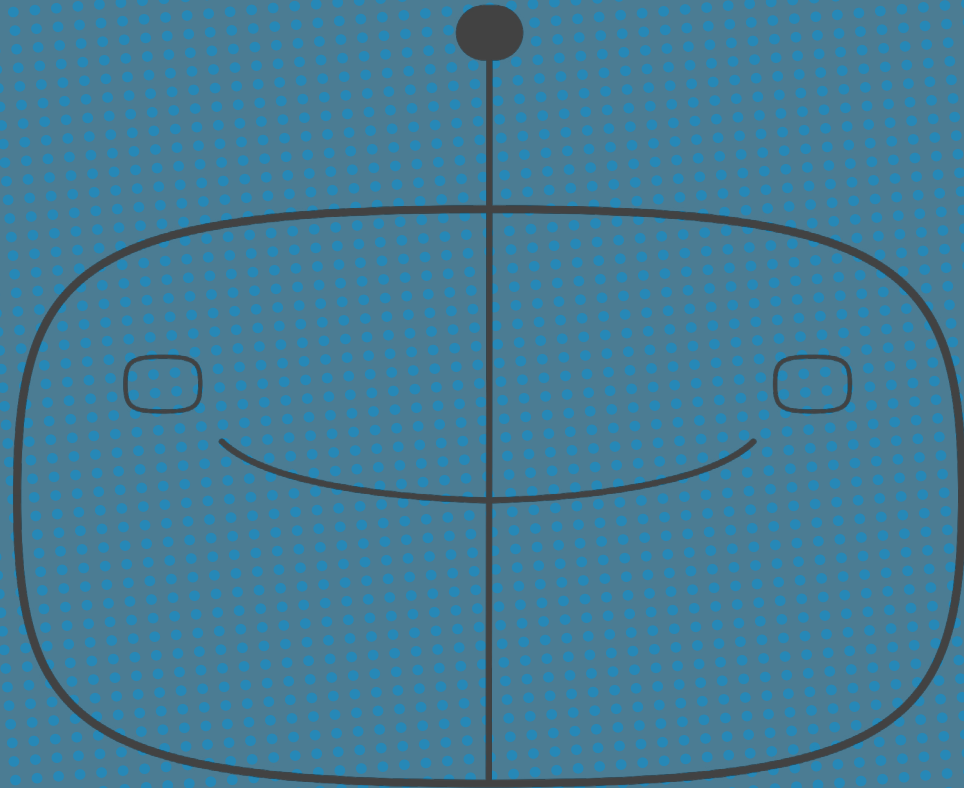


JULY 2017



Ethics Analysis of Predictive Algorithms

An Assessment Report for BC Safety Authority

generation R

Final Report

Delivered on
July 31, 2017

Prepared for
BC Safety Authority
505 - 6th Street, Suite 200
New Westminister BC Canada
V3L 0E1

Prepared by
Generation R Consulting
6163 University Boulevard
Vancouver BC Canada
V6T 1Z1

EXECUTIVE SUMMARY

The aim of the work presented in this report is to help BC Safety Authority (BCSA) take informed and proactive measures to innovate without compromising the organization's values and the value BCSA adds to the communities it serves. Well-designed predictive algorithms can deliver myriad of benefits to organizations across the world. However, recent advances in machine learning have led to discoveries of ethical challenges associated with predictive models. Evidence suggests that it is much harder to revert negative effects of predictive models that are already deployed in a community than to prevent undesirable effects during the design and development of the technology. Generation R Consulting Inc. helps businesses take a proactive approach to addressing these challenges early in the design and deployment of a technology, such that organizations can mitigate and manage possible undesirable effects.

BCSA is currently in the process of transforming a customized software-based decision support system developed in-house, called the Resource Allocation Program (RAP). The computation of an output for the older, existing RAP (RAP 1.0) can be described as linear with a fixed set of factors and parameters, and its scientific validity and performance has been a source of frustration for BCSA employees who use the system daily. In contrast, the new RAP (RAP 2.0) seeks to take advantage of BCSA's data assets, using data science and machine learning techniques, to improve the ways in which BCSA serves public safety. As one of the programs integral to BCSA's operation, RAP 2.0 is being designed to assess a number of technology assets inspected under BCSA's jurisdiction, and assign each one of them a probability value that represents the likelihood of finding a high hazard upon inspecting the asset. As such, the RAP probability output is a prediction that can help BCSA allocate its inspection resources more efficiently, thus helping to ensure public safety more effectively.

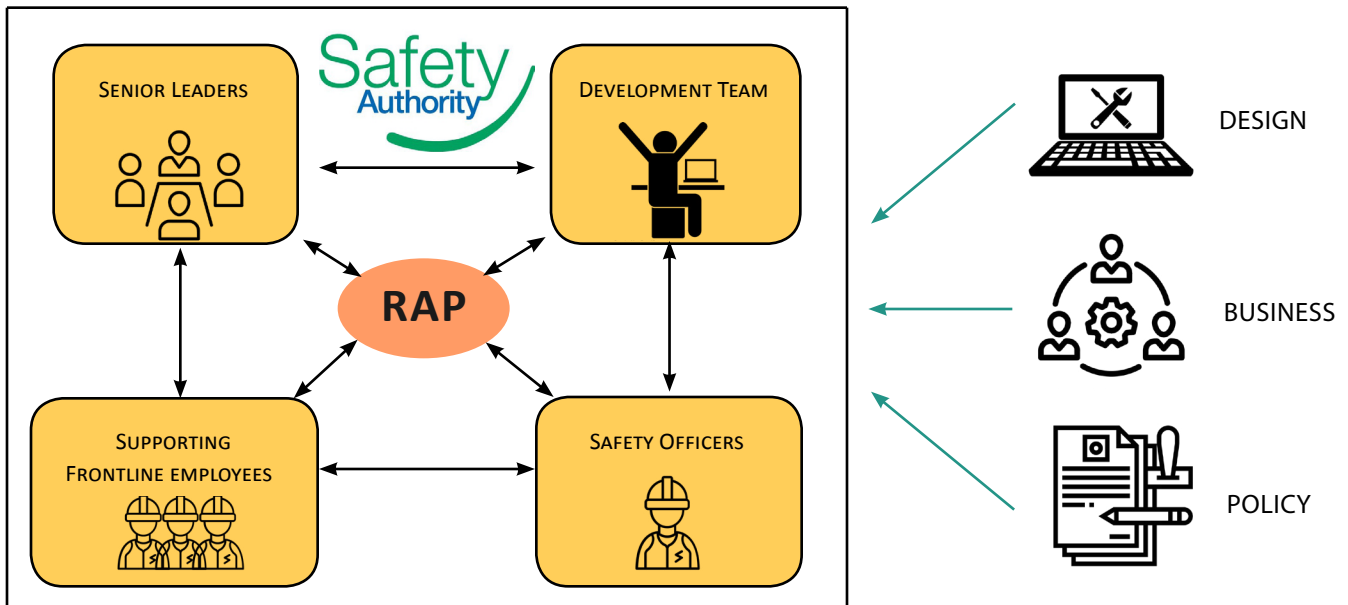
Systems such as RAP is crucial to organizations that use a risk-based approach to oversee public safety. While some jurisdictions operate on a 100% inspection model (i.e., every asset under their jurisdiction is inspected on a regular basis), BCSA operates on a risk-based model. A risk-based model relies on knowledge-based insights about risks to oversee public safety. In a risk-based model, the task of ensuring safety shifts from assuming every asset to be of equal hazard to prioritizing assets that, for one reason or another, tend to pose a higher risk, thereby requiring more attention from safety officers.

All technologies are designed, built and used in social contexts. Understanding what network of values a technology such as RAP 2.0 implicates helps the stakeholders of the technology identify and anticipate potential sources of conflict, and act to mitigate those conflicts. Therefore, Generation R identified key stakeholder relationships within BCSA and interviewed 21 stakeholders at BCSA using questions designed to elicit professional reflections centered around different value categories. Those we interviewed include: safety officers, senior safety officers, members of Information Technology team, members of the Data Analytics and Decision Science team, regional business leaders, safety managers, and other senior leaders within BCSA.

We identified how RAP 2.0 could trigger value conflicts in those relationships and analyzed them as challenges to be considered and addressed. Managing the impacts of value conflicts is important, especially since value conflicts can directly and indirectly result in the undesirable use of RAP 2.0 or rejection of RAP 2.0 by its users. Therefore, combined with the potential benefits RAP 2.0 can deliver to BCSA and its clients, addressing the value conflicts throughout its development and deployment process can help maximize the direct positive impact RAP 2.0 can have on public safety.

The result presented in this report provides foresight into fundamental challenges related to the design and deployment of RAP 2.0. For each challenge identified, we describe and outline recommendations to mitigate or address foreseeable value conflicts upstream in the development and use lifecycle.

The following list provides a high-level summary of key recommendations stemming from Generation R's detailed ethics analysis of RAP 2.0:



1. Clear Objective Setting:

Clearly define the objective(s) that RAP 2.0 will help BCSA accomplish. Currently, RAP 1.0 attempts to serve multiple purposes. Those objectives could be developed via a multi-stakeholder consultation and should be made transparent throughout the organization. This helps to develop internal guidelines, via a multi-stakeholder consultation, that clearly outline limits on what kinds of predictions or conclusions RAP 2.0 output can, and cannot, accurately support.

2. Transparency in Design:

Engage stakeholders with frontline expertise (e.g., safety officers) in the RAP 2.0 design process not only to test the functionality and increase the usability of RAP 2.0, but also to practice inclusive design processes that supports transparency and user autonomy. Noting the inherent opaqueness of data-driven systems, such inclusive design practices can also help identify and provide the type of information that can best address safety officers' need for transparency about RAP 2.0. This can also be accomplished by complementing the current expertise at BCSA with additional system design and user interface/experience design expertise.

3. Decisions about Machine Autonomy:

Acknowledge the fact that the deployment of RAP 2.0 adds more machine autonomy into BCSA's operation. BCSA will need to make an explicit choice about the level of machine autonomy BCSA desires to incorporate in its organization, and the risks associated with it. This includes making of policy decisions about how different kinds of erroneous outputs, inherent to all predictive algorithms, should be handled based on the different risks they pose on BCSA.

4. Monitoring Practices:

Monitor the effectiveness of RAP 2.0. This includes implementation of metrics that help gauge transparency of the system and how much safety officers trust the system in their daily use. The system should also be monitored to detect possible exacerbation of discriminatory practices.

5. Communication Practices:

Actively communicate, upon multi-stakeholder consultations, limits on how RAP 2.0 output can be interpreted and what kinds of conclusions it can/cannot support. Safety officers will also need to understand the quality of the data they collect and enter impacts the predictive performance of RAP 2.0.

Table of Contents

Executive Summary	1
1. Introduction	5
1.1. Motivating Story: A Glimpse of the Future	6
1.2 What is an Ethics Analysis?	8
1.3 Why Perform an Ethics Analysis?	9
2. Main Ethical Challenges and Current Approaches	10
2.1 Transparency and Interpretability of Predictions	11
2.2 Discrimination and Fairness	11
2.3 Public Perception and Awareness	12
2.4 Data, Privacy and Individual Autonomy/Consent	12
2.5 Responsibility and Accountability	13
2.6 Impact on Jobs or Expertise	13
3. Generation R's Methodology	14
4. Findings and Associated Items of Consideration	16
4.0.1 A Historical Context: RAP 1.0 as the Predecessor of RAP 2.0	17
4.0.2 Understanding RAP 2.0	18
4.1 Use Case 1: RAP 2.0 as a Decision Support Tool for Safety Officers	21
4.1.1 Transparency and Interpretability	22
4.1.2 Professional Autonomy, Oversight and Impact on Jobs	27
4.1.3 Discrimination and Fairness	31
4.1.4 Responsibility and Accountability	36
4.1.5 Public Perception/Awareness of BCSA and RAP 2.0	41
4.2 Use Case 2: Strategic Decision Making	43
4.2.1 Interpretability, Transparency and Trust	44
4.2.2 Autonomy and Jobs	47
4.3 Use Case 3: Public Reporting of the RAP 2.0 Output	49

1. INTRODUCTION

BC Safety Authority (BCSA) is the leading authority in British Columbia (BC) that offers licensing, certification and assessment services of key technological systems that affect safety of the inhabitants of BC. BCSA has been using a software-based system, called the Resource Allocation Program (RAP), that helps BCSA safety officers prioritize their daily inspection activities. BCSA has recently started exploring the use of predictive, data-driven algorithms to bring significant improvements to RAP and learn from patterns that BCSA's records of inspection has to offer as a means to better maintain public confidence of safety in BC. The newly established Data Analytics and Decision Science department has been given the responsibility to spearhead this endeavour of developing a new and data-driven version of RAP.

From predicting duration of travel during traffic hours to projecting weather patterns for an upcoming weekend, many Canadians today are already familiar and sometimes dependent on applications of predictive algorithms. However, there are social and ethical risks associated to integrating these data-driven algorithms into an organization and these risks are inherently different from integration of other technological artifacts. Generation R Consulting Inc. (Generation R) specializes in performing ethics analyses on intelligent technologies, such as robotics and artificial intelligence. Generation R has been employed to deliver BCSA with a foresight of potential social and ethical issues of designing and deploying the next generation of RAP (RAP 2.0), such that BCSA can take proactive approach to innovating its operations without compromising the organization's values.

This report provides a summary of the main ethical challenges being discussed by users and developers of predictive algorithms around the world today, as well as a presentation of the results of our ethics assessment on three different use cases of RAP 2.0 BCSA has identified for this project. Presented as part of the results of our analysis are proactive measures and recommendations that can help support BCSA's ethical use and deployment of data-driven algorithms in its operation.

The remainder of Section 1 provides a glimpse of what an everyday work for a safety officer could look like in 5 to 10 years from now with RAP 2.0 as a motivating story. This is followed by the rationale behind performing ethics analyses. Section 2 provides an overview of the main ethical challenges around algorithmic decision making. It also provides the latest proposals of principles, strategies and best practices for navigating these key issues. Section 3 outlines Generation R's three-part methodology used to conduct the ethics analysis of RAP. Section 4 details the findings from our analysis on the three use cases. Our results are presented in the order of use cases and key values pertinent to the use case. For each use case, we present potential issues and associated recommendations for the future development of RAP.

1.1 MOTIVATING STORY: A GLIMPSE OF THE FUTURE

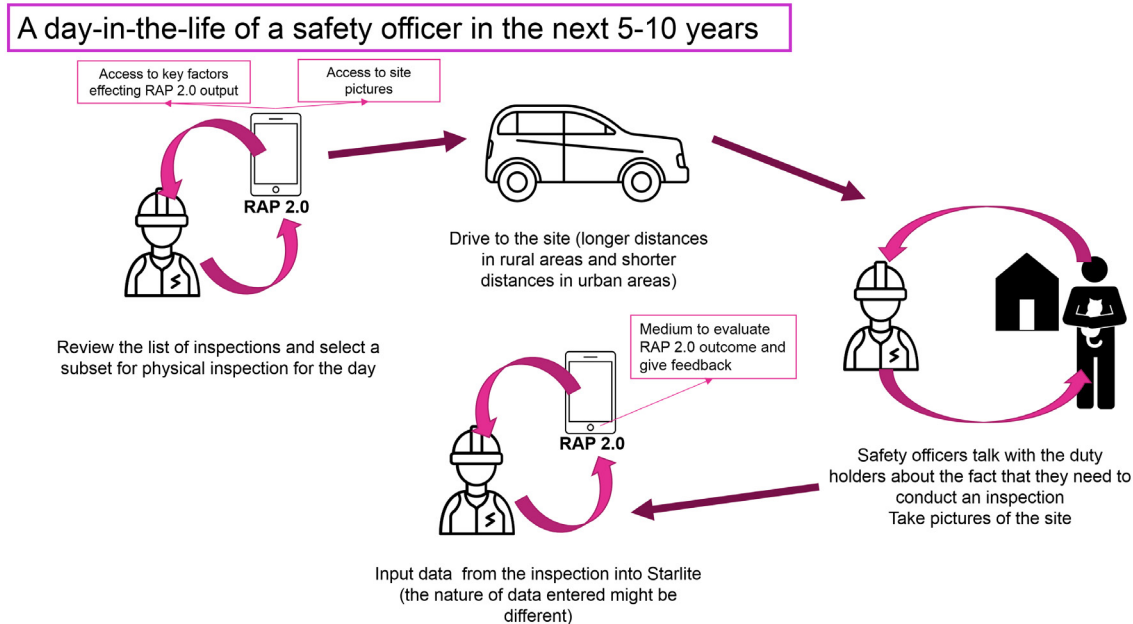


Figure 1.1 This figure depicts a day in the life of a safety officer envisioned by Generation R based on our understanding of RAP 2.0 and the stakeholders.

This section provides a short fictional day in the life of a safety officer in the next 5 to 10 years. What would a safety officer's day look like and where would RAP 2.0 fit in? Rather than providing a catastrophic scenario, we provide the following description as a motivating story for the analysis presented in this report, and to present a positive outlook of RAP 2.0 and its impacts on BCSA.

Jill is a new safety officer at BCSA. She started working for BCSA six months ago, and has since been responsible for electrical inspections in the rapidly growing township of New Westminster. With the overpopulation of metro Vancouver, the city of New Westminster has been booming with new residential and office buildings. She sits down for breakfast and launches the Starlite program on her tablet. The first thing she sees is her personalized inspection dashboard.

The system has learned her usage patterns and has detected that she always starts her day by looking at the map of RAP 2.0 probabilities in the areas closest to her. They are displayed like a heatmap with the warmest to coldest colours indicating the highest to lowest probabilities. Now that the system has learned what she does, it's the first thing she sees on the dashboard when she starts Starlite, and the clusters of the warmest colours are outlined for her.

The heatmap is the feature of Starlite Jill likes the most. When she was being trained to use the interface by Frank, a senior safety officer, he was proud to explain to her how he asked for that feature during one of the regular meetings with the development team back in 2018 when he was a safety officer, and that, as a result, it was implemented later that year. Frank now works with the safety intelligence team, which looks at patterns of noncompliances found in the area. Safety officers who perform everyday inspections, like Jill, don't have access to the noncompliance patterns the safety intelligence team finds. But the team

members, like Frank, come up with new educational and training programs to provide targeted material to contractors and field safety representatives.

RAP 2.0 seems to think there will be quite a few technical failures near downtown. Ever since BC Hydro integrated their data with BCSA, Jill noticed that the RAP 2.0 predictions of technical failures have improved. It could be the unusually high temperature predicted for the day, making everyone turn on their air conditioning, combined with the fact that the main electrical system in the area was installed years before she immigrated to Canada. She doesn't know exactly what combination of these things was picked up by RAP 2.0, and its sister programs, to compute today's probabilities. She could find out by opening up BCSA's internal webpage, where the developers provide the latest information on RAP 2.0 and its underlying models, but she would rather get the updates next week at the safety officers' meeting. Meanwhile, her gut feeling says she should drop by the downtown area first thing today.

Jill knows the downtown area well. There's a senior care centre that opened a couple of months ago in the area of what used to be an old inn, and she has been wanting to make sure that there's no serious safety hazard there. RAP 2.0 probability is at 55% for the care centre, which is quite a bit higher than she expected. She taps on the care centre to see the full list of permits from the facility, and notices a photo of a new boiler the facility manager seems to have submitted to BCSA yesterday. She notices on the side of the photo what seems to be uncovered electrical wiring. "Hmm.. That could be a problematic," she says to herself.

Before she heads out, she clicks on a button that says Policy-Priority and patches of new dots appear where BCSA prioritizes inspections based on policy decisions. They are mostly near Aldergrove, quite a drive away from New Westminster, and it looks like a number of homeowners finished work on a handful of new houses that need inspection today. BCSA has a standing policy to inspect all the electrical work done by homeowners themselves, regardless of the RAP 2.0 probability. She won't have time to both drop by the care centre and inspect all of the Aldergrove area today. Well within the limits of her professional autonomy, Jill decides to make a quick stop at the senior care centre anyway, and plans to inspect as many of the new homes as she can afterward.

At the care centre, she asks the facility manager to show her the new boiler. The facility manager is excited that a safety officer dropped by so quickly after he submitted the photo. He had heard about the new photo feature BCSA added to the client portal and how it uses machine learning to approve things faster. The old boiler apparently failed late last week, and the tenants have been aching to get a new one installed. He is quickly disappointed to find that she is not there to approve the installation of the new high-tech boiler, but to check on the electrical system.

In the boiler room, she quickly realizes that what she saw on the photo was not electrical wiring but a poster on a wall with a picture of open wiring. She opens the Starlite system, clicks on the feedback button next to the RAP 2.0 probability, and puts a quick note explaining what she found. The RAP 2.0 development team will get the note and see if the photo was what raised the RAP 2.0 probability.

She asks the facility manager if she can take additional photos of the boiler for her colleagues to approve, and he agrees. She remembered that her colleague from the boiler team really cared about good photos of the connectors. She takes a few photos and adds additional details about the boiler situation at the facility. She recalls from her training that the more photos and descriptions that are available for a permit, the more likely RAP 2.0's sister programs are to kick in and better assess if a physical inspection is needed. Just as she heads out the care centre, she sees the facility manager check his phone and say "Oh, it has just been approved! That was so fast! I need to turn it on right now."

Jill smiles and drives towards Aldergrove.

1.2 WHAT IS AN ETHICS ANALYSIS?

An ethics analysis provides a systematic way of accounting for the ethical, or value-based, dimensions of a technology. It focuses on exposing underlying values, and value conflicts, that are embedded in a technology or the various processes and policies surrounding it. Identifying the ethical dimensions of a technology allows designers, regulators or policymakers to anticipate stakeholder reactions to a new technology. An ethics analysis also allows decision makers to shape those reactions, by designing specific values, such as privacy protections or specific power relations, into the technology or the many policies surrounding it. These values-based decisions, stemming from a thorough ethics analysis, can ultimately help an organization smooth the rollout of a new disruptive technology, engender trust in the technology, and even gain a competitive edge through enhanced design practices.

1.3 WHY PERFORM AN ETHICS ANALYSIS?

Every technology is designed to address a series of needs. Design and implementation of any new technology impacts its direct stakeholders (e.g., primary users) and indirect stakeholders (e.g., managers overseeing the users). Each use case of a technology carries the potential to disrupt those stakeholders' values, expectations, and established ways of accomplishing certain familiar tasks or work. For example, a technology might undermine certain stakeholders' privacy expectations in the workplace, or shift power relations between stakeholder groups in order to make larger efficiency gains. Thus, technology can elicit a range of ethical (value-based) reactions among different stakeholders, and each of those reactions can influence the way a technology is used, ultimately impacting the success of that technology

At the same time, maintaining the status quo does not necessarily avoid these ethical challenges. For example, continuing to use old technologies that undermine or no longer serve the rights of the employees (e.g., privacy) or values of the organization (e.g., secure storage of sensitive client data) can be sources of value conflicts within the organization and between stakeholder groups.

Seen this way, the ethical dimensions of a technology can pose challenges to an organization. But they also open new opportunities for designers and engineers. Exposing the ethical dimensions of technology through an ethics analysis can help designers and engineers manage shifts in liability, unexpected stakeholder perceptions or reactions to the technology, or even changes in public perceptions of, or trust in, an organization.

An ethics analysis of technology, therefore, can help an organization make the decision to embrace or reject new technologies and, when given the decision to embrace a new technology, help take a proactive approach to navigate the landscape of ethical issues in moving forward with the decision.

2. MAIN ETHICAL CHALLENGES AND CURRENT APPROACHES

In this report, the term “predictive algorithm” refers to a computational algorithm that produces a prediction score from an explicit set of rules and/or a set of data. Data-driven approaches, such as machine learning techniques, can be used to compute the probability of a future event (e.g. a safety officer finding equipment that poses a high safety hazard) based on observed past events.

There is no doubt that predictive algorithms can assist human decision making and deliver many benefits to an organization. Our analysis is not meant to undermine these benefits. At the same time, predictive algorithms raise a number of ethical challenges that have been documented in the academic literature, and that designers, policy makers and regulators need to be aware of, in order to ensure that new predictive systems are deployed successfully and well managed. Here, we outline a few of the main challenges discussed in the literature.

2.1 TRANSPARENCY AND INTERPRETABILITY OF PREDICTIONS

One of the main concerns actively voiced in Artificial Intelligence (AI) communities is the need for transparency. It is possible for machine learning systems to suffer a lack of transparency in the way that predictions are associated with the underlying data space. Developers tend to have a good understanding of the link between data and predictions. End-users, on the other hand, often do not constitute a core part of the system design activities and, therefore, might not have ready access to explanations. The lack of end-user knowledge and engagement can contribute to the perceived lack of transparency between the stakeholder groups. Ultimately, the issue of transparency can lead to a lack of trust when using the system.

In addition, there are the so-called “black-box” algorithms producing predictive outputs that are inherently difficult, or even impossible, for the developer to interpret or explain in satisfying terms. Depending on the application of the predictive algorithm, the decision to use these opaque techniques can directly conflict with the end-user’s need for an explanation. Being able to explain the inner workings of a system, and justify the validity of its output, are crucial aspects of being pre-emptively prepared for incident investigations that pertain to the use of the algorithm.

The idea of increasing transparency in predictive algorithms has recently been proposed as a guiding principle for designers. There are no established standards for how this principle should translate into practice, although active community efforts to define such standards are underway. That said, some ways to consider transparency in development and deployment of predictive algorithms include: the decision to implement a more explainable algorithm over a less explainable algorithm; implementing a healthy communication strategy about the algorithm with the end-users; and ensuring that the design of the algorithm has theoretical underpinnings within a specific application area.

2.2 DISCRIMINATION AND FAIRNESS

Machine learning systems learn from a given set of data. Training data sets that contain biases have been known to propagate those biases to predictions, which can lead to unintentional discrimination against an individual or a group of individuals. Blind interpretation of these biased predictions can cause systemic harms that are inherently different from more familiar types of harm, such as physical, property, privacy and psychological harm.

Sometimes biases can be easily identified. For example, if a group of contractors are falsely marked as having had safety violations because of a recurring technical error in a data entry system, then correcting the bias might be as simple as replacing the faulty system and making sure that the algorithm’s performance is subsequently monitored to catch similar errors. If, however, a biased output from an algorithm is the result of some systematic bias in a larger system feeding into that data (e.g. institutional bias), then the bias can propagate unchecked.

Issues of discrimination and fairness have been recognized within the community of predictive algorithm experts. Currently, there are some preliminary solutions proposed by researchers in this field. However, there is no consensus on what is the best approach for all applications of predictive algorithms. The overall recommendation is to define and understand discrimination and fairness within a specific application of predictive algorithms. This understanding could lead to appropriate design choices or policy changes.

2.3 PUBLIC PERCEPTION AND AWARENESS

Based on recent surveys, public awareness about machine learning is relatively low despite a relatively high awareness of the applications that employ it (e.g. speech recognition and recommender systems). Not surprisingly, the low familiarity of the technology translates into a relatively unfixed public opinion of machine learning. On the one hand, the public seem to recognize the benefits posed by machine learning, such as: machines being perceived as more objective, accurate, and efficient than humans; the economic opportunity machine learning represents; and the ability for machine learning to tackle large-scale social challenges like climate change.

Yet, the public expresses concern over such issues as: potential harms caused by autonomous systems (e.g. driverless cars); job losses; the “depersonalization” that accompanies human-machine interactions; and a general narrowing of choices open to an individual. These facts about the public seem to underscore the importance of good stakeholder engagement strategies when designing and/or deploying systems that use machine learning.

2.4 DATA, PRIVACY AND INDIVIDUAL AUTONOMY/CONSENT

Machine learning and “big data” go hand in hand. Large data sets, or connected streams of data sets, are often used to train, test and continually improve the quality of the predictive output. It is no surprise, therefore, that machine learning raises privacy concerns, including concerns over surveillance (in public, private, or in the workplace) and the ability to consent to having one’s data (or data generated by one’s activities) used as inputs to a machine learning system.

The existing practice of obtaining informed consent from those who provide data is often considered inadequate, especially since the same data set can be often repurposed or reanalyzed with consequences unknown or undisclosed at the time of obtaining consent. Experts in the big data community have proposed a diverse set of recommendations to address this challenge, however the community has yet to reach an agreed-upon solution.

2.5 RESPONSIBILITY AND ACCOUNTABILITY

Using an algorithm to make decisions or support decision makers can blur or transform traditional notions of responsibility for those decisions. Predictive algorithms typically produce predictive output with a degree of uncertainty. Yet, end-users who act on the predictive output are often uninformed of the level of uncertainty associated with the output and can habitually over trust the technology. When an algorithm generates a false predictive output that results in an undesirable consequence (e.g., an accident occurs at a site that an algorithm predicted to be low hazard), how should the responsibility for the false prediction be distributed between the decision makers? What kind of a trust relationship between the technology and the user is appropriate for a given application?

These are some of the many questions that need to be considered by organizations as they design and implement predictive algorithms in their workflow. Currently, there are no universal answers to address questions that relate to issues of responsibility and accountability of a predictive algorithm. However, application-specific analysis of the algorithm and its use by stakeholder groups can help generate solutions customized for a particular use of the algorithm.

2.6 IMPACT ON JOBS OR EXPERTISE

As algorithmic predictions improve and are able to outperform humans in specific tasks or areas of expertise, there can be an increased pressure to further implement algorithmic solutions and simplify work performed by humans. As the scope of work performed by an algorithm broadens, this pressure will only increase. Unlike the automation of repetitive physical tasks, the automation of cognitive tasks raises concerns that machines will be occupying the “last refuge” of human labour, leaving more jobs permanently displaced or job descriptions significantly changed. In other words, the fear is that machines will be occupying roles that are uniquely human for the long haul. Currently, there is increasing pressure on corporations and policymakers to acknowledge this prospect and plan for a broad displacement of knowledge economy workers. Some advocate for the need to introduce re-training programs for workers, such that the displaced workers are trained for new jobs. Others emphasize the need to develop technologies that aim to assist and enable human workers rather than replace them. The practicality of the two approaches depends on the use of the technology in question, the nature of the jobs being affected, and the skillsets of those whose jobs are most impacted.

3. GENERATION R'S METHODOLOGY

Keeping in mind the high-level issues outlined in the above sections, Generation R conducted an ethics assessment that aims to inform BCSA's design and deployment of a predictive algorithm that is being designed to supersede an existing system, RAP. While BCSA does not distinguish the legacy system from the new system, for the purposes of clarity and simplicity, we refer to the existing program as RAP 1.0, and distinguish it from the new version of RAP, RAP 2.0, which is the focus of our assessment.

Generation R's ethics assessment methodology involves a detailed investigation of three different facets of a technology: components of the technological system, social and organizational dynamics of the stakeholders of the technology, and the network of stakeholder values that interact and pose value tensions when individuals encounter and use the technology. What is presented as a result of our method (presented in Section 4) is an orchestration of the findings from our analysis that provide an insight on foreseeable challenges that the technology can pose on an organization.

The results from our three-part analysis allowed us to discover a comprehensive set of foreseeable challenges specific to RAP, which ultimately led us to generate recommendations to address the potential challenges presented in Section 4.

Through a discussion with members of the Data Analytics and Decision Science department and our initial interviews, we identified a number of stakeholder groups. We interviewed a total of 21 participants from across the stakeholder groups except for the public and duty holders. Those we interviewed include: safety officers, senior safety officers, members of the Data Analytics and Decision Science division, members of Information Technology team, regional business leaders, safety managers and employees with who focus on policy and privacy, members of the stakeholder engagement team and senior leaders of BCSA. The data collected from these interviews provided us with a rich amount of information sufficient for our analysis. Engaging the public and duty holders was deemed to be outside the scope of our analysis of RAP 2.0, especially given our focus on the primary use case of RAP 2.0. In the remainder of this report, we use the term RAP development team to refer to individuals who directly work to development of RAP 2.0. This includes members of the Data Analytics and Decision Science department who develop the algorithm and members of the Client Experience department who provide IT support for RAP 2.0.

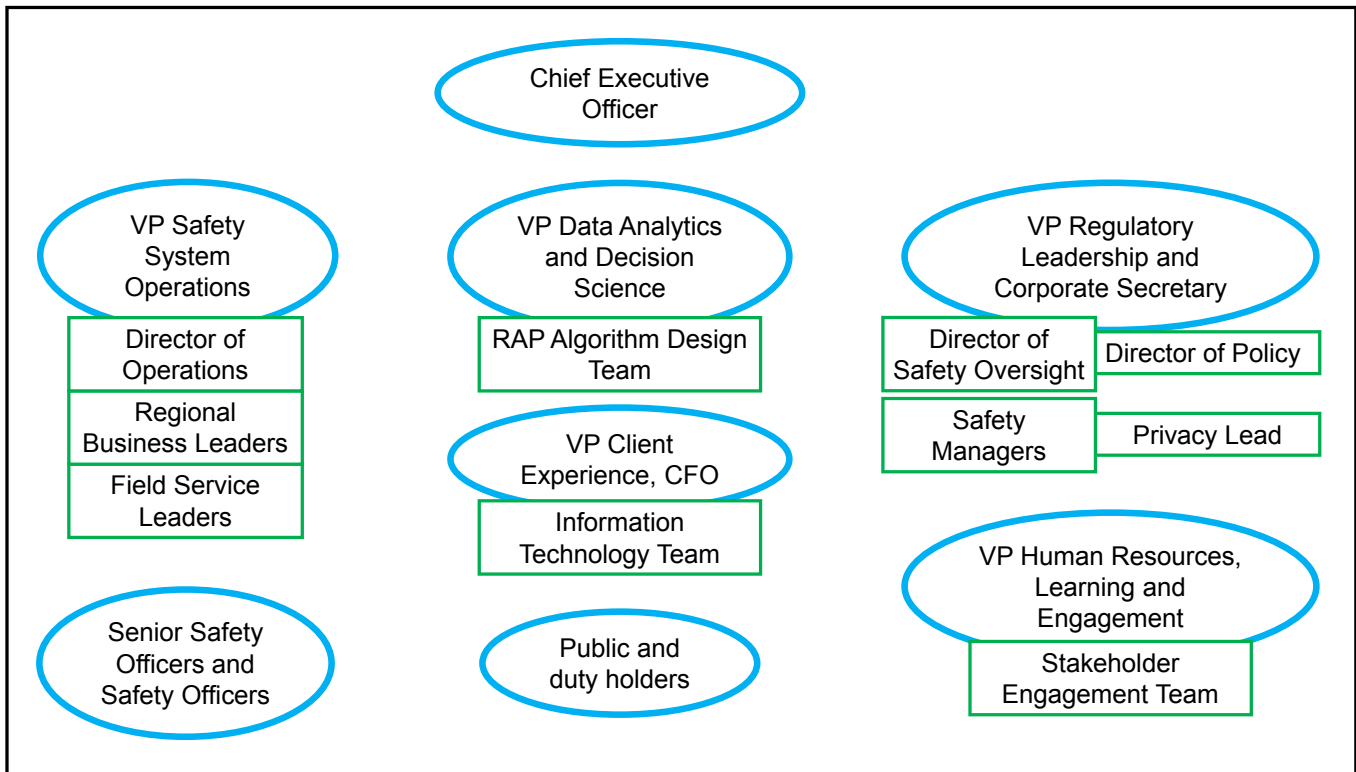


Figure 3.1 This diagram illustrates a list of identified set of stakeholder groups.

BCSA provided us with a list of factors used for RAP 1.0, a demonstration of and screen captures from the user interface currently used to conduct inspections and collect data by safety officers, and a short description of the output that RAP 2.0 is being designed to deliver. We also perused readily available application forms duty holders use to apply for electrical permits from BCSA.

We find that RAP 2.0 – which, at the time of writing this report, is still undergoing pilot trials – is being perceived and discussed by individuals throughout BCSA as an extension of RAP 1.0. Noting the significance of the historical link to RAP 1.0, we expect that much of the same social challenges and value tensions will carry over to RAP 2.0. Therefore, our analysis and recommendations are inherently contextualized with RAP 1.0 in mind.

4. FINDINGS AND ASSOCIATED ITEMS OF CONSIDERATION

In this section, we present the main findings from our analysis. We begin with a brief historical background of RAP 2.0, the predictive algorithm that BCSA is developing, in order to provide context of our analysis. Three RAP 2.0 use cases were identified in consultation with BCSA, which set the scope of our analysis. Use Case 1 (Section 4.1) pertains to safety officers' use of RAP 2.0 as a decision support system. Use Case 2 (Section 4.2) relates to the use of RAP 2.0 and its output for strategic decision making by BCSA's senior leaders. Finally, Section 4.3 provides a short summary of Use Case 3, which involves public reporting of RAP 2.0 outputs.

We paid special attention to Use Case 1, as this has been identified as the primary use case for RAP 2.0. Readers will find that many of the challenges raised in Use Case 1 also apply to Use Cases 2 and 3. Where there are repeated findings of challenges across the use cases, we present them as part of Use Case 1. Sections 4.2, and 4.3, therefore, contain challenges that are unique to those use cases. Use Case 3 is the least well-defined of the three use cases, and require further investigation after clear objectives and target stakeholders have been identified.

4.0.1 A HISTORICAL CONTEXT: RAP 1.0 AS THE PREDECESSOR OF RAP 2.0

Current day-in-the-life of a safety officer

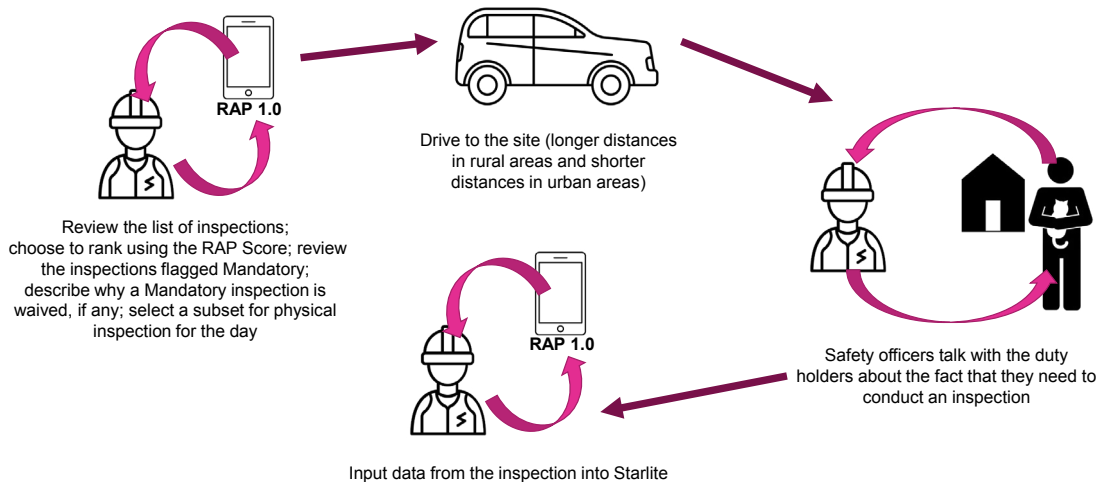


Figure 4.1 This diagram depicts what a safety officer's typical day looks like with respect to their use

RAP is a software-based decision support system that is primarily used by safety officers. It was developed to help prioritize the physical permit inspections of technical systems that BCSA conducts to oversee public safety within its jurisdiction. RAP's main output is a RAP Score, a quantitative score (a positive numerical value) computed for each permit subject to an inspection. RAP 1.0 is built on a linear model; its output represents a summation of a set of factors, each of which are multiplied by a fixed numerical multiplier.

In the case of electrical installation permits, each RAP Score is classified into one of three Priority Rating categories – Mandatory, Discretionary, and Low-Priority -- based on a set of threshold values that BCSA has predetermined. For example, if the threshold values are set to 80 and 200, then a permit with a RAP Score below 80 will be marked as Low-Priority and automatically waived by the system. Those above 200 will be automatically flagged as Mandatory, while those between 80 and 200 will be considered Discretionary.

Data collected from permit applications and permit inspections vary across the seven technology sectors that BCSA oversees. Hence, data fields used to compute RAP Scores also vary from one technology sector to another. We expect these cross-technology variances to exist in RAP 2.0. Therefore, for the purpose of this report we focused our analysis on the Electrical technology sector, which has been identified as a priority for deploying the new RAP.

Safety officers, the primary users of RAP 1.0, are not obligated to conduct physical inspections for permits flagged as Mandatory by RAP 1.0; they can waive them through the user interface by providing a justification. Safety officers encounter the outputs as they review permit inspection tasks within their geographical region of responsibility using either a touchscreen-based interfacing system called Starlite, or a desktop application called Star. These inspection tasks include newly installed assets that require immediate assessments, incidents related to assets that require an investigation, and assessments of assets flagged as Mandatory or Discretionary by RAP 1.0. All safety officers have access to the RAP Score associated with each task, but they do not have direct control over the threshold values used to classify inspection tasks as Mandatory, Discretionary or Low-Priority.

4.0.2 UNDERSTANDING RAP 2.0

Safety officers, as the primary users of RAP 1.0, have had years of history with the decision support system and its role in their daily work. **RAP 2.0 is currently being developed as an improved version of RAP 1.0 that can better support safety officers. Using data-driven, machine learning methods, RAP 2.0 is being designed to extract patterns from the data available to BCSA, as a means of predicting the probability that safety officers will find high level of hazards at a particular permit site.** One of the many factors that motivated the decision to use machine learning for RAP 2.0 is the availability of the As-Found Hazard, a standardized hazard ratings framework that safety officers currently use as part of their physical inspection. This framework allows safety officers to rate observed hazards from a physical inspection on a scale of 0 to 5, where a score of 3, 4, or 5 are considered high hazard.

In contrast to RAP 1.0, which provides a RAP Score and a Priority Rating category, the planned main output of RAP 2.0 is a probability indicating the likelihood that a safety officer will find a high hazard on a permit inspection, that is, that they will submit an inspection report with an As-Found Hazard rating of 3, 4, or 5 upon conducting a physical inspection of the permit.

While the modeling technique (computation of RAP Score) used for RAP 1.0 is fixed and can be described as a summation of a predetermined set of factors and their associated weights, the selection of a predictive algorithm and the factors to be used to develop RAP 2.0 remains open-ended. Table 4.1 provides a brief comparison between RAP 1.0 and RAP 2.0. At the time of writing this report, we are aware that a pilot study of RAP 2.0 is being conducted. Noting that the design and deployment of RAP 2.0 will undergo iterative design and testing processes, we conducted our analysis with the assumption that any permit-related data currently collected by BCSA, including those not used in RAP 1.0, could be used to develop and improve RAP 2.0 in the future, and that several machine learning algorithms could be employed to produce the probability output. As such, rather than providing constraints on the scientific and creative processes that can help innovate RAP 2.0, we aim in our analysis to highlight the unique set of challenges and recommendations that can help BCSA make informed design, deployment, policy, and operational decisions.

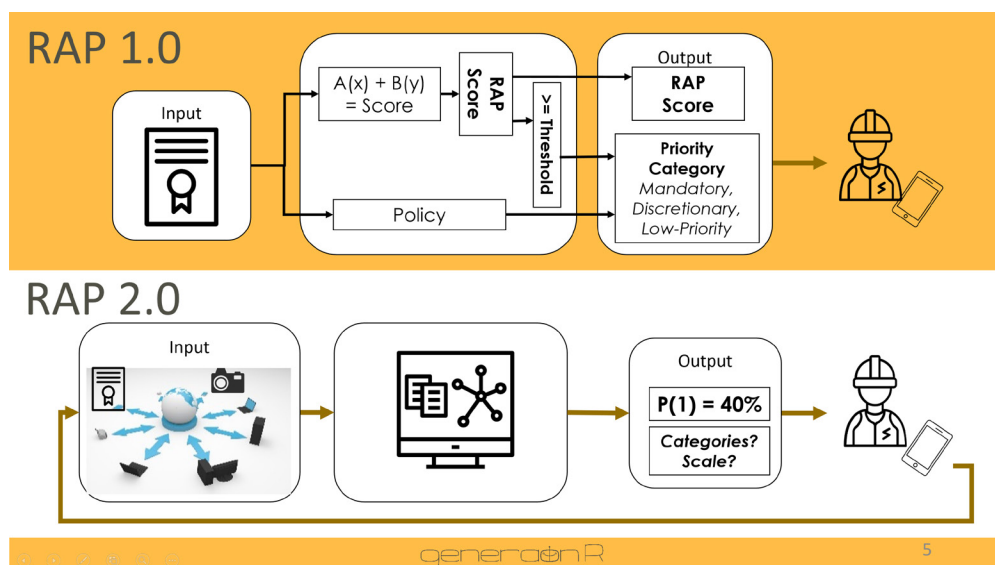


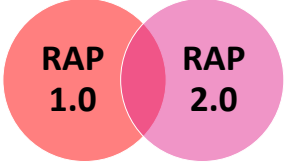
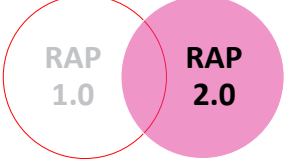

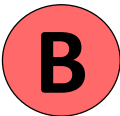

Figure 4.2 An illustration of key differences between RAP 1.0 and RAP 2.0. While RAP 1.0 can be described as a linear system with a fixed set of parameters, RAP 2.0 is being designed as a data-driven system that seeks statistical patterns from existing data, and can constantly improve its performance as safety officers enter new data from their inspections.

Table 4.1. Summary of key technical differences between RAP 1.0 and RAP 2.0

	RAP 1.0	RAP 2.0
INPUT	Permit data and the latest inspection data	Permit data. Larger set of data (cross technology) available to BCSA. Possibility to integrate data from other sources (e.g., BC Hydro)
OUTPUT	RAP Score RAP Priority Rating (Mandatory, Discretionary, Low Priority)	Probability of finding hazard level 3, 4, 5 Categorization of RAP Probability (assumed)
NATURE OF MODELING TECHNIQUE	Linear. Fixed weights are assigned to a fixed set of factors	Data-driven. Parameter values used for the algorithm can be optimized and changed based on findings from data.
SELECTION OF FACTORS (E.G., TYPE OF INSPECTION, VOLTAGE) USED FOR COMPUTATION	Discussion with safety officers from unspecified number of years ago.	Statistically derived. Multiple factors may be combined in statistically meaningful ways.
MODEL UPDATE FREQUENCY	Unspecified	Update frequency to be determined. Can be made as frequently or infrequently as desired (e.g., hourly, daily, monthly)

Some of the challenges we identified from our assessment are either unique to the introduction of RAP 2.0, or are inherited from RAP 1.0. Our recommendations or items for consideration to address these challenges are framed as design considerations, business decisions, or policy considerations. We distinguish the challenges and recommendations using visual labels shown in Table 4.2.

Table 4.2. Visual glossary

VISUAL	MEANING
	<p>Challenge relevant to both RAP 1.0 and RAP 2.0</p>
	<p>Challenge relevant to RAP 2.0</p>
	<p>Design Consideration</p>
	<p>Business Consideration</p>
	<p>Policy Consideration</p>

4.1 USE CASE 1: RAP 2.0 AS A DECISION SUPPORT TOOL FOR SAFETY OFFICERS

The primary use case of RAP 2.0 is to help safety officers prioritize their physical inspection tasks. This is an extension of the function RAP 1.0 is designed to serve (see Figure 4.0.1). As is the case with RAP 1.0, we expect safety officers to access RAP 2.0 probability outputs through the Starlite or the Star interface where they can browse and select permit inspection tasks within the geographical region assigned to them. In this use case, safety officers are the primary users of RAP 2.0.

SUMMARY OF KEY CHALLENGES AND RECOMMENDATIONS

In our assessment of RAP 2.0, we have identified transparency as one of the key values that must be considered for a successful launch of RAP 2.0. Issues of transparency and interpretability relate to a specific set of stakeholders who require or desire a certain set of information and how readily accessible and understandable the information is to them. In the case of RAP 2.0, safety officers' limited access to, and lack of understanding of, the internal components of RAP have been identified as two of the key issues that must be considered. These issues are closely related not only to the need safety officers have to make informed decisions based on RAP outputs, but also to their ability to communicate and use the output to the duty holders in a consistent, clear and confident way. Transparency is inherently coupled with the trust dynamics stakeholders have with RAP 2.0. Increased transparency tends to increase trust and confidence in a technology, whereas lack of transparency can result in distrust that can be difficult to overcome.

Figure 4.3 illustrates where transparency is relevant within this use case with respect to its key stakeholders and RAP 2.0. Our recommendations focus on improving key transparency links.

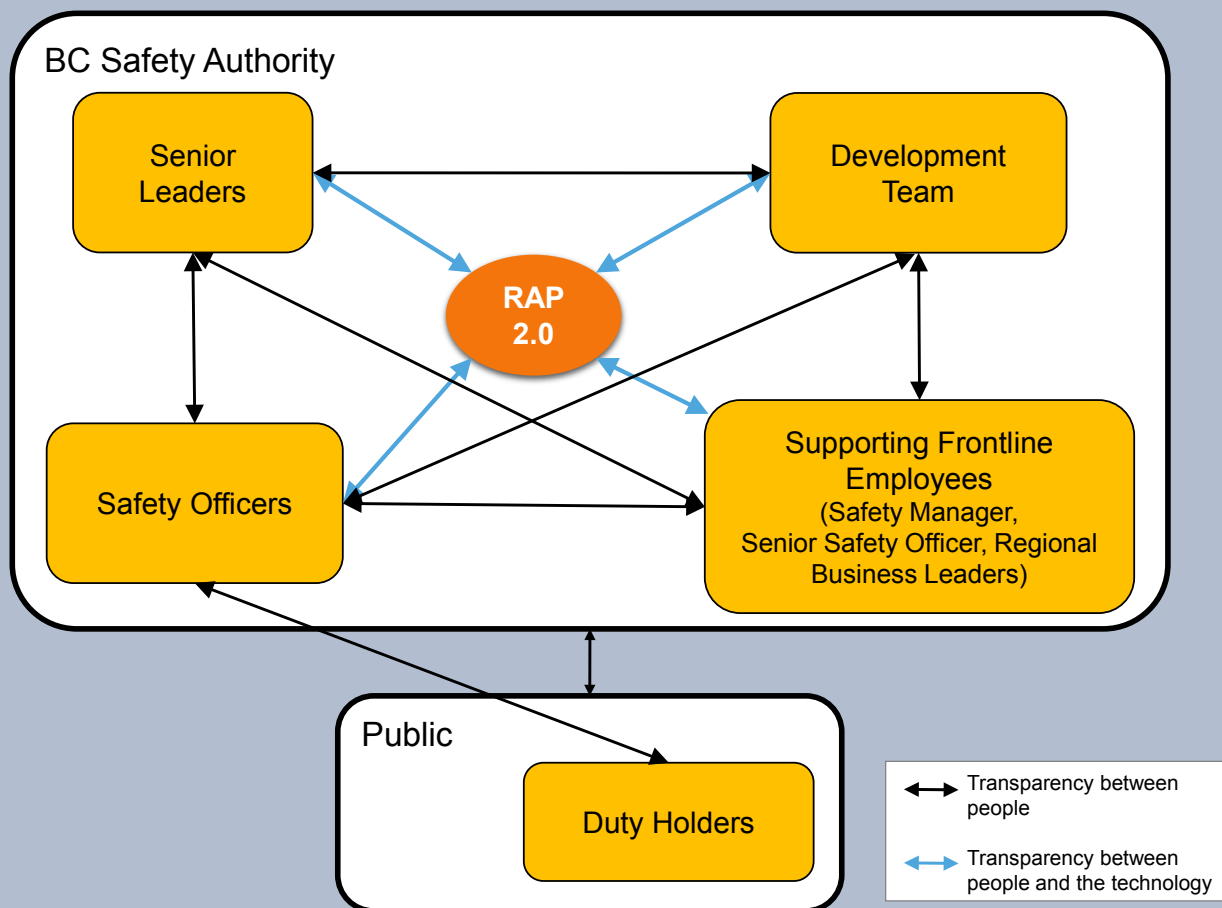
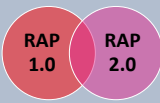


Figure 4.3 Transparency can be fostered or hindered at the information exchange link between a pair of stakeholder groups. This diagram illustrates a mapping of the network of transparency in BCSA with respect to the use and development of RAP 2.0.



Challenge 4.1.1.1: The interpretability of RAP outputs and its impact on Safety Officers' decision making process

RAP 2.0 is meant to be a decision support tool for safety officers that enables them to prioritize inspections based on a combination of their own frontline expert knowledge and observations, and the RAP 2.0 output. One thing we learned from our interviews is that to ensure that the RAP 2.0 output is useful for the safety officers in their decision-making process, it is critical that they are able to interpret the output. A RAP output is interpretable when the safety officer has a good idea of what the output represents in their frontline context. For example, does it represent the probability of finding an “as found” hazard of 3-5 in the inspection site? Does it represent the probability of finding non-compliance in the inspection site? Or does it represent the behavioural differences between different contractors working in a region? The score and the factors leading to it need to be transparent and contextualized for the safety officers so that they can effectively bring that output into their own expert decision making process.

Currently, RAP 1.0 places permit inspection tasks into one of three categories based on the score: mandatory, discretionary or automatically waived. The RAP Score, and a breakdown of it, are provided to the safety officers in the Starlite system. According to our findings from interviews, RAP 1.0 output is not easily interpretable by safety officers and this issue could carry over to the implementation of RAP 2.0. Most of our study participants noted that the RAP Score is not transparent to them, and that they do not understand what the score indicates. Interestingly, when we asked the safety officers to describe how RAP Scores are computed, none of them referred to the breakdown of the RAP Score that is currently provided in the Starlite system. One participant expressed concerns over the fact that changes are made all the time to RAP 1.0, but that safety officers tend to be unaware what those changes are or why they were made. In addition, the interviews suggested that there is an inconsistency with how each one of the safety officers is interpreting the RAP Score. Some believe it is an indication of level of risk present at an inspection site, while others believe it is a score that simply prioritizes their work. They were not aware of the linear model of RAP 1.0 or its details. It is noteworthy that the senior safety officers had more in-depth knowledge of RAP 1.0 because they had been more involved in shap-

ing the linear model. It was clear from the interviews that the safety officers are not able to easily and accurately interpret the information that they are given from RAP 1.0. This prevents them from truly using RAP 1.0 output in their decision-making process.

Recommendations:

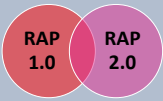
D 4.1 Engage safety officers on the level of transparency that is necessary for them to interpret and use RAP 2.0 output to prioritize their daily work. It would be useful to engage senior safety officers in this discussion given their vested interest in understanding and improving RAP. This will also lead to necessary interface design decisions on how RAP 2.0 output should be communicated to the safety officers to help them easily recognize and comprehend the information presented to them.

B 4.2 Develop a metric/scale that allows the RAP development team to gauge how transparent the system is with respect to their primary users (safety officers). This process can involve identifying the type of information each stakeholder requires and assessing the accessibility of the information by the stakeholders. It can also involve surveys or behavioural analyses. Currently, no universal transparency metric exists that can be used across application domains. Therefore, a customized metric is recommended.

Items for Consideration:

D 4.3 Provide a feedback mechanism in the RAP 2.0 interface that allows safety officers to indicate their level of understanding they have on the received RAP 2.0 output and their agreeableness of the output.

P 4.4 Develop internal guidelines with key stakeholders about what it means for BCSA to make RAP transparent. Internal guidelines can help guide consistent RAP development, and ensure that key stakeholders have a better understanding of the tool and its outputs.



Challenge 4.1.1.2: Trusting RAP

RAP 2.0 is a decision support tool, the output of which is meant to be incorporated into safety officers' daily decision making processes. Therefore, it is imperative that safety officers have a balanced trust dynamic with the output of RAP 2.0 that does not lead them to over- or under-trust the system. This is especially important since predictive algorithms built using machine learning techniques heavily rely on having a high-quality data input for continuous performance improvements, and a majority of this data is collected and entered by safety officers who directly interact with duty holders and the geographical community BCSA serves. As both the primary users of RAP 2.0 and sources of input data, safety officers' acceptance of RAP 2.0 is crucial for the successful integration of the technology. However, in our interviews these stakeholders were also the ones who strongly expressed their distrust of RAP 1.0 and its output (i.e., RAP Scores). Figure 4.4 illustrates the trust links between the relevant stakeholders and RAP.

There are three elements that currently affect safety officers' trust in RAP 1.0, which are likely to impact their trust relationship with future RAP versions:

1. The level of understanding that safety officers have of the RAP model/algorithms: currently, safety officers consider the internal workings of RAP 1.0 opaque

2. The perceived validity of the RAP output

a. Safety officers, who have frontline expertise accumulated through their day-to-day inspection work, do not feel that their frontline expertise is reflected in the RAP Score (the main output of RAP 1.0)

b. RAP Scores have often led safety officers to inspection sites that are low hazard, which have led to a culture of under-trusting, and thereby not valuing, the RAP 1.0 output in their daily work

3. The actual validity of the RAP output: safety officers who have frontline expertise have expressed a tension between balancing the perceived objectivity of data (and the need for scientific validity of RAP Scores) against the contextual (subjective) nature of that very same data

It is important to note that while safety officers and senior safety officers expressed their distrust of RAP 1.0, it is also clear that the safety officers, along with other stakeholder groups, acknowledge the need and potential utility of RAP 2.0 in their daily work.

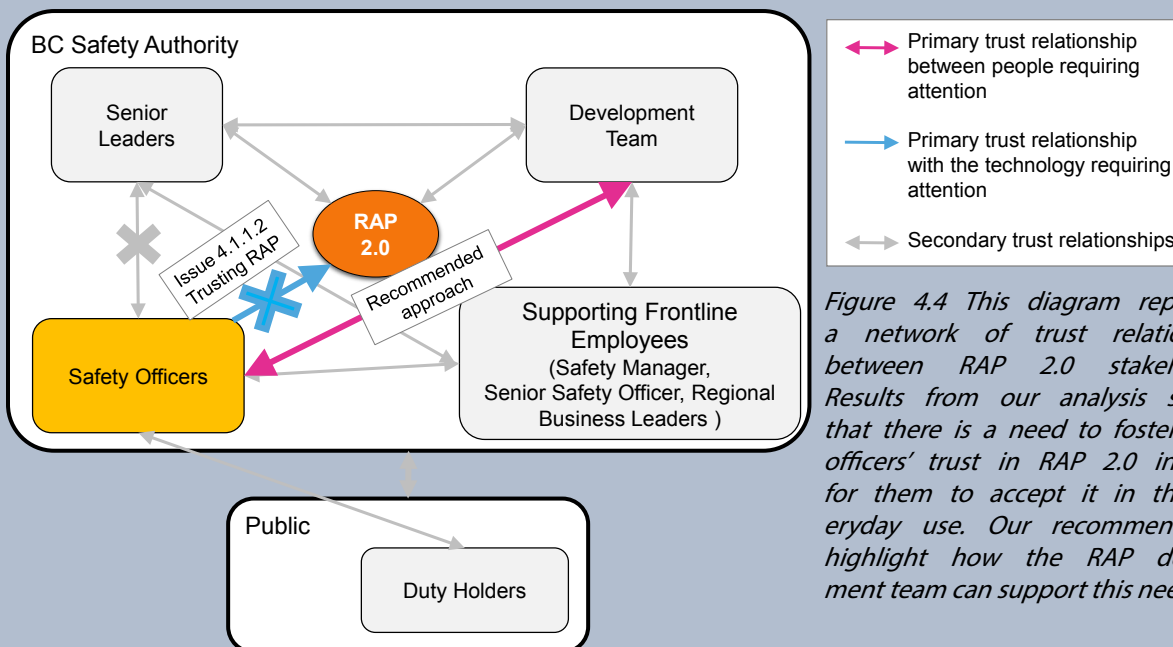


Figure 4.4 This diagram represents a network of trust relationships between RAP 2.0 stakeholders. Results from our analysis suggest that there is a need to foster safety officers' trust in RAP 2.0 in order for them to accept it in their everyday use. Our recommendations highlight how the RAP development team can support this need.

Recommendations:

B 4.5 Clarify and communicate the objectives of the RAP 2.0 program. Part of the perceived lack of transparency with respect to the inner workings of RAP 1.0 stems from the fact that RAP 1.0 has been designed to serve too many objectives (e.g. communicate policy-driven priority inspections for safety officers, provide a heuristic of risks associated with permits waiting to be inspected, and manage the limited safety officers' use of time to be used where their attention is needed most). Therefore, the safety officers are confused about what a RAP Score is supposed to represent. Communicating and clarifying the key design objective that RAP 2.0 is built to serve will help mitigate this confusion, improve transparency, and could positively affect the trust dynamics with safety officers and other stakeholders.

D 4.6 Maximize inclusive design practices that specifically include safety officers (and other stakeholders) in the design process. Predictive algorithms are often designed by statistically exploring patterns available in a given set of data, which necessarily requires the expertise of data scientists and engineers on the RAP development team. That development team needs to be able to function independently from the primary audience of the predictive algorithm's output. However, it is clear from our investigation that, while the development team may view RAP 2.0 as a predictive algorithm (a piece of software), the nature of how the technology affects safety officers takes the form of a technical system that cannot be separated from the trust dynamics users have with the output as well as the user interface that affect this dynamics: RAP is a socio-technical system. Therefore, inclusive design practices common among user interface and user experience designers can help improve both the usability and the user acceptance of the technology. In particular, we have found that BCSA's safety officers take pride in their technical abilities and aptitudes. Given this understanding of the user group, it would help to involve them as much as possible in the design process (be it during the process of interface design, or providing opportunities for them to provide feedback on the usability/performance of the system) in order to instill a sense of ownership of RAP 2.0 among the safety officers.

SIDE NOTE

Opportunities for safety officers to provide direct feedback to the development team can take the following (among other) forms:

- designing an interface that specifically asks feedback from safety officers
- brainstorming sessions during interface design phase of the system to highlight how they would use RAP 2.0 in their day-to-day
- openly requesting volunteers to help test-drive new algorithms/interfaces
- holding regular training sessions on appropriate use of RAP 2.0 in day-to-day operation of safety officers that clarify what the input and the output of the system represent
- co-designing and utilizing customized metrics that help measure what safety officers view as important in trusting and actively using the system

Challenge 4.1.1.3: Trusting RAP



As frontline employees, safety officers interact regularly with BCSA clients. Clients sometimes question why they are being inspected.

In some cases, safety officers incorporate the RAP Scores into their explanations to the clients (e.g. explain that the client’s permit has been flagged as a “Mandatory” inspection” by RAP 1.0), requiring safety officers to sometimes describe what RAP is and provide the RAP Score to the duty holder in terms that the duty holder can understand. In those interactions, any misunderstanding safety officers have of RAP can be passed on to the clients. The clients might also receive different explanations depending on which safety officer they are talking with. A few of the participants described how RAP Scores are currently discussed with duty holder on an ad hoc basis, which could create confusion in interactions, and could contribute to inaccurate perceptions of RAP both from the client’s perspective, and from the safety officer’s perspective. Given that RAP 2.0 is likely to be not only more complex to understand but also harder to explain in lay language than RAP 1.0, it is likely that safety officers will have an increased need to understand RAP 2.0 to be able to explain it to duty holders. Figure 4.5 illustrates the transparency link that poses an explainability challenge and our recommended approach.

Items for Consideration:

D B 4.7 Addressing the issues of interpretability (Section 4.1.1.1) and trust (Section 4.1.1.2) would help make RAP 2.0 more explainable for the BCSA clients.

D B 4.8 Developing guidelines in consultation with key stakeholders (e.g. safety officers) on how to describe RAP 2.0 and its output to duty holders would help mitigate possible misunderstandings duty holders may develop about BCSA’s use of predictive algorithms in the future.

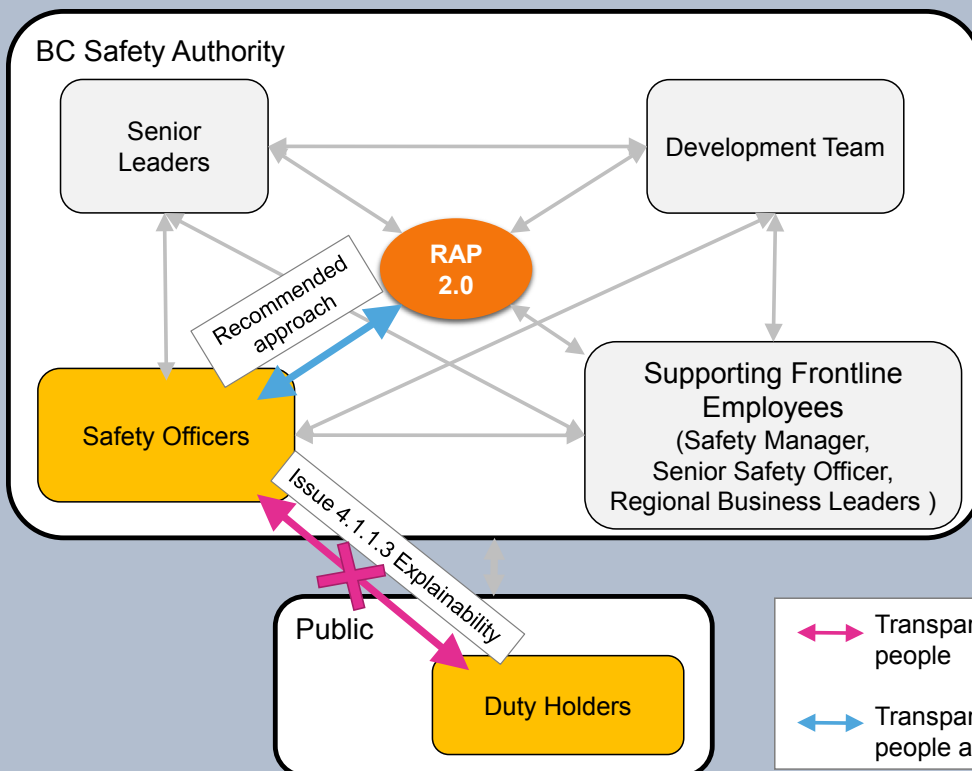


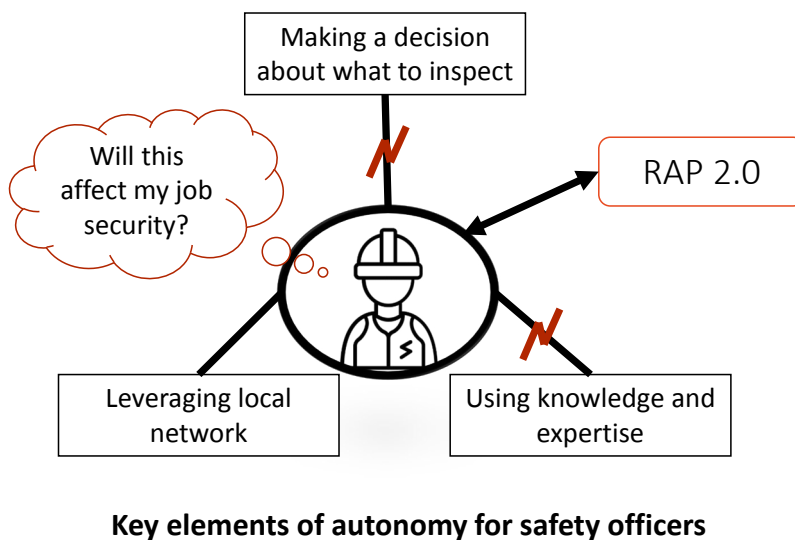
Figure 4.5 One of the issues of transparency arises from the fact that safety officers sometimes use RAP Scores to explain reasons for their inspections to clients. As representatives of BCSA who directly interact with duty holders, it is expected that safety officers will have an increased need to understand, and be able to clearly describe to the duty holder, the function and probability output of RAP 2.0.

SUMMARY OF KEY CHALLENGES AND RECOMMENDATIONS

Serving public safety is a common objective for safety officers, the RAP development team, and the senior leaders of BCSA, among others. Participants from these three stakeholder groups echoed the notion that they are “partners in safety” for the duty holders, rather than enforcers of safety rules and regulations. Indeed, the organizational value that the participants of our analysis consistently mentioned was the recognition and endorsement of public safety as a key value in their daily workflow. Stakeholders’ ability to have a positive impact on public safety includes (and enhances) the freedom to exercise their professional autonomy, that is, the use of their own means, expertise, sense of control, and creativity in performing their work to serve public safety. Activities that interfere or hinder this exercise of autonomy, on the other hand, can contribute to feelings of frustration and dismissal, which could lead to low levels of acceptance of RAP 2.0 and a decrease in the quality of data collected for RAP 2.0.

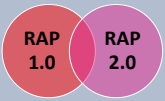
Professional autonomy is important for both safety officers, who are the primary users of RAP 2.0 and maintain and utilize valuable frontline knowledge, and the developers of RAP 2.0, who extract values with “centralized” (as opposed to “frontline”) knowledge. The differences between frontline and centralized knowledge that each of the stakeholder groups brings to the table are useful for achieving the shared objective of public safety. However, the differences between the knowledge categories is currently a source of tension between stakeholder groups because of how RAP relates to the stakeholders’ exercise of professional autonomy.

In forecasting the decision support role that RAP 2.0 will play in a safety officer’s daily workflow, developers of RAP 2.0 would benefit from acknowledging the intricate impact RAP 1.0 has already had on many safety officers’ sense of autonomy. For instance, certain inspections are more preferred by safety officers than others due to factors such as proximity to the safety officer and familiarity with the type of inspection. RAP Score, in part, nudges safety officers to visit sites that may not be desirable or preferred by safety officers but nonetheless requires their expert attention. This forms a tension between safety officer’s autonomy and machine autonomy, where public safety depends on a careful balance of the two. The challenge moving forward is to work toward building features into RAP 2.0, and the workflows surrounding it, that improve stakeholders’ ability to work toward the shared value of public safety, but that also appropriately maintain or enhance their professional autonomy.



While professional autonomy for the RAP development team takes the shape of having the flexibility to design and formulate algorithms using their expertise, for safety officers professional autonomy includes the following key aspects: a) the ability to use their expertise to directly contribute to public safety; b) the ability to understand and influence RAP output, and c) the ability to provide feedback on the continuous development and improvement of RAP and the workflows surrounding it.

Figure 4.6 For safety officers, professional autonomy includes their ability to use their expertise to directly contribute to public safety, the ability to understand and influence RAP output, and the ability to provide feedback on the continuous development and improvement of RAP and the workflows surrounding it.



Challenge 4.1.2.1 : Confusion with the categorization of RAP Scores

There is an important sense in which RAP 1.0 impinges on a safety officer's ability to exercise their expertise and, thus, their professional autonomy. This stems from the fact that RAP Scores are categorized as Mandatory, Discretionary, or Low-Priority, which are interpreted as commands upon which safety officers are to act. Factors used to compute RAP Scores were selected based on safety officer insight. Unfortunately, in practice, RAP Scores often do not reflect the task prioritization needs and interests of safety officers -- a fact that was widely acknowledged by a majority of the stakeholders we interviewed within the organization. The tension between the low reliability of RAP Scores and the command-like language used to group RAP Scores into task priority levels (e.g., Mandatory) has been problematic. As one safety officer put it:

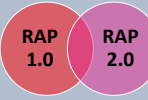
"It drives [me] crazy...to be driving past things that I know I should be doing, and would have a great impact on safety, to go do something that a computer says is important, that I know full well is not important at all."

Furthermore, reporting the categorization of RAP Scores to safety officers seems to have served a somewhat superficial role. Even an inspection categorized as Mandatory can be waived by safety officers as long as a valid reason is provided, and RAP Scores seem to have led safety officers to attend to inspections that they correctly do not consider as high hazards. Even though the tasks assigned by RAP 1.0 only forms a minority of total inspection tasks assigned to safety officers, our interviews suggest that RAP Score categorizations tend to be considered an annoyance, often dismissed by the safety officers.

It was clear from our interviews that safety officers understand the need for, and possible utility of, a decision support tool such as RAP, and that the redesign of RAP stems from management's desire to support safety officers' need to better prioritize their daily inspection tasks. It was also clear that the safety officers we spoke to are very much in favour of using RAP, especially if RAP can serve the function of providing additional (and enabling) information to them, that is, if it helps them make good decisions about which inspections most likely require their expert attention.

Items for Consideration:

- D** 4.9 Separate command-like categorizations from the enabling information - A model that could increase a sense of autonomy for the safety officers while providing a useful support system for efficient resourcing decisions, is one that separates the command-like inspection categories (Mandatory, etc.) that take away from safety officer autonomy, from RAP 2.0 probability output. That way, RAP 2.0 output can serve as useful enabling information that safety officers can take into account, while the ultimate decision to attend to an inspection or not remains in safety officers control.
- D** 4.10 Provide a scale (e.g., high, medium, low) that communicates an internally consistent method of understanding the severity of the RAP 2.0 probability output. Such a scale would not be meant to dictate where safety officers must go, but can allow them to make more sense of the information provided.



Challenge 4.1.2.2 : The need for safety officers to provide feedback on RAP output and performance

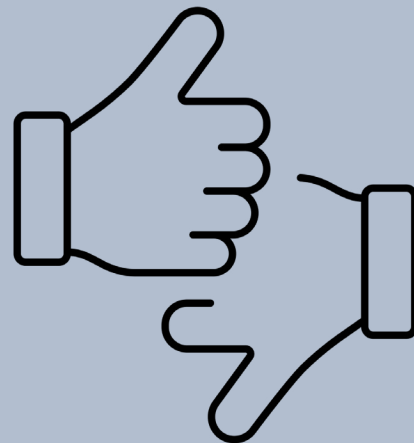
BCSA has a need to prioritize inspections to maximize public safety. Our interviews indicated that safety officers share this goal. Furthermore, safety officers fully acknowledge the complexities involved with using data-driven prediction models to prioritize work. As one participant put it, “I don’t know how you make something more predictive of human behaviour if you can’t input the fine points of human behaviour.” At the same time, frontline stakeholders indicated a willingness to help improve RAP, sometimes describing a general need for a lot more safety officer input into the algorithms, other times indicating the need for continuous safety officer validation as a means of maintaining trust in the system, yet other times describing how safety officers’ unique contextual knowledge of the data will be invaluable for the success of RAP 2.0.

Though RAP 1.0 was supposed to serve as an efficient prioritization tool, many safety officers are confused by the output that RAP 1.0 generates, where they attend to an inspection with a high RAP Score only to find no high hazard item there. While safety officers understand that these algorithm-based tools are not meant to be perfect, those frustrations have manifested themselves as the need to be able to provide input that corrects the underperforming RAP Scores.

It is noteworthy that the machine learning approach to designing RAP 2.0 can be adaptive in nature, which can improve its predictive performance over time. Collecting explicit feedback on RAP 2.0 from safety officer is not necessary to improve RAP 2.0 predictive performance. However, providing a convenient means (e.g., having a thumb up or thumb down button next to RAP 2.0 output) for safety officers to provide feedback on the performance of RAP 2.0 can help address their need to influence and evaluate the system that affects them daily.

D 4.11 Develop clear metrics or shared mechanism among stakeholders that allows for the monitoring of how well RAP 2.0 is performing, and review whether its performance is improving over time.

D 4.12 Provide safety officers a direct feedback mechanism on RAP 2.0 output (e.g., such as thumb up/down button). Regardless of whether the collected feedback is directly incorporated into RAP 2.0 algorithm or indirectly considered in improving the system, giving safety officers a convenient means to provide feedback on RAP 2.0 would help safety officers exercise their autonomy towards influencing the system they use on a daily basis. Such practice can also foster inclusiveness in RAP 2.0 design and improvement process while providing safety officers with a means to handle frustrating erroneous RAP 2.0 outputs that are likely to be frequent in the beginning stages of RAP 2.0 deployment.



Items for Consideration:



Challenge 4.1.2.3: Impact on Jobs and Expertise

We do not foresee the role of safety officer being replaced by RAP 2.0.

The observations and physical visits are essential to conducting physical inspections, and RAP 2.0 does not operate as an autonomous system, one that combines comprehensive sensing and action capabilities, and capable of visiting sites and autonomously collecting hazard-related data. The development of a fully autonomous system is a future possibility, although it is not currently being considered by BCSA. BCSA clearly values its frontline human resources and the face-to-face time that they spend with their clients.

Even though safety officers have their own ways of using frontline expertise for decision-making, they acknowledge that RAP 2.0 has the potential to improve their workflow and reduce their workload if it functions well. Furthermore, they see that their role can be shifted to one including more educational and in-person service that would allow them to attend to client triggered requests. However, as highlighted in issues 1 and 2, there are some friction points with how RAP 1.0 is integrated to their workflow. These issues, combined with the fear that some safety officers have of RAP replacing their jobs, need to be addressed to ensure the fluid integration of RAP 2.0.

Items for Consideration:

- D** 4.13 Foster internal discussions on a regular basis among key stakeholders on how RAP 2.0 can positively contribute to safety officers' workflow, and how it can enable them to act on their values of efficiency, educating the public and building constructive relationship with the BCSA clients.
- D** **B** 4.14 Address the issues of transparency and autonomy mentioned above to ensure that safety officers can effectively employ RAP 2.0 in their daily decision making process.

SUMMARY OF KEY CHALLENGES AND RECOMMENDATIONS

One of the key challenges of employing empirically-driven predictive algorithms in the domain of machine learning is that it is challenging to foresee and monitor how the output from the algorithm can lead to potential discriminatory and unfair practices. This is especially true of centralized decision support systems in which the training dataset is taken from data entered by people. In this section, we highlight potential types of discriminatory practices that RAP 2.0 can unintentionally enable given the dataset BCSA currently has available to them. The notion of discrimination is often coupled with the value of fairness. Our analysis indicates that fairness is a foundational value that both BCSA's safety officers and the senior leaders share. We found that what is considered to be fair varies from stakeholder to stakeholder, as well as from stakeholder group to stakeholder group. These variations in notions of fairness can lead to value tensions, especially if the function of a predictive algorithm reinforces one stakeholder group's understanding of fair practice but not the other's.

We highlight how proactive monitoring of patterns in RAP 2.0 output and its use by safety officers, coupled with intentional objective setting, can help ensure the development and use of RAP 2.0 as an enabling technology for BCSA and the public it serves. We also discuss potential issues that can arise in repurposing the collected dataset for training RAP 2.0 in the future.



Challenge 4.1.3.1: There is a need to monitor potential emergence of discriminatory practices resulting from predicting human behaviour

Currently, the computation of RAP Score includes variables that relate previous records of contractor or FSR performance, such as the pass/fail ratio of their previous permits, to the number of noncompliances recorded under their permit applications. Selecting these performance-related data in a predictive algorithm for training a predictive algorithm is a logical step, especially given that RAP 2.0 aims to predict the probability of finding a high As-Found Hazard (levels 3, 4, or 5) on a permit inspection. The expressed assumption in making this design choice is that an individual or organization that has a history of noncompliant safety practices is likely to violate safety codes again in the future. This assumption may be reasonable given that the majority of As-Found Hazards ratings directly map to the presence and type of noncompliances observed during a physical inspection, rather than technical failures of systems that naturally fail or deteriorate over time.

Using machine learning would not only test this assumption statistically, but also lead to discoveries of patterns in the data that may be unexpected. For example, patterns may emerge that relate to specific FSRs, contractors, asset owners, geographical regions, types of installation and permit, other details about the permit, and any combination of these factors. Employing behavioural and performance-related factors into the deployment of empirically-informed predictive algorithms can strengthen the predictive ability of RAP 2.0 and, at the same time, lead to discriminatory practices if the objectives of determining such patterns of noncompliances are not made clear by the senior leaders and development teams (see the side note example).

A HYPOTHETICAL DISCRIMINATORY PRACTICE SCENARIO

After using RAP2.0 for a while, safety officers responsible for a particular region realized that RAP2.0 tend to have higher probability output on permits in regions where contractors of a specific ethnicity have dominated the market. Safety officers develop a stereotype that the ethnic group often violates safety codes. With the intention of supporting public safety, the safety officers in the region conduct more inspections on permits with duty holders of the ethnic group. This unevenly increases the number of noncompliances recorded under the permits associated to the ethnic group, which the machine learning algorithm adapts to and further increases RAP2.0 probability output in the region to be higher. As a result, the duty holders of the ethnic group feel discriminated against and unfairly penalized as they compare the frequency of their interaction with safety officers with contractors of another ethnic group. When such pattern of inspection goes unnoticed, it can lead to unintended discriminatory practices that, unlike the linear model used in RAP 1.0, the empirically-driven nature of RAP 2.0 can catalyze. A clear and positive objective setting by the senior leaders and the RAP development team can interrupt the negative cycle of discriminatory practice without jeopardising public safety.

The possibility of discriminatory practices is a concern, not because of any known discriminatory practices at BCSA, but because safety officers are the main sampling mechanism for data collected by BCSA as well as the primary users influenced by RAP outputs. This means that which inspections the safety officers decide to conduct directly translates to the selection of inspections sampled for the predictive algorithm that, in turn, frames the dataset used to train the algorithm.

Safety officers' decisions to conduct physical inspections on a permit are informed by the front-line expertise they gather through working with the local community. This expertise can include knowledge such as whether a contractor is a recent immigrant practicing with different safety standards. Due to the diversity in the nature of the communities individual safety officers serve, different safety officers are likely to have different ways of selecting permits to conduct physical inspections. In monitoring RAP 2.0 and its use by safety officers, it is important to keep in mind these variabilities across regions and individual safety officers.

Recommendations:

B 4.15 Actively monitor sampling patterns and RAP 2.0 output probability patterns that emerge in order to avoid implementing or catalyzing potential discriminatory practices through the use of RAP 2.0. A review process could be developed that to take into account, and help to avoid, discriminatory pitfalls. Such monitoring practices, while keeping in mind potential factors that contribute to discrimination, could help disrupt the spread of discriminatory or unfair practices.

SAMPLING PRACTICES AND VARIABILITIES

It is likely that the inspection selection patterns (sampling behaviours) by safety officers who serve urban areas are different from those who serve rural areas due to the distance between the permit sites that they need to inspect as well as differences in nature and size of the communities.

It is also important to note that, while some safety officers already recognized how quality data entry is a crucial aspect to making RAP2.0 a successful part of the organization, these individuals also voiced their concern that there are inconsistencies on how different safety officers tend to enter data differently (especially with respect to the As-Found Hazard rating).

Items for Considerations:

B 4.16 When the individuals monitoring the system come across secondary findings, especially as they pertain to patterns about individuals or groups of individuals, the management team could help set a clear objective about how the information should be used by BCSA as an organization. This could help avoid the use of information that can potentially damage the organization's shared set of values. One approach to this is demonstrated in the motivating story (Section 1.1). In the motivating story, patterns discovered about individuals or groups of individuals are handled by trained individuals – separate from those who handle everyday inspection tasks – with the objective of developing targeted educational programs or interventions that support improved safe practices for the target individuals or groups.



Challenge 4.1.3.2: There is a need to monitor potential emergence of discriminatory practices resulting from predicting human behaviour

BCSA is collecting descriptive text and photos as part of their data entry system. Descriptive texts are provided by the duty holders (permit applicants) and safety officers, while photos (as far as we are aware) are collected by safety officers during site visits.

Currently, safety officers are aware that the data they enter through the Star and Starlite system is being used to compute in RAP Scores. Duty holders who apply for permits from BCSA are also given a consent statement on how their data will be used. However, part of the data entered by safety officers includes photos from inspection sites as well as text-based entries that can, in the future, be analyzed as part of the machine learning system implemented in RAP 2.0.

Our analysis suggests that duty holders and safety officers are unlikely to be aware of the possible secondary uses of the data they provide to BCSA or their implications, such as the use of image recognition or natural language processing algorithms to better inform RAP 2.0. As far as we are aware, BCSA does not have immediate plans to perform these secondary analyses. However, should BCSA consider the secondary use of such data, depending on the nature of the secondary analysis the stakeholders who provided the information should be made aware of this and perhaps given the option to opt out.

Depending on the type of processing/analysis techniques and libraries used to analyze such data in the future (currently not implemented in RAP 1.0), developers will need to be cautious of discriminatory effects or characterization of individuals/assets in an undesirable or stereotyped manner (see the example from the field).

AN EXAMPLE FROM THE FIELD

For example, in one experiment, sentiment analysis algorithm was used on the texts available on restaurant reviews. The results of the analysis revealed that Mexican restaurants were predicted to receive particularly low ratings because the training set used to develop the sentiment model employed words from the web that associated the word “Mexican” to have a negative connotation.

Recommendations:

P 4.17 Update the consent process in the permit application process where clients currently provide data that will be used for RAP 2.0. This would include a clear statement indicating the intended use of the data, which relates to a lay expression of the data analytics goals of BCSA, along with the option to opt out of the particular use of the information they provide.

Items for Consideration:

P 4.18 Develop policies around transparency of primary and secondary data use to inform safety officers about how the collected data is used for RAP 2.0. It is more likely that safety officers will be more engaged in the data collection process and contribute to entering more high quality data if they a) see the improved performance of RAP 2.0, b) feel confident about the use of input data they provide and how it can affect RAP 2.0, and c) have the power to decide which information to record or not, depending on their value and context-based judgement related to the inspection or the duty holders.



Challenge 4.1.3.3: Fairness in human vs. machine decision-making

As a decision support system, RAP 1.0 has been behaving as a proxy to communicate what the senior leaders and BCSA as an organization wish to prioritize at the operational level. Policy-driven decisions to prioritize certain types of permits (e.g., public schools) have been translated into the assignment of higher RAP Scores, which are categorized as mandatory inspection for safety officers. We found that this type of influence on safety officers' decisions is likely to be helpful only if the safety officers agree with the suggested decision, not only on the basis of whether RAP is useful in suggesting where risks are to be found, but also on the basis of whether the suggestion helps them conduct their work in a fair manner.

We find fairness to be considered a foundational value to various stakeholders within the organization, perhaps in part due to the fact that the organization is a regulatory body, which necessitates the prioritization of fair practices. However, our interviews suggest that different stakeholders seem to contextualize, or have open questions about, the notion of fair practice differently. For the senior leaders, fairness is expressed as operationalizing the fair distribution of safety officers' time – a limited and highly valued resource – allocated to permits that require attention. While it was clear that safety officers also value efficiency in their work, to them fairness is expressed in terms of being able to provide a sufficient amount of attention to all the geographic areas assigned to them – which can be sometimes at odds with operational fairness as expressed by the senior leaders. In addition, safety officers decide whether to enter certain data based on their value judgment of what is fair for the duty holder (e.g., asking and answering the question “would penalizing this contractor lead to a fair outcome?” or “would educating this duty holder, rather than making a penalizing data entry, be a better option?”). Though these conceptions of fairness differ, all of the stakeholder groups we interviewed echo the importance of embracing safety officers' frontline knowledge and human skills that are an essential part of BCSA's business and commitment to public safety.

As RAP 2.0 is a decision support tool for safety officers, it is important for the senior leaders as well as the safety officers to have a clear understanding of their shared, but often conflicting,

notions of fairness to understand how the algorithmic decisions made with RAP 2.0 can support the notion, rather than introduce tensions with human decisions made by safety officers.

Items for Consideration:

- B** 4.19 Work towards an institutional understanding of fairness that explicitly takes into account what fairness means to various stakeholders in BCSA. Establishing an agreed-upon, multifaceted but shared understanding of fairness can help in evaluations what role RAP 2.0 can play in BCSA's operations in terms of promoting more fair practices, or hindering such practices. Taking this step could help bring stakeholder groups to the same understanding about the notion of fairness that is important for BCSA, and ensure that a) RAP 2.0 supports the notion of fairness valued by the organization and b) help safety officers accept RAP 2.0 if its design serves their fairness objective.

SUMMARY OF KEY CHALLENGES AND RECOMMENDATIONS

Introducing machine learning into a decision support tool like RAP 2.0 will impact the nature of responsibilities and accountabilities. Understanding these shifts is critical within the context of an organization with the primary mandate of overseeing safety of technical systems within regions of BC. As expressed by one of the interviewees, BCSA has the role of providing safety oversight and the duty holder has the ultimate responsibility of ensuring that the technical system that they own, install, or operate is safe. This fundamental assignment of responsibilities are not expected to change with the introduction of RAP 2.0. However, RAP 2.0 helps discover new knowledge, use and handling of which BCSA will be responsible for.

Both BCSA senior leadership and RAP development team take on added responsibility of ensuring that RAP 2.0 achieves the program objective. As an extension of this responsibility, they need to investigate how erroneous outputs from RAP 2.0 affect the day-to-day operations of BCSA and its overall mandate. One aspect of data analytics is that it creates centralized knowledge within an organization. BCSA is interested in leveraging this centralized knowledge in their operations. However, the existence of this new set of knowledge means that various stakeholders will need to act upon it. In addition, this centralized knowledge is derived from a set of data. Manipulation of this data set could lead to inaccuracies in the information that emerges from RAP 2.0, which could impact BCSA's public reputation.

In order to ensure that data analytics is implemented responsibly within an organization, it is critical to devise communication strategies and practices around how the new centralized knowledge base will be handled within the organization. These strategies should be developed in collaboration with all of the stakeholders that are impacted by the new set of knowledge. In addition, RAP 2.0 inputs, algorithms and outputs need to be continuously monitored for biases and potential problematic cases.



Challenge 4.1.4.5: Implication of RAP 2.0 output performance

The RAP 2.0 can produce outputs that suggest a high probability of finding a high hazard when in fact there are none or low hazard to be found on site. Similarly, RAP 2.0 can produce a low probability output when in fact there is a high hazard. Encountering these types of misleading output – which statisticians refer to as false positive and false negative, respective – is unavoidable and is inherent to probabilistic, data-driven systems such as RAP 2.0. To illustrate, a 98% probability of finding a high hazard still indicates that there is a 2% probability of not finding high hazard. However, if no high hazard is found upon inspection of the site, a safety officer is likely to perceive the RAP 2.0 probability output to have misled him/her. Our interviews indicate that these outcomes come with their own set of implications.

Visiting a low hazard site due to a high RAP 2.0 probability output can be seen as a waste of valuable and limited human resources. It can further frustrate safety officers, primary users of RAP 2.0, and undermine their trust in the RAP 2.0 output. Our interviews suggest that the number of RAP 1.0 outputs that have misled safety officers (e.g., a task categorized as Mandatory when there is no or only low hazard to be found) could lead to acceptance issues of RAP 2.0, as BCSA is likely to confront the legacy of performance of RAP 1.0 outputs, which is considered relatively poor. Due to this precedence, the development team may face organizational barriers in attempting to build trust relationship between safety officers (and senior safety officers) and RAP 2.0.

FALSE POSITIVE VS. FALSE NEGATIVE

There are two types of outputs that can be used to measure the performance of systems such as RAP 2.0. Statisticians call them false positive and false negative.

In the case of RAP 2.0, a false positive refers to RAP 2.0 output suggesting a high probability of finding a high hazard, when in fact there are none or only low hazard to be found on site.

Similarly, if RAP 2.0 suggest that there is a low probability of finding a high hazard when in fact there is a high hazard to be found on site, such output would be called a false negative.

It is also important to note that the performance of data-driven algorithms, such as RAP 2.0, heavily depend on the quality of data used to develop the system. That is, the performance of RAP 2.0 and the usefulness of its output is linked to the quality of data safety officers provide to the system. Deployment of RAP 2.0, therefore, requires safety officers to understand the role they take on as one of the sources of input data.

Recommendation

B 4.20 Monitor the predictive performance of RAP 2.0 over time and determine how much risk BCSA is willing to accept for which type of misleading (false positive/negative) findings from the system. While one type of misleading output (false positives) can contribute to frustrations by safety officers and inefficient use of safety officers' time, the type of incidents that occur due to false negative outputs are associated with a varying amount of risk. For example, if RAP 2.0 output leads safety officers to waive an inspection of a suburban single-family dwelling resulting in BCSA to miss a high hazard condition in the home, this would pose a different level of risk than if RAP 2.0 leads safety officers to miss high hazard conditions in a public educational facility. Given that having false positives and negatives are inherent to RAP 2.0 by the probabilistic nature of the system, BCSA will need to map the false negatives to the level and type of risk they pose on BCSA and public safety according to factors such as asset and inspection type. Explicit decisions will need to be made to determine what level of risk BCSA is willing to accept and implement it into its operation and RAP 2.0 design. For inspections that require the most conservative approach to risk posed by RAP 2.0, additional policy decisions can be made to enforce certain inspections regardless of RAP 2.0 output. For example, currently, BCSA has a policy that enforces 100% inspections on certain type of permits (e.g., homeowners) regardless of RAP Score output, which provides a safety buffer against the performance of RAP 1.0. Taking the highly conservative approach for all inspections that does not pose high risk would drive BCSA's business closer to a 100% inspection model and away from the risk-based model it currently operates.

Items for Considerations:

B 4.21 Safety officers are one of the key sources of data that RAP 2.0 is designed to model. Predictive performance of a data-driven RAP 2.0 cannot improve if the quality of data the system is designed to model is poor. Therefore, it is imperative for the RAP development team to communicate this relationship between safety officer's data entry practices to RAP 2.0. In turn, it is important for safety officers to recognize their role in improving the performance of RAP 2.0 through entering of high quality data in their daily tasks.



Challenge 4.1.4.6: Burden of Knowledge

Discovery of new information can not only serve as an enabling mechanism for those who discover it, but it can also put the burden on them to use the information responsibly and appropriately. This burden of knowledge within an organization changes as an organization learns and grows. With the advancement in data analytics it is important to consider what various stakeholders need to do with the new insights provided by the algorithm. Currently, at BCSA, the development team is the first group to observe the RAP 2.0 output and confront possible new insights. The challenges lie in deciding who they decide to communicate the information to and determining what actions BCSA should take.

Acting on a piece of information that is discovered through the algorithm is especially challenging for those not involved in the discovery of the information. For example, if RAP 2.0 focuses on predicting a probability of noncompliance (rather than a technical failure, for example), the probability output would inherently put the burden of knowledge on those who are given access to the information, including safety officers. What should they do when confronted with the insight that RAP 2.0 is predicting a high probability of noncompliance for a contractor? Should a safety officer pre-emptively check up on the contractor? This would not be perceived kindly by the duty holders, especially if it further contributes to a cycle of discriminatory practices.

Items for Consideration:

- B 4.22 Identify the sensitivity of the information produced by RAP 2.0 and train target audience of the information on how to handle sensitive cases. Unlike outputs from RAP 2.0 designed to predict probability of technical failures of assets based on technical details (e.g., voltage and load), outputs from RAP 2.0 to predict noncompliant behaviours using behavioural and identifying factors (e.g., contractor address, field safety officer names) carries a highly sensitive information. Use of the latter type of RAP 2.0 output would require safety officers to be trained on how to handle high RAP 2.0 probability output as an information they can use in their workflow, and how to handle their interaction with duty holders associated with the high RAP 2.0 probability. Echoing earlier recommendations, a clear objective setting of RAP 2.0 can help foresee the type of burden of knowledge BCSA is likely to encounter.



Challenge 4.1.4.7 : Destructive cycles and automation bias

One of the inherent problems with machine learning algorithms is that the inclusion of any bias or manipulative information can have cyclic negative consequences for the organization. It is not to undermine the fact that people behave and make decisions based on their own set of biases and stereotypes. For example, safety officers may prefer inspecting sites that are closer to them than to travel to an inspection site far in a rural area. Individual safety officers may have discriminating stereotypes against specific geographical regions or contractors. RAP 2.0 can mitigate these effects by directing safety officers to areas that would otherwise be less attended to. However, RAP 2.0 can also pose new challenges. Based on our interviews we identified two key challenges that would be most relevant for BCSA and further development of RAP 2.0.

1. Depending on safety officer's understanding and perception of machine learning algorithms, it is possible that safety officers would be more inclined to assign an inspection to be of a high hazard (As-Found Hazard category of 3, 4, or 5) if they had made the decision to attend to the inspection because of a high RAP 2.0 output. This psychological bias (automation bias) can lead to a self-fulfilling prophecy that can act as an artificial amplifier of the performance of RAP 2.0 and the features that have been selected for its training.

2. BCSA's stakeholder engagement process includes having advisory boards for each of the technology sectors it oversees. However, the advisory boards tend to represent larger clients, since larger clients tend to have more resources to participate in these board meetings. If the influence from these larger companies is reflected in RAP, it could create a perception of undue influence. There can be biases built into the trained model due to the fact that larger companies have more equipment/facilities to be inspected and are therefore more likely to generate a larger number of samples within BCSA's database than smaller companies.

Items for Consideration:

D 4.23 Actively monitor potential biases in the data that could result in the abovementioned destructive cycles. Trusting the performance of RAP 2.0 with the potential types of bias in mind can direct the monitoring practice.

SUMMARY OF KEY CHALLENGES AND RECOMMENDATIONS

Issues of managing public perception may pose a challenge as RAP 2.0 becomes more integrated within safety officers' workflow. Based on our interviews and the cautionary tales observed in the field, we foresee two main challenges. First, BCSA's clients and the public might grow concerned about the use of advanced technology in the regulation of public safety. Independent from BCSA's operations, this could stem from a general distrust of advanced technologies and the public's perception of advanced technologies as a means to replace human labour and decision-making. Second, BCSA's clients might feel uncomfortable about the level of oversight RAP 2.0 would provide. As discussed above, we believe that both of these challenges can be minimized by addressing the issues of transparency mentioned above and making an explicit effort to engage BCSA's stakeholders in the development and implementation process.

In order to ensure that data analytics is implemented responsibly within an organization, it is critical to devise communication strategies and practices around how the new centralized knowledge base will be handled within the organization. These strategies should be developed in collaboration with all of the stakeholders that are impacted by the new set of knowledge. In addition, RAP 2.0 inputs, algorithms and outputs need to be continuously monitored for biases and potential problematic cases.

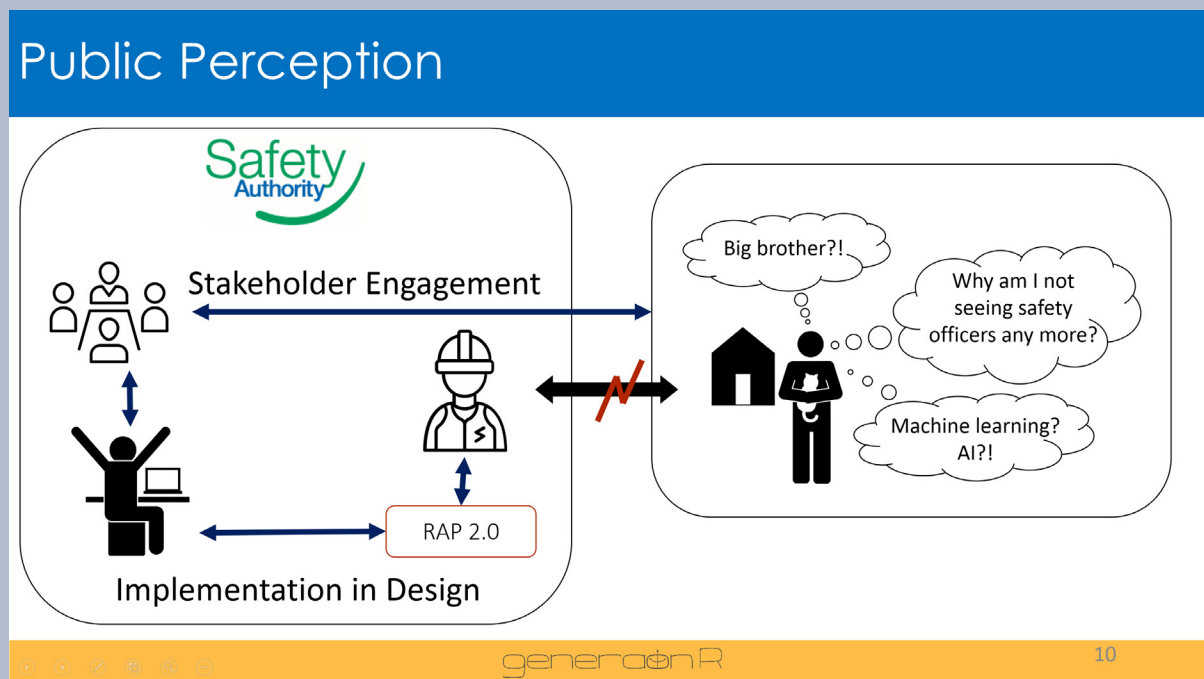


Figure 4.8 Potential challenges of public perception could be addressed by holding stakeholder engagement sessions and implementation or acknowledgement of feedback received from the engagement sessions.



Challenge 4.1.5.1: Perception of the Public towards Advanced Technology

The public perception of the use and implementation of AI and machine learning systems can vary drastically from one community, culture, country, to another. Some people have an optimistic outlook on these technological advancements, while others are pessimistic about how the technology will affect their autonomy, jobs and privacy. Negative perceptions about predictive algorithms could negatively interact with the public's perception of BCSA today.

For an extreme example, an asset owner might overly grow suspicious of a BCSA safety officer who is taking photos of their asset, regardless of whether BCSA peruses photos as part of RAP 2.0 training dataset. In another case, an asset owner might wonder why they are losing their face to face time with BCSA safety officers, and further question how well BCSA is maintaining safety oversight.

It is noteworthy that, as of now, there have been no sticking points with the people who have participated in the stakeholder engagement process held by BCSA. The general sense seems to be positive -- stakeholders seem to realize that technology is changing our lives and that it is accepted as an important part of societal progress. However, the stakeholder engagement representatives have communicated the public's desire to know what RAP is, how it impacts decision making at BCSA (at different levels) and how it affects them.

Items for Consideration:

- B** 4.24 Host RAP-focused stakeholder engagement sessions to identify the concerns and questions that various stakeholders have throughout the RAP 2.0 development process.
- D** 4.25 Ensure that concerns that are brought up during the engagement sessions are acknowledged and considered in the design and implementation process. It is possible that not all concerns or requests brought up by stakeholders can be addressed or implemented as a technical or policy solution. However, the expressed acknowledgement of these concerns can help mitigate development of public suspicion.

4.2 USE CASE 2: STRATEGIC DECISION MAKING

BCSA's senior leaders have expressed an interest in using outputs from RAP 2.0 to inform their decision making at the senior management level. Supplementing the issues discussed in Use Case 1, this section highlights issues that are unique to the use of RAP 2.0 output for supporting strategic and operational decisions.

Based on our interviews, it seems possible that various individuals at BCSA could use the RAP 2.0 output in their decision-making processes, including: senior safety officers; safety managers; regional business leaders; directors of policy; operations and legal; and the executive team. We distinguish this use case to be different from Use Case 1 in that the nature of the decisions (e.g., strategic, operational, policy decisions etc.) that RAP 2.0 can inform for the target users in Use Case 2 is different from the daily inspection decisions safety officers (target users of Use Case 1) would make.

SUMMARY OF KEY CHALLENGES AND RECOMMENDATIONS

Some of the key challenges with using RAP 2.0 output for strategic decision making are related to values of interpretability, transparency and trust. Our interviews suggest that the senior leaders have many objectives tied to the development and implementation of RAP 2.0. Interpreting RAP 2.0 output appropriately becomes more challenging if these objectives are not clearly defined and prioritized. As the senior leaders start to integrate RAP 2.0 outputs into their decision making, it will be important for the senior leaders to be transparent with BCSA's employees about how RAP 2.0 affects their decisions. Perceived or actual lack of transparency might result in an erosion of trust between stakeholder groups within BCSA. Finally, if the context and/or limitations of RAP 2.0 and its outputs are not recognized and duly acknowledged, the senior leaders may run the risk of over trusting RAP 2.0 outputs.

First, we recommend establishing a clear communication link between the development team and the senior leaders to ensure that RAP 2.0 output and system limitations are clearly articulated and acknowledged. Second, it will be imperative to have consistency between the stated objectives that RAP 2.0 is designed for and the context of senior leader decision making in order to avoid undesirable mission creep effects.

MISSION CREEP

Mission creep is a term that is used to describe the way that tools can be intended for one use then used for an increasing number of unintended uses that can be problematic if the tool isn't quite capable of those new uses.



Challenge 4.2.1.1: The objectives of the senior leaders and the interpretability of the RAP output within the strategic decision making process

Our interviews uncovered four clear objectives for using RAP in strategic decision making that were expressed by senior leaders:

1. RAP 2.0 output could identify factors that lead to higher probabilities of finding non-compliance and a higher As-Found Hazard: Identifying such factors can lead to productive strategic decisions around education, enforcement and certification. For example, RAP 2.0 output could show that the number of non-compliances recorded over time, and identifying information about FSRs, are some of the most promising factors related to finding As-Found Hazard categories 3, 4, or 5. In interpreting such a result, the development team may need to exercise caution in extracting any causal relationships – which are much harder relationships to establish than correlations – between individual field safety representatives, their contractors, and senior safety officers who have certified/licensed them. Making unjustified interpretations of the RAP 2.0 output can lead to the development of inappropriate company policies and programs.

2. RAP 2.0 output could be used to inform the public of major safety issues: Pursuing this objective would mean that the BCSA senior leaders would further extend and publicize their interpretations of RAP 2.0 output. This could have different consequences than just using the information to improve BCSA's internal processes. It could enhance public safety by informing the public but it could also negatively impact a group of contractors/stakeholders. These contractors and stakeholders could, in turn, question their confidence in how BCSA

is making these claims (interpretations of RAP 2.0 output). This objective is more explored in use case 3.

3. RAP 2.0 output could improve the efficiency of allocating human resources, particularly safety officers: RAP 2.0 output could allow the senior leaders to make decisions about where they should allocate their human resources, specifically safety officers. For example, the interpretation of the output could provide the justification for a decision to allocate more human resources to certain regions, technology or inspections.

4. RAP 2.0 output could support the risk-based model that BCSA uses for inspections: BCSA is the only regulator of public safety within British Columbia that is following a risk-based model as opposed to a 100% inspection model. The interpretation of the RAP output, whether it is probability of technical risk or probability of non-compliance, affects how well the company can support their overall approach to public safety.

Interpreting the RAP 2.0 output for each of the above objectives requires its own considerations and has its own implications. An interpretation of RAP 2.0 outputs for one of the objectives might not translate smoothly for the other objectives. For example, the objective of effectively allocating resource for physical inspections would motivate the leadership to maximize the time safety officers spend on tasks that necessitate their attention (e.g., physical inspection) and minimize the time spent on tasks that do not. However, having a policy for safety officers to attend to permits with highest RAP 2.0 probability outputs in the order of the probability values may not directly equate to an efficient nor optimized use of safety officers' time, especially if these high RAP 2.0 probability permits are often geographically spread apart.

Recommendations:

B 4.26 The senior leaders should engage and seek guidance from the development team in determining how, appropriately, to interpret RAP 2.0 probabilities. This could help to avoid making strategic decisions that unfairly discriminate against individuals or groups. Contextualizing the RAP 2.0 output is important as there could be many false assumptions about what a specific number represents, and about the limits of what that output can be interpreted as. Due to the nature of developing predictive algorithms, modeling processes for RAP 2.0 are likely to lead to the discovery of various factors that are highly correlated to “As-Found Hazard categories 3, 4, and 5”. These new discoveries can only be interpreted as correlations, rather than causations, unless the discoveries are explored further. The senior leaders’ efforts to subsequently explore possible causal relationships in detail in making strategic policy decisions can improve public safety and their role in BC as partners in safety.

Items for Consideration:

B 4.27 Review a prioritized list of objectives for developing RAP 2.0 and identify whether the objectives are better served by a separate data analytics program. This can help address senior leaders’ need to make informed decisions and providing tailored means to support the need, independent from the primary objective of RAP 2.0.

RAP 2.0. This would include a clear statement indicating the intended use of the data, which relates to a lay expression of the data analytics goals of BCSA, along with the option to opt out of the particular use of the information they provide.

**Challenge 4.2.1.2 Trust and Transparency**

It is important to consider the level of transparency on how the senior leaders uses RAP 2.0 output in their decision making. Transparent use of RAP 2.0 output could help safety officers and other employees understand objectives for RAP 2.0, beyond those that relate directly to their own work. It could also raise their awareness of conditions under which their work could be implicated by RAP 2.0. Currently, management uses their experience and expertise to make decisions. Without clear objective setting for using RAP 2.0 and transparent communication about the objectives, the introduction of RAP 2.0 could convey the appearance that BCSA leadership is replacing human experience with data-driven decision making, which could trigger an erosion of trust among BCSA employees.

Recommendations:

B 4.28 Consult the development team in setting objectives of using RAP 2.0 outputs for senior management decision making, and ensure that the performance of RAP 2.0 and the scope of what RAP 2.0 output represents are appropriate for the objective.

Items for Consideration:

B 4.29 Develop a comprehensive communication strategy so that the employees of BCSA are aware of, or have easy means to be informed of what RAP 2.0 is and how it integrates within the workflow of the senior leaders at BCSA.

SUMMARY OF KEY CHALLENGES AND RECOMMENDATIONS

In the foreseeable future, the issues discussed in use case 1 surrounding autonomy and jobs could be similar for other positions within BCSA (RAP 2.0 implications for the role of safety officers is discussed in use case 1). The two issues that the senior leaders should be made specifically aware of are the use of RAP 2.0 outputs in performance evaluation, and the impact of RAP 2.0 in the daily workflow of various employees. We recommend that the senior leaders discuss these issues pre-emptively, and in collaboration with other employees, to arrive at strategic decisions of whether, and how, RAP 2.0 might be integrated into performance evaluations and the daily workflow of those employees.



Challenge 4.2.2.1 : Evaluation of performance and effect on roles

The use of RAP 2.0 can allow the senior leaders to monitor and evaluate how various contractors, field safety representatives, safety officers, senior safety officers, regional business leaders and safety managers are performing in their respective positions. Any decision and processes that affect whether any of these individuals is promoted or demoted due to their performance needs to be carefully contextualized, validated and explained. Each of these individuals might alter their roles such that it optimizes their performance based on their understanding of how RAP 2.0 is being used in performance evaluations. This can have both positive and negative impacts for public safety, depending on what factors affect RAP 2.0 and how the RAP 2.0 output are used as an incentive for these individuals.

Items for Consideration:

B 4.30 Carry out a more detailed analysis of the potential impacts that use of RAP 2.0 output for performance evaluations could have on the organization. In particular, imprudently using RAP 2.0 for such purposes could have perverse effects on some BCSA values, such as having employees with high job satisfaction or, ultimately, public safety.



Challenge 4.2.2.2: Changing nature of jobs

The roles of each one of the individuals in the pipeline of services that BCSA offers could change with a more extensive use of RAP. This is most prevalent in the case of safety officers, which we described extensively in use case 1. Other roles could also change. For example, the role of the development team could shift towards finding factors that are more predictive of risk/hazard. The role of the RBL could shift to require including the RAP 2.0 probability of the region when they are making business decisions. The senior leaders should oversee if, and how, various other roles change and whether these changes hinder or facilitate the objectives that the specific positions serve within the organization.

Items for Consideration:

B 4.31 Work collaboratively with various employees to refine the definition of their role as RAP 2.0 is becoming more relevant in their work process. Consider engaging the development team in these discussions so that RAP can be improved with these employees' needs in mind.

4.3 USE CASE 3: PUBLIC REPORTING OF THE RAP 2.0 OUTPUT

BCSA's senior leaders expressed an interest in exploring the case of publicly reporting the RAP 2.0 output. Based on our interviews and discussions surrounding this specific use case, we believe that public reporting could take few different forms. Urgent public reporting could be accomplished through the mass media, if disclosure of specific problematic trends are observed from RAP 2.0. An annual report on RAP 2.0 probability trends for the year could be produced and made available on BCSA's website. Public reporting could take on a more individualized format, such as a safety officer showing the RAP 2.0 output to duty holders. Each one of these communication choices has its own ethical ramifications. Based on our analysis there are two key challenges that are unique to publicly reporting RAP 2.0 outputs.

First, a challenge arises when the reporting is done outside of a well-defined objective. Without the accompaniment of a clear message about the nature of RAP 2.0 outputs and the objective it serves to BCSA and the public, making public RAP 2.0 outputs that can be traced back to individuals and specific organizations could lead to dangerous misinterpretations by the public. The appropriate combination of the medium of reporting, target audience, and the information provided to the audience would be different for different objectives. For instance, providing information about the state of safety to individuals who cannot act on the information change can lead to unnecessary frustration. In pursuing this use case, we recommend clearly identifying the objectives and target audience in order to ensure the information can be used for the purposes intended. Further, we recommend identifying possible value conflicts that can arise from the specific reporting of RAP 2.0 results, keeping in mind that the possibilities of mission creep (see Use Case 2 for a definition).

Second, reporting predictions from RAP 2.0 output will have a different effect if the predictions become publicly available. For example, duty holders who are marked having high hazard assets are likely to demand explanation from BC Safety Authority, and perceive the practices of BCSA to be unfair or untrustworthy. This perception would be even stronger if no serious hazards were found upon inspections. BCSA would benefit from having communication strategies to handle such public backlash.

Analysing the public reporting of RAP 2.0 output will require a more clearly defined objective and further interviews with external stakeholders outside of those we have interviewed within the scope of this project.