trustworthy AI Software Development by Cho et al.

APPENDIX C

Survey to examine adoption barriers for Responsible AI

Responsible Artificial Intelligence (AI)

Responsible AI/Ethical AI is artificial intelligence that adheres to well-defined ethical guidelines regarding fundamental values, including such things as individual rights, privacy, non-discrimination, and non-manipulation.

Artificial intelligence is the name given to any computer system taught to mimic intelligent human behaviors. AI covers all aspects of machine learning and systems that adapt themselves.

There are 13 questions in this survey.

Demographic Information

1. What best describes your role in the organization? *
Please choose all that apply and provide a comment:

- Technical
- Administrative
- Leadership
- Other

2. What stage best describes the level your team is currently at with respect to adopting Ethical AI?
* Please choose all that apply and provide a comment:

- Innovators
- Early Adopter
- Early Majority
- Late Majority
- Laggards

If you don't work on a team, you can use the next biggest organizational structure such as branch, division, ministry or organization.

User Acceptance Factors (Individual level characteristics)

3. Please select the answer that best reflects your response to the following statement:
"An ethical AI application would be useful in my current job."
* Please select at most one answer

- Disagree
- Somewhat disagree
- Neither agree, nor disagree
- Somewhat agree
- Agree

4. Please select the response that best reflects your response to the following statement:
"An ethical AI application would make my job easier."

***** Please select at most one answer

- Disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Agree

Leadership characteristics

5. How would you describe the general attitude of leadership in your team towards using Ethical Artificial Intelligence in the workplace?
* Please choose all that apply and provide a comment:

- Strong concerns
- Some concerns
- No concerns
- Somewhat open to it
- Strongly encouraging its use

If you don't currently belong to a team, you can use branch, division, ministry or organization to reflect your answer.

6. How would you describe your team's leadership with respect to newer, innovative technologies?
* Please select at most one answer

- Wants to be the first to try new tech
- Embraces new tech
- Waits to see if works first, before adopting it
- Adopts new tech after it's established and popular
- Generally dislikes new tech

Environmental characteristics

7. How would you describe the level of government support that currently exists for adopting Ethical Artificial Intelligence?
* Please choose all that apply and provide a comment:

- No support
- A little bit of support
- Not sure
- Moderate support
- Strong support, adoption is actively encouraged

8. How would you characterize the influence of external pressure, such as media, public perception or public expectation for the provincial government to adopt Ethical Artificial Intelligence?
* Please select at most one answer

- The expectation is that government will not adopt it
- The expectation is that government will adopt it, but will also govern its use
- Don't know
- The expectation is that government will adopt it
- The expectation is that government will adopt it and actively encourage its growth

IT expertise

9. What do you think about the current level of IT expertise in your organization as it relates to supporting and using Ethical Artificial Intelligence?
* Please select at most one answer

- No expertise
- Some expertise
- Don't know
- A lot of expertise
- Too much expertise

10. In terms of how many human resources are allocated to technical roles (IT); how would you describe the size of that group within your team?
* Please select at most one answer

- very small (1 or 2 people)
- small (between 2 and 10)
- medium (between 10 and 30)
- large (between 30 and 50)
- very large (over 50 people)

Relative advantage

11. How would you characterize any advantages that Ethical AI technology could bring to your specific area of work?
Comment only when you choose an answer.
* Please select at most one answer

- It would be detrimental to my work
- There would be little advantage to my work
- It wouldn't change my work at all
- There would be some advantage to my work
- There would be great advantage to my work

12. How compatible would an Ethical AI implementation with your existing line of business?
Comment only when you choose an answer.
* Please select at most one answer

- Incompatible
- Somewhat incompatible
- Not sure
- Somewhat compatible
- Very compatible

Other comments

13. Please feel free to use this space to add any additional comments with respect to barriers, concerns or benefits related to your organization's adoption of Ethical AI.
* Please write your answer here:

APPENDIX D

User Flow Diagram comparison

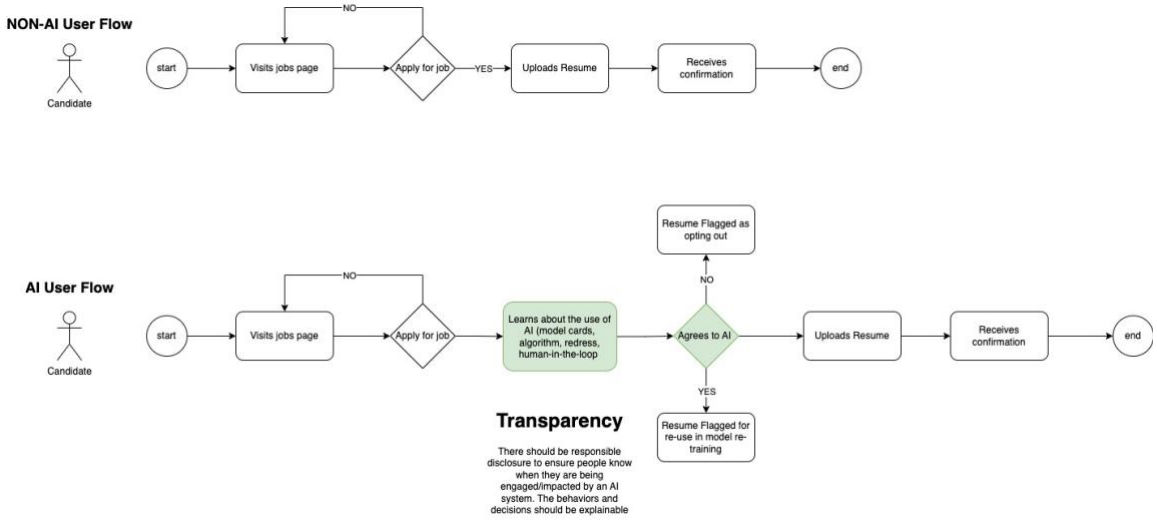

Figure 15. User Flow Diagram comparison

APPENDIX E

Action Research Workshop

Cycle 1 - Problem Exploration

Building ethically aligned AI systems with ethical user stories; a socio-technical approach to requirements elicitation

This workshop will be an introduction to the landscape of ethical considerations for an AI-enabled recruitment system. The exploration of the problem will be framed as working towards the goals of:

1. transparency
2. accountability
3. fairness
4. data and individual agency
5. human oversight
6. well-being
7. safety and security

Exploration Problem: **How can high-level ethical principles translate into practical features?** In the context of gathering requirements for an automatic redaction and job matching AI system, how do we minimize the negative impacts of the AI system?

Cycle 1 - Action (Options)

Requirements Elicitation

- ECCOLA (ethical user stories)
- interviews
- scenarios
- requirement workshops
- interactive demos/prototypes

Requirements Analysis

- requirements classification (group ethical principles as non-functional and functional requirements)

Requirements Specification

- make ethical requirements quantifiable/measurable
- clearly define scope of responsibility
- determine ethical requirements verification/validation techniques
- determine data requirements

- provide examples of desired outcomes
- negative requirements

Requirements Verification/Validation

- monitoring and analysis to continuously perform requirements validation
- awareness of potential mismatches between training data and real-world data (detection of model drift)

APPENDIX F

Taxonomy of currently existing forms of bias (Richardson and Gilbert)

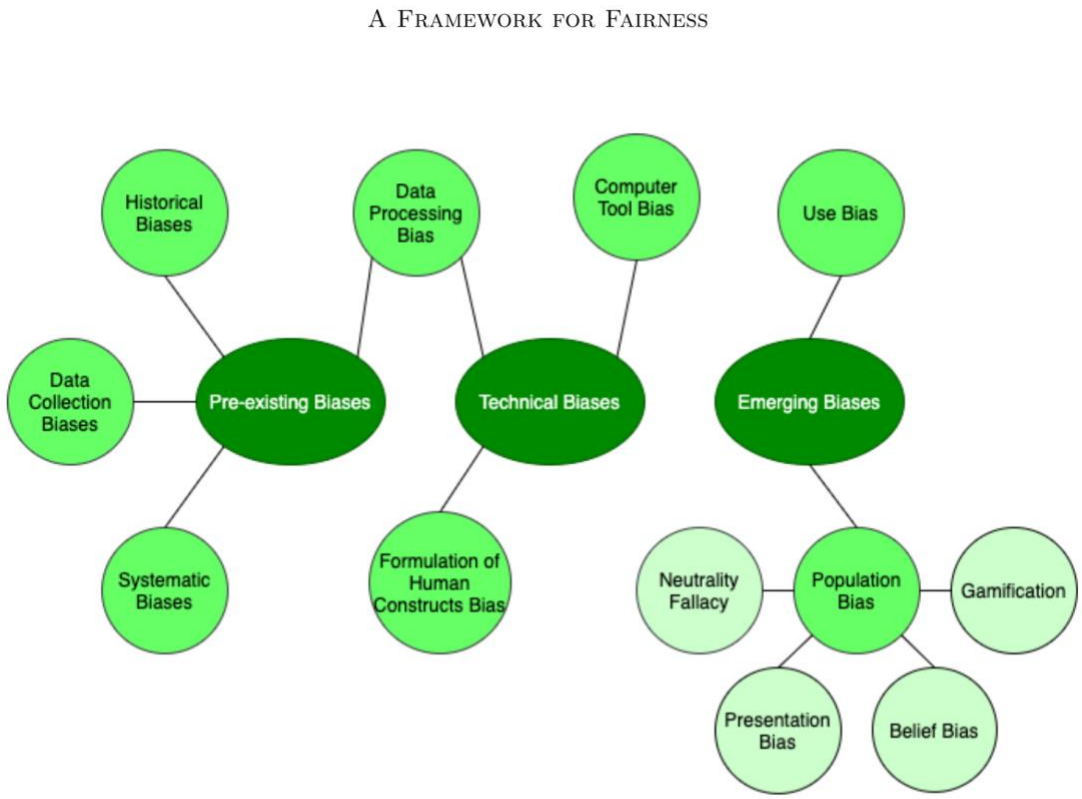

Figure 1: A taxonomy of the currently existing forms of bias.

Figure 16. Taxonomy of currently existing forms of bias (Richardson and Gilbert)

APPENDIX G

System Architecture Design for Automated Redaction

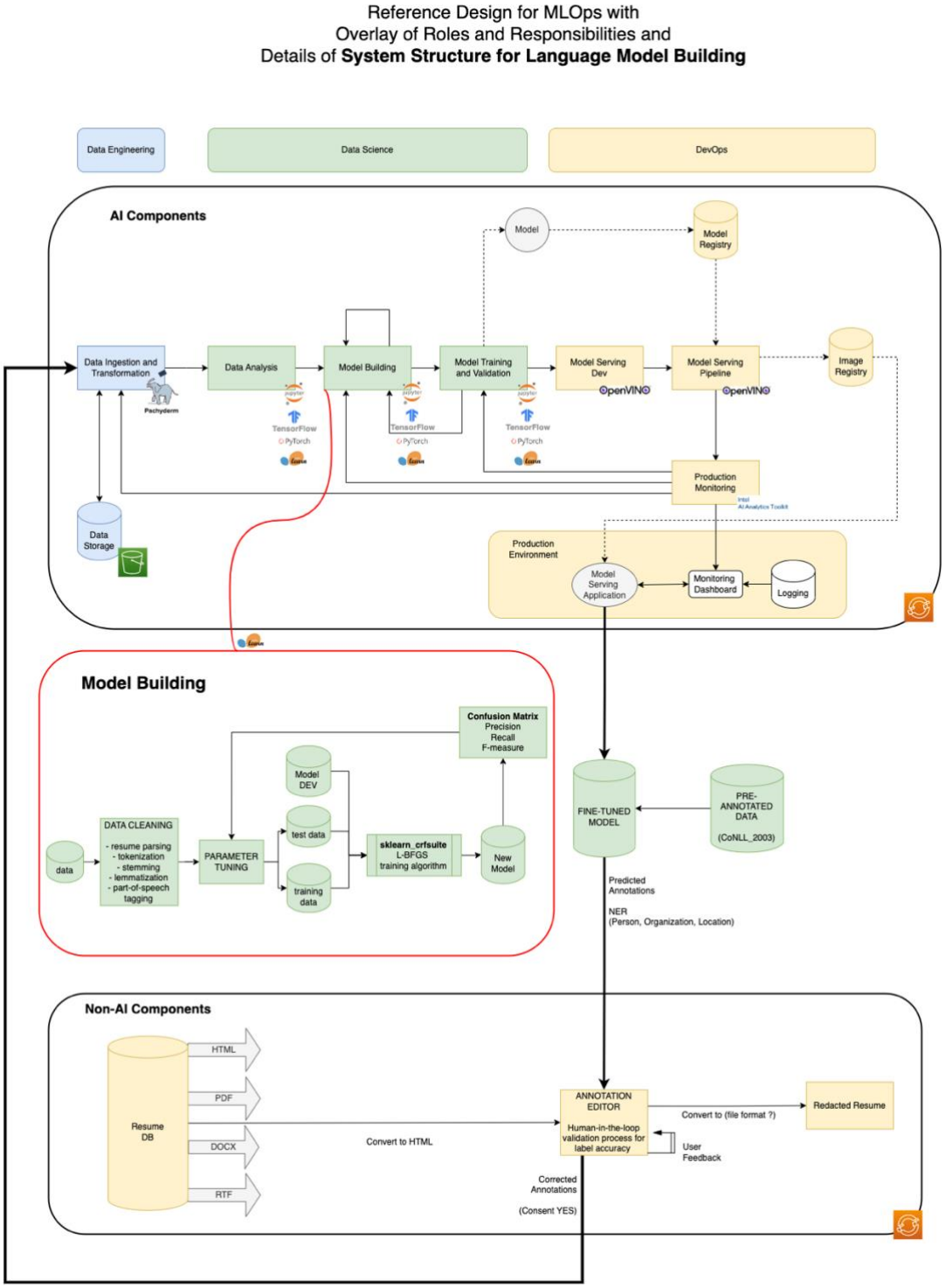

Figure 17. System Architecture for integrating AI and Non-AI components in RedHat OpenShift