NSW GOVERNMENT

Ethical use of artificial intelligence in the workplace final report

Centre for WHS

THE UNIVERSITY of ADELAIDE

Flinders UNIVERSITY

This report and the work it describes were funded through the Workers Compensation Operational Fund. Its contents, including any opinions and/or conclusions expressed, are those of the authors alone and does not necessarily reflect SafeWork NSW policy.

The Changing World of Work: Ethical Deployment of AI in the Workplace (Final Report)

Prepared by:

Dr Andreas Cebulla[1]

Dr Zygmunt Szpak[2]

Dr Genevieve Knight[3]

Dr Catherine Howell[4]

Dr Sazzad Hussain[5]

June, 2021

[1] Australian Industrial Transformation Institute, Flinders University

[2] Australian Institute for Machine Learning, University of Adelaide

[3] South Australian Centre for Economic Studies, University of Adelaide

[4] Independent Consultant / University of Adelaide

[5] Centre for Work Health and Safety, NSW Department of Customer Service

Executive Summary

Background and Aims

More and more Australian workers are experiencing the introduction of Artificial Intelligence (AI) in the workplace, affecting role design, task allocation, time management, organisational structure, and communication. While AI can change the work environment significantly, there is limited research that has examined the impact of AI on Work Health and Safety (WHS). There are gaps in the understanding of potential risks and hazards to workers, as well as a lack of resources for assessing and mitigating WHS risks in using AI in the workplace. This research sets out to address the key gaps with the following aims:

1. To understand the potential WHS risks related to AI use in the workplace.
2. To understand the current WHS management practices of organisations that had recently introduced or were in the process of introducing AI in the workplace.
3. To develop a novel risk assessment tool (i.e., an AI WHS Scorecard) to assist businesses in identifying and assessing WHS risks related to the use of AI in the workplace.

Method

This research adopted a qualitative and practice-led approach, and collected evidence by completing a literature review and conducting a series of consultations with AI experts, WHS professionals, regulators and policymakers, representatives from organisations adopting or having adopted AI, and others with knowledge in the field. In developing the risk assessment tool, the research incorporated feedback from the consultations and made continuous refinements for improvement. The specific components of this research included the following:

- Consultations with experts and stakeholders through interviews and two online workshops to gauge awareness of and concern about AI effects on WHS and further develop the risk assessment tool based on the feedback.
- Consultations with representatives from organisations using or planning to use AI to understand the processes leading to AI use in a workplace, the WHS management practices, and the utility of the proposed risk assessment tool.
- Consultation with WHS inspectors to collect feedback on the scorecard from a WHS practitioners' perspective.

Results and Discussion

Potential WHS risks related to AI use in the workplace

The findings suggested that harm from AI use was more likely to impact workers psychologically than physically. However, workers' physical safety and health might still be impacted if the use of AI influences the intensification of workflows or surveillance in the workplace, causing workers to accelerate their pace of work and thus creating new hazards. The consultations also highlighted expectations that AI would partially automate tedious and repetitive tasks; therefore, impacted employees would have to adapt to new workflows and learn how best to integrate AI solutions into their daily routines.

AI would also be used for work augmentation. That is, employees would improve the quality of their work owing to features and functionalities provided by AI. AI was especially likely to cause deep changes to how organisations schedule or allocate workloads for their employees. Thus, AI capabilities are starting to take over from traditional managerial tasks, and the consultations highlighted concerns that AI tools might create barriers between workers and managers. This may then challenge WHS, which requires effective communication between workers and managers.

WHS management practices

Little evidence was found of organisations taking strategic approaches to anticipate the impacts of AI on workplaces beyond the intended process or product change.

Potentially far-reaching organisational implications of AI were acknowledged, resulting in new data-sharing arrangements, new job descriptions and the creation of new positions. However, potentially harmful implications of AI to WHS were more typically late considerations, commonly raised at the point of AI use rather than at the design stage.

Proposed risk assessment tool – the AI WHS Scorecard

The proposed risk assessment tool, the AI WHS Scorecard, integrated principles of the ethical use of AI with Safe Work Australia's WHS concepts of hazards and risks.

The initial draft of the scorecard combined two existing frameworks that both intend to guide the design, development, and implementation of AI: (i) the Australian Government's AI Ethics Framework and (ii) the AI Canvas. The AI Ethics Framework covers eight Ethics Principles designed to encourage AI use for the benefit of Australian society. The AI Canvas, originally developed by researchers at the University of Toronto, Canada, identifies seven core stages of an AI system's design and development lifecycle and its implementation.

Based on participants' feedback, the scorecard was shaped around a simplified framework describing AI Ethics Principles that evolved from eight to three broad categories (Human Condition, Worker Safety, Oversight). We also aggregated the seven stages of the AI Canvas into three higher-level steps (Ideation, Development, Application).

We then incorporated Safe Work Australia's concept of the Characteristics of Work from their "Principles of Good Work Design", and WHS hazards and risks into the scorecard. Finally, we introduced a risk rating to assist organisations in determining the possible likelihood and consequences of WHS risks in the use of AI. We also prepared an AI WHS Protocol to accompany the scorecard explaining its roots and providing guidelines for its use.

Conclusion

The outcome of this research contributes to a better understanding of AI use in the workplace and its impact on workers. We developed an evidence-based risk assessment tool (i.e., AI WHS Scorecard) and accompanying Protocol, which can help organisations adopt AI with a WHS focus.

Table of Contents

List of Tables

List of Figures

Introduction

Artificial Intelligence (AI) has the potential to change the landscape of work fundamentally. In the context of this research, AI refers to software systems or machines that (i) adapt and learn by identifying patterns as they encounter new information and (ii) use these patterns to make predictions or recommendations. For example, they may predict worker performance based on activity and behavioural patterns, identify the most diligent workers for given tasks or recommend workflow optimisation. There is emerging evidence that the use of AI is driving changes in workplaces across multiple domains, including worker role design, organisational structures, and management strategy (Safe Work Australia, undated; Griffin et al., 2019). Recruitment and retention, task allocation, time management, how workers communicate with one another and with managers, and how workers are incentivised, supported and rewarded in the performance of their jobs are all impacted by the introduction of AI (O'Neill, 2016).

With the emerging growth of AI use cases and its adoption (Perrault et al., 2019; Hajkowicz et al., 2019), the discussion around AI use has mainly focused on its potential economic benefits, for example, increases in productivity and cost or time savings. Although there is an emerging emphasis on the impact of AI solutions on the general population (i.e., consumers) and its ethical implications, little attention has been given to the impact AI systems might have in the workplace and on the health and safety of workers. In fact, there has been little research examining work health and safety (WHS) risks associated with AI implementation in businesses, and a lack of resources and tools for assessing and mitigating WHS risks.

This research started to fill this gap by (i) investigating the perceptions of AI use and its impact on the workplace and (ii) developing a risk assessment tool for AI adoption with a WHS focus. The output of this research aims to help businesses adopt AI solutions while championing the health and safety of workers. Specifically, the research contributes to the following:

1. The understanding of WHS risks associated with the use of AI in the workplace.
2. The understanding of the current WHS management practices of organisations that had recently introduced or were in the process of introducing AI in the workplace.
3. A risk assessment tool (i.e., AI WHS Scorecard) to assist businesses in assessing the WHS risks related to the use of an AI systems in the workplace.

This report presents the research undertaken to understand the knowledge gaps and develop the AI WHS Scorecard. It starts with an overview of the relevant AI, ethics, and WHS background, and reviews existing resources, along with some of the emerging research on the topic of AI impact on workers. We then outline our research methods, which involved a series of consultations (qualitative interviews and workshops). This is followed by the presentation of the findings from each of the methodological components of the research, including the AI WHS Scorecard that has evolved iteratively from the research.

We close with a summary and direct readers to relevant supporting materials included in our appendices.

Background and Rationale

This chapter sets out the empirical background based on the literature review to present the current WHS status and gaps in adopting AI in the workplace, as well as the rationale to address key gaps in safeguarding the anticipated risks of AI use on workers. These topics are structured as follows.

1. AI and potential WHS risks of its use.
2. Current WHS management practices.
3. Existing resources for risk assessment and mitigation.
4. Rationale for developing a novel AI risk assessment tool.

AI and potential WHS risks

AI can pose a multiplicity of risks, ranging from threats to personal data security and privacy owing to the increased use of big data (Dawson et al., 2019), to societal vulnerability to an unprecedented use of AI designed machinery (Devitt et al., 2020). In a recent review, global leaders in AI research, business and policy making expressed concern that AI will profoundly affect how we live and work (Pew Research Center, 2018). In particular, they anticipated that with the use of AI expanding and machine algorithms determining what and how we do things, individuals might lose control over their lives or jobs and experience a reduction in their cognitive, social and survival skills. The novelty of AI means that its risks and their impact on humans remain hard to foresee and categorise. This uncertainty is particularly the case in a workplace context, where the phenomenon is only beginning to be explored. WHS, in contrast, already has a historically grown understanding of hazards and risk, which seeks the protection of workers from physical injury and psychological harm. A similar purpose of harm prevention may need to be applied to the use of AI, the associated WHS hazards and risks of which have yet to be comprehensively identified.

Technological innovations are associated with operating environments with elevated levels of uncertainty and the possibility of hazards to workers or the wider public emerging only after their implementation or launch. AI-based systems are agent-like and replace human actors in certain domains, aiming to make predictions or recommendations. AI shifts the nature of work in the workplace by minimising human involvement and oversight of traditional operational processes or systems. Moreover, the human interaction with AI is no longer a simple, mechanistic model of an operator inputting data into a system or machine, which then processes the information and generates an expected output. The interaction is more complicated because an AI-system evolves over time as it adapts and learns through an ongoing process of identifying patterns that continually shape its predictions or recommendations. This evolutionary nature makes AI systems significantly less transparent and explainable than traditional systems or machines.

Loss of transparency and lack of explainability of AI-based predictions or recommendations may cause anxiety and stress to workers, as does AI use for performance monitoring, surveillance and tracking of employees (Moore, 2018; Horton et al., 2018). On the job, workers may experience AI as competence-enhancing, but AI may also be competence-destroying, for instance, as a result of task automation (Paschen et al., 2020). Workers may thus face

deskilling, a loss of control over work schedules and tasks, or redundancy (e.g., IEEE, 2016; OECD, 2019; Australian Human Rights Commission, 2019). In emerging evidence, workers report that they have little or no say in how business processes and employee roles are changed through AI and that there is insufficient communication between employees and employers about the complex technological changes often associated with AI (Commission on Workers and Technology, 2019).

Kellogg et al. (2019) explored in considerable detail some of the ways that AI can affect workers. They suggest that AI enables organisations to direct, evaluate, and discipline workers. AI can direct workers by restricting and recommending information or actions. For example, an AI could generate a script that a call centre employee must follow, or an AI could automatically recommend responses to an email that a client sent. Scripting call-centre workers may result in demotivation and absenteeism of employees. Employers can use AI to evaluate how workers perform tasks and assess their activity and behavioural patterns, and determine which employees are best suited for different tasks that a workforce needs to complete. Invasive surveillance of worker performance through remote application of AI can increase stress. Loss of autonomy can lessen employees' sense of enjoyment and accomplishment in their work. Management can use AI to find opportunities for optimising workflows and identify the most diligent employees, and even discipline workers. For instance, AI can discover erratic and dangerous driving behaviour in taxi drivers or detect safety violations such as not wearing appropriate safety attire when entering restricted areas. AI applications such as these can be used for positive purposes, but the effects of AI on workers can still be negative and less apparent.

To date, the literature suggests that harm from AI systems seems more likely to impact workers psychologically than physically. Workers' physical health may nonetheless be impacted if the intensification of workflows through AI or surveillance through AI induces them to accelerate their pace of work, creating new physical safety and health hazards (Moore, 2018). It has been argued that maintaining worker autonomy over the execution of their tasks may be critical to sustaining physical and psychological health in the workplace. Hence, the design of digital solutions such as AI must consider these issues (Calvo, 2020).

WHS management practices

The rapid ascent of ICT and AI creates new challenges for existing WHS risk assessment models and current Codes of Practices for managing workplace risks. An essential element to consider here is how technology may involve a transformation of work roles and functions. The creation of new workflows through ICT may entail important elements of work redesign for human actors (i.e., workers). The Safe Work Australia Handbook, "Principles of Good Work Design", notes that:

> *In most workplaces the information and communication technology (ICT) systems are an integral part of all business operations. In practice these are often the main drivers of work changes but are commonly overlooked as sources of workplace risks.* (Safe Work Australia, undated: 15).

Managing the hazards and risks associated with AI use means scrutinising the decisions that are made in the workplace based on AI outputs (cp. Autor et al., 2020). Besides accommodating a level of human autonomy, AI risk management may also require enhancing the capacity of workers through digital literacy and feedback mechanisms that help them

to manage new technologies (Donati, 2020). The empirical literature suggests that these concerns may, for now, not be well articulated and recognised in workplaces. A survey by McKinsey Digital (2020) detailed a range of AI risks acknowledged by commercial businesses, led by higher-level organisational risks, such as cybersecurity, regulatory compliance and personal/individual privacy. Between 40 per cent and 60 per cent of the organisations surveyed by McKinsey Digital expressed concern about each of these AI risks. In contrast, AI risks that directly affect workplaces appeared much less of a concern. Thus, only 31 per cent of organisations surveyed by McKinsey Digital expected AI to lead to workforce displacement and only 19 per cent anticipated physical safety threats arising from AI. Moreover, McKinsey Digital (2020: 9, emphasis added) concluded that only "a minority of companies recognise many of the risks of AI use, and *fewer are working to reduce the risks*".

In the Australian employment context, the Fair Work Act (2009) creates specific rights and obligations for employers and workers, designed to protect employees' workplace rights. In the wording of the Act, it is intended "to provide a balanced framework for cooperative and productive workplace relations that promotes national economic prosperity and social inclusion for all Australians" (Australian Government, 2009: Division 2). The Act emphasises workplace relations laws that are "fair, relevant, and enforceable". Its emphasis on "fairness" is an underlying ethical principle that is mirrored in the Work Health and Safety Act (2011), which also speaks of "fairness" specifically about "providing for fair and effective workplace representation, consultation, co-operation and issue resolution in relation to work health and safety" (Government of Australia, 2011: Division 2). Moreover, government agencies such as Safe Work Australia, act as regulatory bodies to ensure WHS and provide guidelines to business.

The Safe Work Australia "Principles of Good Work Design" specifies under Principle 4 that:

> *"Good work design addresses physical, biomechanical, cognitive and psychosocial characteristics of work, together with the needs and capabilities of the people involved" (Safe Work Australia, undated: 9).*

Change and innovation in work practice potentially affect each of these characteristics of work (Figure 1), individually or in combination, and thus call for a systematic approach to hazard management. Each characteristic of work, in turn, is subject to specific hazards and risks.

Figure 1*: Key Characteristics of Work. Adapted from Safe Work Australia, undated: 9.*

WHS management schemes tend to be best suited for, and to date mostly focused on, addressing situations where there is a straightforward connection between the cause of a specific safety hazard and its resolution. For example, to manage the risk of a robot colliding with a human, the robot could be placed behind a fenced area. In general, WHS management schemes tend to favour the regulation of physical safety-related hazards. They are less well-suited for scenarios where the hazard is ambiguous and its resolution complicated and multi-faceted, as is the case with AI in the workplace.

Existing resources - guidelines, frameworks and tools

This section presents some existing resources found in the literature to inform the design of an AI WHS risk assessment tool, drawing on: (i) AI ethics and principles, (ii) AI implementation strategy, and (iii) generic AI risk assessments.

AI ethics and principles

There is limited research with a specific focus on the WHS aspects of AI and also a lack of guidance on how to manage WHS in workplaces increasingly adopting AI. However, there is continuing discussion of the *ethics* of AI, which provides avenues for understanding the potential WHS hazards associated with AI.

In response to widespread concerns about the potential for negative impacts of AI on society, guidelines for ethical AI have been developed around the world. Some AI ethics guidelines initiatives have been government-led (for example, Australia, Canada, and Singapore), while others have been industry-led (for example, Microsoft, Google, and the Open

Data Institute). Hagendorff's (2020) recent review of AI ethics guidelines identified 22 examples. The legal and regulatory status of these guidelines differed by jurisdiction, although generally their adoption was optional. This led Hagendorff to assert that AI ethics "lacks mechanisms to reinforce its own normative claims" (Hagendorff, 2020: 99). Nevertheless, the creation of national and international ethics guidelines for AI offers a framework to guide 'good work design' where AI is involved.

Australia is a signatory to the Organisation for Economic Co-operation and Development's (OECD) "Principles on AI", endorsed by 42 countries in 2019 and subsequently adopted by the G20 (OECD, 2019). The OECD Principles on AI are as follows:

- *AI should benefit people and the planet by driving inclusive growth, sustainable development and wellbeing.*
- *AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society.*
- *There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them.*
- *AI systems must function in a robust, secure and safe way throughout their life cycles and potential risks should be continually assessed and managed.*
- *Organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles.*

Source: OECD, 2019

In Australia, the Commonwealth Scientific and Industrial Research Organisation (CSIRO)/data61's Strategic Insight team, in partnership with the Australian Government Department for Innovation, Industry and Science (now: Department of Industry, Science, Energy and Resources – DISER), and the Office of the Queensland Chief Entrepreneur, led a project to develop an AI Roadmap and Ethics Framework under the banner of "Building Australia's artificial intelligence capability". The framework was published in April 2019 (Dawson et al., 2019), and has subsequently been adopted by Federal Government (DISER, undated) and State and Territory governments, including the NSW Government (NSW Government, 2019).

The DISER (undated) statement defines AI ethics principles as aspiring to ensure or further:

- *"Human, social and environmental wellbeing: Throughout their lifecycle, AI systems should benefit individuals, society and the environment.*
- *Human-centred values: Throughout their lifecycle, AI systems should respect human rights, diversity, and the autonomy of individuals.*

- *Fairness: Throughout their lifecycle, AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups.*

- *Privacy protection and security: Throughout their lifecycle, AI systems should respect and uphold privacy rights and data protection, and ensure the security of data.*

- *Reliability and safety: Throughout their lifecycle, AI systems should reliably operate in accordance with their intended purpose.*

- *Transparency and explainability: There should be transparency and responsible disclosure to ensure people know when they are being significantly impacted by an AI system, and can find out when an AI system is engaging with them.*

- *Contestability: When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or output of the AI system.*

- *Accountability: Those responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled."*

Source: DISER (undated)

The Australian national AI ethics framework, like the OECD's, is a framework that helps one broadly to explore the extent to which AI may affect individual wellbeing, values and rights. But the ethical principles at their core are abstract, and in their current form, not immediately suitable for assessing workforce WHS concretely.

AI implementation strategy

We broadened the scope of our review beyond tools designed specifically for measuring ethical or WHS AI risks to tools for scoping AI design. Agrawal, Gans and Goldfarb (2018a) developed one such tool called the AI Canvas. The AI Canvas is a practical decision support tool for businesses and organisations considering using AI. Developed by a team of researchers at the University of Toronto, Canada, its purpose is to help business leaders and managers understand whether adopting AI will enable them to achieve their strategic goals. It does so by mapping the processes to follow and questions to ask when deciding on the utility, design and operation of AI. The AI Canvas is based on the researchers' experience working with AI entrepreneurs and helping to seed successful AI start-ups in their business incubation lab (Agrawal et al., 2018a).

The idea of a canvas as a tool for mapping the various stages of an IT development project is not new. What is specific to Agrawal, Gans, and Goldfarb's AI Canvas is their economic understanding of AI as a *prediction machine*. The central insight of their work is that AI can improve decision-making under uncertainty, by enabling better and cost-effective predictions. At the same time, it also increases the value of judgment to an organisation: that is, understanding whether and in what circumstances predicted outcomes might deliver a reward or profit (Agrawal et al., 2018b). Together, these predictive capabilities can provide a strategic advantage to organisations.

The AI Canvas proposes a set of seven categories of questions that decision-makers would need to ask themselves to determine whether adopting AI will advance their overall strategy. The seven categories and associated questions are:

1. Prediction: what does the AI need to predict?
2. Judgment: how do we value correct versus incorrect predictions?
3. Action: how do the predictions affect what we do?
4. Outcome: how do we measure the performance of the AI?
5. Input: what data does the AI need for deployment?
6. Training: what information does the AI need for training?
7. Feedback: how can we use outcomes to improve the AI continually?

Responding to these questions helps to clarify the purpose of a proposed AI system. It gives an organisation an overall picture of capabilities and potential gaps concerning their AI strategy, resources and ambitions. It enables a preliminary dive into data issues, which are core to any organisation grappling with the potential of AI (e.g., what data are needed for a particular purpose, where data come from in the organisation, what the data lifecycle is, who monitors and evaluates data quality).

The AI Canvas summarises a proposed AI system design. It identifies its core components/stages, but it does not answer the question of whether the proposed system aligns with organisational values, ethics, and with WHS principles. Thus, it does not cover whether a proposed AI system is fair, ethical, or safe for workers and users. The AI Canvas, in its current form, is focused on the potential of a proposed AI system and its *technical* underpinnings. It is less able to identify the human factors necessary to its functioning or to evaluate the context of organisational and human relations in which an AI system is used. The AI Canvas's "Action" category comes closest to considering the context of how using AI would affect work practices. But it does not explicitly raise the question of a system's impact on workers or work roles, and nor does it enable a risk assessment of associated WHS issues.

AI risk assessment

Whilst our literature review did not identify examples of dedicated AI risk assessment tools for WHS, some generic instruments for assessing fair and ethical AI exist, such as the Canadian Algorithmic Impact Assessment, or AIA (Government of Canada, undated) and Mantelero's Human Rights, Ethical or Social Impact Assessment (Mantelero, 2018). Mantelero's Human Rights, Ethical or Social Impact Assessment was designed for use in the European context and focuses on data protection and the ethical use of data. Data protection and the rights of European citizens to access and manage their data have been a strong focus in European public life and lawmaking, with a data protection package adopted in 2016. It does not specifically include WHS or worker safety.

In the Canadian context, the AIA is available as an online questionnaire. It was designed principally for use by organisations tendering for government-funded work, particularly public service provision. Two advantages of the AIA

are that it is designed around a scoring logic that includes mitigation measures adopted by the completing organisation (that is, specific actions taken to mitigate risk), and it delivers a score to users on completion. The AIA does include specific questions about the health and wellbeing of individuals or communities, but these questions are focused on end-users, not on workers using AI as part of their jobs.

Rationale – key gaps and conceptualising a risk assessment tool

Key findings from reviewing the literature suggest that awareness of the risks and potentials of AI in society and the economy is not matched by a similar understanding of the effect that AI may have on workers and WHS. Notably, we found no WHS tools ready to observe, address and manage AI risks in the workplace.

In assimilating the existing resources, an AI risk assessment tool with a focus on WHS was seen to be feasible by incorporating the key concepts – ethical principles, implementation strategy, and WHS principles. This research has built on the original AI Canvas as a tool for understanding the strategic processes of implementing AI in a workplace. It has mapped AI Ethics Principles onto the AI Canvas to explore which, if any, AI Ethics Principles may come into play at each of the AI Canvas's stages. Just as the AI Canvas has been used as a tool for understanding AI implementation processes, the AI Ethics Principles have served as a lens for capturing the complexity and range of risks that may be associated with AI in a workplace.

Ethics principles aim to help to make behavioural choices that are right and acceptable within a shared social or cultural context. Whilst ethics principles are formulated at an abstract level, their violation can nonetheless have concrete WHS effects on those using or otherwise exposed to the use of AI in a workplace. This relationship to WHS of AI uses is conceptually and empirically established in this study by linking the AI Ethics Principles to Safe Work Australia's Characteristics of Work framework and associated workplace hazards and risks (Figure 2).

The two ethics principles of "privacy protection and security", and "reliability and safety" arguably resonate most clearly with Safe Work Australia's principles of good work design. The principles of good work design urge that attention be given to physical, biomechanical, cognitive and psychosocial characteristics of work to avoid or minimise risks of harm to workers. These harms may come about as the result of one or more workplace hazards. This research set out to establish connections between the eight AI Ethics Principles and Safe Work Australia's list of workplace hazards, and in adopting the AI Canvas.

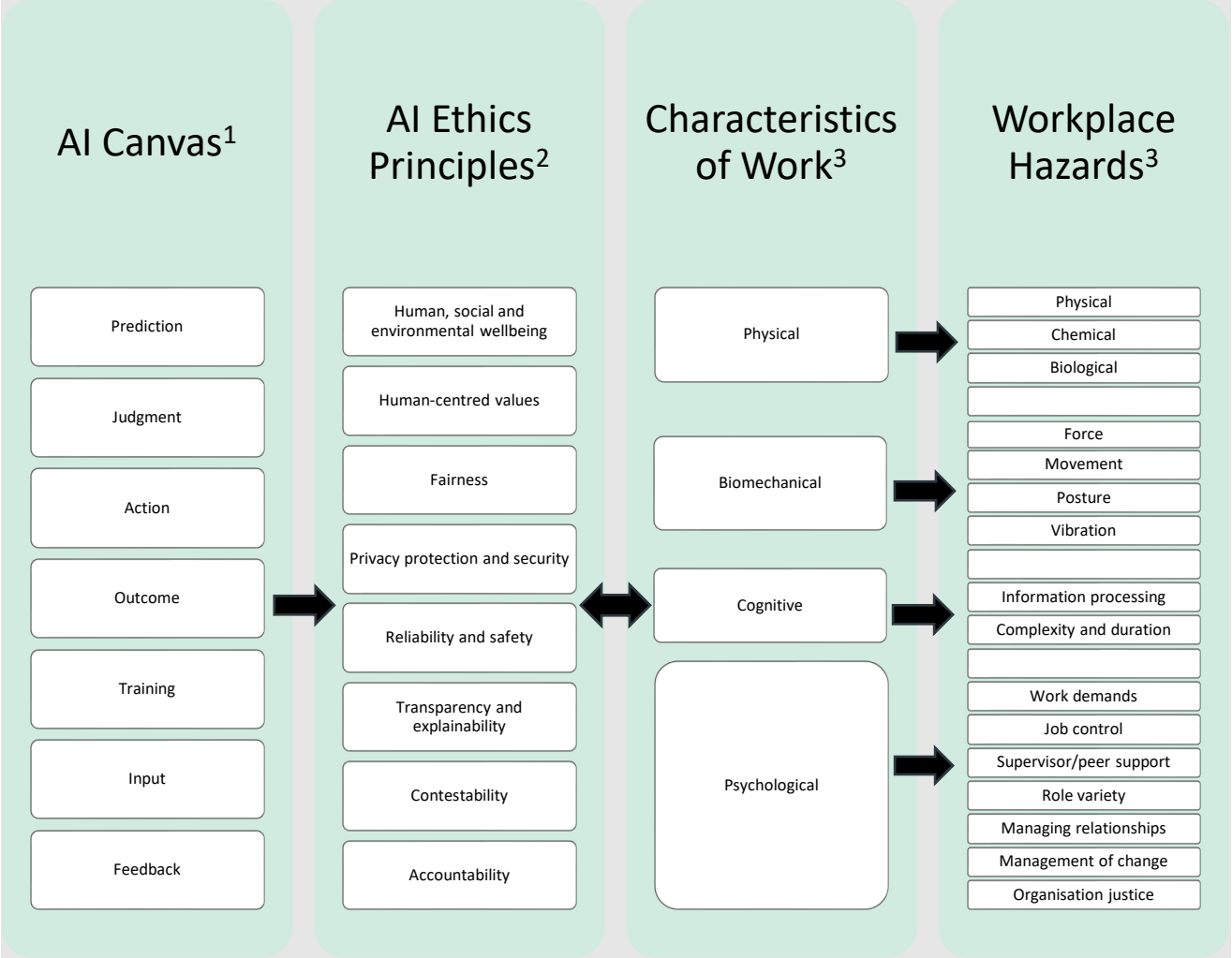

Figure 2: Conceptual integration of AI Canvas, AI Ethics Principles and Safe Work Characteristics. Adapted from [1]Agrawal et al., 2018a; [2]DISER, undated; [3]Safe Work Australia, undated.

Based on these considerations, an initial AI WHS Scorecard was developed (Scorecard v1.0), combining the seven AI Canvas dimensions with the eight DISER AI ethics principles in the form of a matrix (Table 1). This version of the scorecard (v1.0) was used at the starting point for consultations in this research.

Table 1: The original AI WHS Scorecard draft (Scorecard v1.0).

	AI Ethics Principles							
AI Canvas	**Human, social and environmental wellbeing**	**Human-centred values**	**Fairness**	**Privacy protection and security**	**Reliability and safety**	**Transparency and explainability**	**Contestability**	**Accountability**
Prediction *Judgement* *Action* *Outcome* *Training* *Input* *Feedback*								

Method

This section describes the main components of the research methods for the consultations, which have included:

- Phase 1: expert and stakeholder interviews, and two online workshops.
- Phase 2: case study interviews with organisations using or planning to use AI, along with further expert interviews.
- Phase 3: an online consultation of WHS inspectors.

Prior to commencing the research, all phases of the fieldwork were reviewed and approved by the University of Adelaide Human Research Ethics Committee (application number H-2020-212).

Background literature review

To gain an overview of current research and best practice in the areas of AI, WHS, and risk management, we conducted a qualitative literature review. This created an evidence-based foundation for the design of the AI WHS Scorecard. The review searched databases including Google, Google Scholar, ProQuest, Harvard Business Review, and Business source premier via EBSCO. The search terms used included: “artificial intelligence”, “AI”, “decision tools”, “risk management”, “risk assessment”, “risk matrix”, “balanced scorecard”, “ethics”, “business process improvement”, “health and safety”, “workplace”, and “wellbeing”, and applied the Boolean operators “and” and “or”. Articles retrieved were individually reviewed for quality and impact factors, including publication source, author affiliation/institution, country of origin, and number of citations achieved. Reference lists of selected articles were checked, and additional hand searches of key journals were conducted.

The review specifically included grey literature in the form of online corporate reports and advisory documents published by IT and management consulting firms, including Microsoft, Cisco, Deloitte, Accenture, and Ernst and Young. These firms have been prominent amongst those issuing high-level guidance documents on the commercial use of AI in parallel to, and sometimes in advance of, public regulatory and legislative reform. Work undertaken and published by CSIRO/Data 61, the Australian Human Rights Commission, the UK Commission on Workers and Technology, the Canadian Government Directive on Automated Decision Making, and the Singapore Government Model AI Governance Framework formed an additional distinct group of sources.

About 250 items of interest were identified; over 150 books, journal articles, monographs and news reports were reviewed in detail.

Phase 1: surveying the AI landscape

The first phase of consultation aimed to gather feedback on the proposed structure and content of the AI WHS Scorecard as well as its layout and presentation. It also sought to understand stakeholder perspectives on how AI is currently being adopted in Australian businesses and organisations, and what, if any, issues this raises for WHS. Phase 1 consisted of a series of qualitative interviews with key stakeholders for AI, and two public workshops. Across Phase

1, the research team engaged with individuals recognised for their expertise in AI matters, and/or their experience working with AI in industry or government organisations. They included academics and other professional or commercial experts using or developing AI in specialist subjects (e.g., health, engineering, computer science), ethicists (in academia and specialist institutions), senior managers, directors, laboratory directors and data analyst working in advanced information technology or end users of AI (including in government or affiliated to AI networks); and WHS practitioners.

Interviews

Participants and recruitment process

The purpose of the interviews was to develop a broad, inclusive overview of current experiences of adopting AI in the workplace, focusing mainly on the knowledge and practices of managers, leaders, and experienced professionals. We were also interested in probing participants' level of awareness of the national AI Ethics Framework and their understanding of how potential ethical issues for AI could relate to, or result in, WHS issues and risks. Participants of the interviews were selected based on their informed perspective on AI and/or their experience with AI use in a business, public sector or academic environment. An initial list was compiled from internet search investigations, together with contacts suggested by the project team and the Centre for Work Health and Safety. A total of 83 individuals were identified as potential participants for interview: 47 were initially proposed by members of the project team (including members from the Centre for Work Health and Safety - CWHS) and a further 36 were identified over time, some by recommendation of individuals who had been approached and/or interviewed. Of this list, fourteen individuals were excluded because they either represented the same organisations or their expertise was found to be outside the scope of this study. Sixty-nine individuals were thus approached for interview by email or via their social media platform (Linkedin) where an email address was unavailable.

A total of 30 interviews were completed; the remaining individuals who had been approached declined their participation in this study. The spread of participants across employment sectors is shown in Table 2. Just over half (16) of participants were working in industry, with about one third coming from the government and WHS sectors (9).

Table 2: Participant interviews by sector.

Sectors	**Interviewed**
Academia	*2*
AI Professional Networks	*1*
Government	*5*
Industry	*16*
Research	*2*
Statutory body	*0*
WHS	*4*
Total	*30*

Interview format and process

Each participant was emailed a Participant Information Sheet and Consent Form. A copy of the draft AI WHS Scorecard (version 1.0; see Table 1) was later emailed upon receipt of consent to be interviewed. An interview topic guide was used to conduct semi-structured interviews (see Appendix A). Interviews lasted between 40 minutes and slightly more than an hour; and were conducted by phone or via video conferencing tools (e.g., Zoom, MS Teams). Fifteen participants agreed to their interview being audio or video recorded. Researchers prepared detailed notes of their interviews.

Workshops

The workshops aimed to reach a wider audience for exploratory discussions about designing a healthy and safe use of AI in the workplace, and for identifying and managing WHS risks arising from using AI. In addition to discussing the purpose, the content and the design of the initial AI WHS Scorecard, participants were asked to suggest examples of potential or known WHS risks that might fit the scorecard's matrix. The workshops thus provided an opportunity for a focused group discussion testing the level of interest in the research themes, and an initial validity check for the first draft of the AI WHS Scorecard.

Participants and recruitment process

Recruitment was undertaken through online promotions, with registrants invited to attend a one-hour online workshop. The workshops were advertised on the websites of the researchers' institutes and promoted by the research institutes and CWHS via social media platforms. There were no selection criteria for participation. In total, 32 registrations were recorded, of whom 22 attended the workshops. Consent to participate in the workshop was sought during the workshops' introduction, along with permission to record the sessions.

Workshop format and process

Each workshop started with a short presentation about the research objectives and an initial outline of AI in the context of workplaces. This was followed by two facilitated breakout sessions (i.e., group discussions) giving participants opportunity to comment on and discuss the AI WHS Scorecard draft (version 1.0). Participants were asked about how they saw the Australian AI Ethics Principles applying to the seven categories of the AI Canvas, and which principles they believed to be most relevant to actual or potential WHS risk dimensions of AI. This was intended to test and develop the project team's understanding of AI WHS risks that had been identified previously during the literature review and interviews. Other discussions included the structure, layout, and categories of the AI WHS Scorecard (see Appendix B for workshop topic guide).

Workshops were about an hour long, with break-out group discussions lasting for about 20 to 30 minutes.

Data analysis

A thematic analysis of the issues discussed in the interviews and workshops was completed using the research notes and the session recordings. The semi-structured nature of the interviews and workshop breakout room discussions permitted exploration of themes depending on the participants' experience and level of expertise with AI.

Phase 1 provided information on the concepts and practices of AI implementation, with specific emphasis on WHS considerations, but also the context of the current, emerging, and future use of AI in workplaces.

Scorecard development during Phase 1

The draft AI WHS Scorecard underwent an initial revision, drawing on (i) the feedback collected in the interviews and workshops undertaken in Phase 1, and (ii) a further detailed examination of a selected number of studies, reports and AI risk assessment tools identified during the literature review process. Those studies, reports and AI risk assessment tools directly addressed or named risks associated with the application of AI that were or could be relevant to the identification of WHS risks of AI at the workplaces.

There is a larger literature on ethical risks associated with the use of AI, however the AI WHS Scorecard solely drew on evidence reported specifically for workplace risks or for which relevance to WHS could be clearly established.

Phase 2: understanding AI in workplaces

The Phase 2 consultations consisted of a series of in-depth interviews with two groups of individuals: (i) experts, i.e., individuals with strong experience in the introduction of AI technologies in organisations; and (ii) employees, i.e., individuals who had directly experienced, or were about to directly experience, the introduction of AI in their organisation. The broad aim was to use these interviews to build complementary case studies of AI adoption and its impact on workers and workplaces.

The objectives of Phase 2 consultations were to:

- Understand the process of introducing AI in a workplace, including the roles and responsibilities of the people involved.

- Identify risk factors affecting workers' health and safety during the introduction of AI at the workplace and understand the extent to which those risks are considered.
- Identify principles and practices that champion workers' health and safety during the introduction of AI at the workplace.
- Validate and revise the draft AI WHS Scorecard.

Interviews

Participants and recruitment process

Due to the lack of a register or similar source of information about organisations that use AI in Australia, an investigative approach was taken to identify participants for Phase 2 of the research.

Potential participants were identified through contacts established during Phase 1 or through additional searches of publicly available sources such as AI industry networks, innovation centres and innovation labs (mostly university-based). Other sources included websites advertising, promoting, selling or otherwise exploring and discussing AI, as well as professional social media platforms (e.g., LinkedIn).

In total, 37 individuals from 31 different organisations, that included commercial businesses and government organisations (federal, state/territory, and local), were approached by email, enquiring about their interest in being interviewed. Sixteen individuals (from 13 organisations) agreed *in principle* and were emailed a Participant Information Sheet stating the research objectives and explaining the current consultation phase, as well as a participation Consent Form to be signed and returned. Where appropriate, senior management consent for participation was also sought. A total of 12 individuals from 9 organisations participated in one or more interviews; the other four withdrew their original consent. A breakdown of participants by sector and their occupation and field of expertise is provided in Table 3.

Table 3: Overview of participants in Phase 2 consultation.

Participant Type	Sector	Role in Organisation
Employees	*Local government (QLD)*	*Senior manager (road management)*
		Senior manager (road maintenance)
		Data scientist
	Federal government (ACT)	*Data Mining Scientist*
	Social service provider (SA)	*Chief Executive*
	Manufacturing (SA)	*Health and Safety Manager*
		Senior Production Lead
Experts	*Health partnership (QLD)*	*Chief Executive Officer and affiliate of AI Health Alliance*
	Information Technology business (NSW)	*Head of Technology*
	Data analytics and machine learning business (SA)	*Founder*
	AI ethics advisory business (NSW)	*Founder*
	Information Technology/AI business (NSW)	*Commercial Software Strategist*

Interview format and process

Interviews were semi-structured and conducted using a question topic guide tailored for either the business CEO (Appendix C) or employee (Appendix D). Prior to the interview, participants were emailed a copy of the AI WHS Scorecard draft for use during the discussion.

Interviews were conducted by phone or video conferencing tool and varied in length between 40 and 80 minutes. Interviews were recorded with participants' consent and notes were also taken. Where possible, two researchers attended the interview, with one leading and the other acting as scribe.

Data analysis

The interviews were analysed using researchers' notes. The interviews were examined iteratively for information against the four above-mentioned objectives of Phase 2. The researchers conducted a comparative analysis of emergent themes from each group of interviews, which were cross-checked by those members of the research team who participated in the interviews and also compared with emergent themes from Phase 1. The textual analysis process was supplemented by revisiting the recordings. The case studies offered qualitative insight into how organisations were working or planning to work with AI, and how they perceived and managed associated workplace risks.

Scorecard development during Phase 2

The AI WHS Scorecard was further revised drawing on (i) the feedback collected in the in-depth interviews undertaken in Phase 2, and (ii) the identification and integration of a WHS framework. Additional AI WHS risks were identified and added to the scorecard matrix incorporating concepts and examples provided by the interviewees. Each risk was also linked to a specific scorecard dimension using the Safe Work Australia Framework (namely, the four Safe Work hazard/risk categories of: physical, cognitive, biomechanical, and psychological risks).

Phase 3: incorporating the WHS practitioner perspective

One objective in designing and developing the AI WHS Scorecard was to influence awareness of, and practices around managing, WHS risks arising from the use of AI in workplaces. To gain further insight into how the AI WHS Scorecard may be used to identify and assess such risks, a workshop was organised with SafeWork NSW inspectors. SafeWork NSW Inspectors are agents of the regulator, and work with the business community to help improve workplace health and safety. They issue licences for potentially dangerous work, investigate workplace incidents and, where necessary, enforce WHS, workers compensation and explosives laws in NSW. Inspectors regularly visit workplaces in order to provide advice, respond to incidents or complaints, work with businesses to develop targeted injury prevention programmes, and enforce compliance with legislative obligations.

Workshop

Participants and recruitment process

Fifteen inspectors from a standing SafeWork NSW advisory committee were invited to participate in the group consultation. The committee represents different managing units in SafeWork NSW and members come from a range of WHS specialities (e.g., health psychosocial services, construction, hazardous chemicals, operational practice, engineering, organisation capability, system/process improvements, etc.) and a range of locations (e.g., Sydney metro, regional areas).

CWHS initially approached the Chairperson of the committee with the request to invite the inspectors to the workshop. The workshop was scheduled as part of a routine committee meeting with an agreed agenda item. One week prior to the workshop, the Participant Information Sheet and Consent Form were forwarded to the inspectors. A version of the AI WHS Scorecard along with the set of questions for discussion (Appendix E) were also emailed beforehand.

Workshop format and process

The workshop was conducted via teleconference; verbal consent to participate in the research and record the session was obtained from attendees at the outset. Twelve inspectors attended the group consultation which lasted for approximately 1.5 hours. The workshop was structured around four topics, exploring:

- Participants' impression of the AI WHS Scorecard, in terms of its usefulness and suggested improvements.
- Their perception of mapping the AI Ethics Principles against WHS hazards and risks.
- Their response to the scorecard's approach to rating AI ethics risks in the WHS context.
- The aggregation of the AI Ethics Principles into broader groups to simplify the visual presentation, and overall usability, of the AI WHS Scorecard.

Data analysis and scorecard development during Phase 3

Researchers took detailed notes of the workshop discussion and subsequently revisited the recording. A mind-map of themes and issues arising using Ayoa mind-map software was produced based on the review of the material. A comparative analysis of the themes and issues was developed by topic.

Comments and feedback from the Inspectors were incorporated into the final version of the AI WHS Scorecard.

Results and Discussion

This chapter presents the research findings from the three consultation phases.

First, we report the findings from Phase 1 consultations which gathered information about general perceptions of the use of AI and its effect on workplaces, as well as the awareness of ethical issues and, specifically, of the DISER AI Ethics Principles.

Second, we present the findings from Phase 2 consultations which identified the AI adoption process and challenges encountered by organisations.

Third, we present the feedback received in Phase 3 consultations on the utility of the AI WHS Scorecard and potential barriers to its use from the WHS practitioner perspective.

Finally, we summarise how insights from those consultations, together with insights arising from the review of the literature were used to shape the AI WHS Scorecard, to populate it with hazards and risks scenarios, and associated examples, and to iteratively refine it in terms of content and format.

Phase 1: surveying the AI landscape

This first phase of consultation was divided into two distinct components, namely (i) expert and other stakeholder interviews and (ii) online workshops. The two components had different but complementary roles and are reported separately.

Interviews

The interviews contributed to our understanding of:

- The current and likely future use of AI in the workplace.
- The innovation processes typically associated with AI.
- The general level of awareness of the AI Ethics Principles.
- Feedback on early scorecard design.

Challenges of current and future use of AI in workplaces

Participants anticipated that AI would be used for work intensification so that employees could complete more work in a shorter period of time. The majority expected that AI would partially automate tedious and repetitive tasks. They believed impacted employees would have to adapt to new workflows and learn how best to integrate AI solutions into their daily routines. One illustrative example of how employees would be impacted by AI, mentioned by more than one interviewee, was the use of chatbots to field the most common customer service enquiries, thus permitting employees to focus on the more unique and challenging queries. However, participants also identified that as organisations increased their reliance on AI to complete specific tasks, they would have to raise the quality control for

the AI-generated results. Workers would begin to see AI such as chatbots as employees who also needed to be managed and may view monitoring of the chatbot as essential to delivering the core service of their organisation.

A further expectation was that AI would be used for work augmentation. That is, employees would improve the quality of their work owing to features and functionalities provided by AI. Examples of work augmentation due to AI mentioned by participants were:

- Human Resource (HR) departments using AI to provide individualised support to employees and to generate data to contextualise an employee's accomplishments during performance reviews. Here, AI may create barriers between workers and managers if HR started to view its workforce solely through the lens of the metrics and data that the AI tool provides. Communication between workers and managers is a central principle for WHS, so this would prevent adequate WHS consultation.

- The insurance industry using AI as an opportunity for faster claims processing. AI also helps with the underwriting of insurance by providing insurance workers more information and allowing them to select from multiple models to estimate and set insurance premiums. Employees might react differently to the necessity to adapt to those new workflows for claims processing, and the potential new job specifications. Some may also see these developments as a threat to their employment.

- Employees responsible for procurement and managing inventory increasingly relying on AI to inform their decision making. The degree of autonomy that employees have in deviating from the AI recommendations was flagged as a potential issue, for instance, in response to stock requirements, which an AI program had failed to predict.

- Sales staff using AI to rank business opportunities and to gain insights on their prospect of closing a deal. If organisations insisted that sales staff strictly followed AI recommendations, they might have to reconsider their incentive and performance evaluation processes. The perceived issue was that it might become difficult to attribute sales performance to an employee's talent or hard work rather than the AI tool's predictive accuracy.

Participants felt that AI was especially likely to cause large changes to the ways that organisations schedule or allocate workloads for their employees. An example provided described organisations using AI-powered dispatch systems to assign jobs to drivers that were on standby. The AI sought to minimise costs and travel times, and to increase efficiency. Participants gave other examples where AI scheduled desk work. For instance, organisations may use a ticketing system to keep track of jobs and allocate tickets to employees whilst considering constraints, such as an employee's experience or current workload. Participants saw these AI capabilities starting to take over from traditional managerial tasks and expressed concern that AI tools might create barriers between workers and managers. Organisations would need to introduce policies and practices to bridge that gap.

The rise of AI use in the workplace was also thought to challenge how organisations monitoring WHS standards in Australian businesses operated. SafeWork NSW employees interviewed for the project noted it was often difficult to

understand and anticipate the health and safety implications of AI, especially dynamic AI (dynamic AI systems continually learn and adapt while being utilised). The operational behaviour of dynamic AI was considered unpredictable, which would have consequences for attributing accountability and identifying the root causes of accidents involving dynamic AI. The prospect of AI receiving periodic updates that fundamentally changed how the AI operated was concerning for SafeWork NSW participants who were unsure about how to keep pace with continuously updated technology.

WHS management practices when adopting AI

The key drivers for investing in AI were seen to be expected cost savings and gaining competitive business advantages (e.g., offering a new service or new product features, or by efficiency gains boosting production). Concern with the financial benefits of AI, some participants argued, reduced the attention AI adopters gave to concerns for WHS impacts. Participants noted that the sheer speed with which AI was being adopted appeared unprecedented and set it apart from past innovation cycles. There was a perception among some participants that AI was often being created "in the wild" without adequate risk assessment, and without adequate checks and balances in place.

Introducing AI-driven innovation was described as likely involving significant organisational change and requiring careful change management. One participant realised at the outset of an AI project that it would radically change the nature of work at their organisation. It was anticipated that at least some workers might feel uncomfortable about the changes. To address this challenge, the organisation hired an external consultant to lead the change management. Ultimately, the AI project was met with varying levels of resistance. As a result, some workers left the organisation, while others were re-assigned to new tasks. The participant felt that role redesign and redeployment might be inevitable in some instances of AI use, but argued for consulting workers early, engaging them in planning the new workplace arrangements, and identifying where and how the AI tool could affect them.

However, as some participants argued, the benefits of employee consultation were overlooked when organisations were focussed on the cost benefits of AI, which also meant that they might be late to reach out for guidance on managing the workforce implications of AI use. The example given referred to an instance in which, by the time critical workforce issues were identified, an organisation had spent its AI development budget and consequently was reluctant to undertake a potentially costly redesign. Workforce matters, it was suggested, ought to be considered early in the AI development cycle.

Awareness of AI ethics

Participants were interested in and expressed concern about the ethics and wellbeing impacts of AI. Participants referred to a range of stories and conversations about AI ethics and ethical - or unethical - computing that they had followed in the mainstream media, including data breaches. Citizen and consumer issues relating to AI were frequently cited, including the use of profiling, and threats to privacy and data protection.

There was less recognition of WHS impacts of AI: among those consulted, the majority considered AI ethics from an end-user or consumer standpoint. Few had considered AI ethics from the point of view of employee WHS, although notions of "AI for good" and "Tech for good" had some traction and participants had a general understanding of the

role of ethics in AI. Participants saw merit in exploring and better understanding the WHS impacts of AI to prevent harm.

Few participants had detailed knowledge of the DISER AI Ethics Principles (DISER, undated). Awareness of the DISER guidelines appeared to be most common among those who had previously engaged in work within AI ethics or who, due to their role as a data scientist, were required to think and act ethically, in compliance with the legal and regulatory environments.

Participants frequently positioned ethical issues alongside legal obligations, especially concerning data privacy and AI complying with other forms of legislation. An example was that of a utility company using AI to schedule employee workloads within the confines of its enterprise bargaining agreement. While this company had engaged external developers to build the AI tool, it remained the company's responsibility to ensure the tool met this legal obligation. In another example, the DISER Ethical Principle of *contestability* of AI recommendations – and, by implication, accountability in case of their eventual use – was seen to relate to legal obligations, for instance, if an AI program were to lead to accidents and insurance or indemnity claims.

Technology such as AI was sometimes viewed not only as an ethical problem, but as an ethical solution. This view tended to be held especially by engineering and computing specialists, and technical managers who were actively implementing and designing AI systems deployment. These participants identified specific technologies that were being developed to address ethical or legal issues, such as data privacy. Examples included parallel AI technologies, such as "federated learning", that sought to maintain privacy in large datasets by reducing the need to share or combine secure data from multiple sources.

Participants raised the question of who within an organisation was responsible for ensuring that soft ethical requirements of AI were met, in addition to hard legal and regulatory requirements. They noted that AI developers or their clients may lack knowledge of ethical requirements and could not communicate them as part of their risk assessment. In that case, opportunities to design AI systems that are also ethical would be seriously reduced.

Moreover, it was argued that even if a business or programmer was familiar with AI ethics principles and might identify a problem, they may not be equipped with the necessary skills to address that problem. One participant (an ethicist) expressed the view that those developing or using AI typically did not have the skills required for the challenging and meaningful deliberations that were needed to resolve ethical dilemmas. The participant explained that sometimes a benevolent attempt to address an ethical problem of AI gave rise to more insidious ethical issues previously overlooked. AI developers and users alike might thus mistakenly believe that their design was ethical when, in fact, it was not.

This participant further argued that the clients of AI developers were looking primarily for decision support. Their secondary concern was the degree to which decision support was explainable, which was critical for AI to meet the AI Ethics Principle of *transparency*. A repeatedly noted concern was that many current AI technologies were just not wholly explainable, and least explainable for someone without a specialist understanding of AI.

Other participants noted that organisations wishing to adopt AI, but unsure as to the practical steps they needed to take to do so safely, were turning to third-party providers of AI as an assurance that advice was trustworthy and independent. One suggestion from participants was for organisations to engage a market research style (independent) AI ethics review in a triage involving the AI-introducing business itself and any AI programmers. The market research would systematically assess AI impact risks for different actors directly or indirectly affected by or expected to be working with the AI solution.

One participant with an engineering background advocated the need for an independent body to assure or certify AI technologies. Such a body would have the advantage of standing apart from the interests of commercial firms or the government. The participant considered that the technology sector overall in Australia lacked oversight and that a "laissez-faire" attitude to the development of such oversight prevailed, increasing the risk of ethical breaches due to AI. This lack of oversight meant that ethical violations were only addressed after the event.

It was argued that there was a case for a regulatory or advisory response promoting ethical AI. However, several participants warned that any such intervention ought to avoid being burdensome and time-consuming to comply with, as this would mean that they would fail, especially if they remained strictly optional (i.e., advisory).

Feedback on early Scorecard design

Participants preferred a scorecard that was visually and cognitively more accessible than our initial model (version 1.0, see Table 1). One of the early suggestions was to reduce the complexity of the AI WHS Scorecard, which at that stage included 56 risk dimensions as it tabulated seven AI Canvas stages across eight AI Ethics Principles. The participants found it difficult to distinguish with precision between ethics principles that were conceptually closely related. In particular, the three AI Ethical Principles of "human, social and environmental wellbeing", "human-centred values", and "fairness" were seen in many respects to overlap. It was recommended that the AI Ethics Principles be simplified by aggregation into a smaller number of categories.

The participants also suggested an additional item be added to the AI ethics principles: the capacity of the AI system to "forget" or "learning to forget". This item refers to algorithms being set up so that errors and old data be removed during or after the Training stage. This concern was integrated into the AI WHS Scorecard as an item to consider during the Input and Feedback stages of the AI.

Workshops

The workshops continued the discussion of the current and likely future use of AI in organisations. However, their main objective was to gather views on an early draft (version 1.0, Table 1) of the AI WHS Scorecard and suggestions for its content and design.

Workshop participants agreed on the intrinsic value of an AI WHS Scorecard for workplaces and suggested some modifications to the initial draft (version 1.0, Table 1). Participants echoed the views expressed during the interviews that the early scorecard design was unnecessarily complex, also recommending the AI Ethics Principles be aggregated into a smaller number of categories, as shown in Table 4.

Table 4: Higher-level aggregates of the AI Ethics Principles. Adapted from DISER, undated.

Human Condition	Worker Safety	Oversight
Human, social and environmental wellbeing*: Throughout their lifecycle, AI systems should benefit individuals, society and the environment.* ***Human-centred values****: Throughout their lifecycle, AI systems should respect human rights, diversity, and the autonomy of individuals.* ***Fairness****: Throughout their lifecycle, AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups.*	***Privacy protection and security****: Throughout their lifecycle, AI systems should respect and uphold privacy rights and data protection, and ensure the security of data.* ***Reliability and safety****: Throughout their lifecycle, AI systems should reliably operate in accordance with their intended purpose.*	***Transparency and explainability****: There should be transparency and responsible disclosure to ensure people know when they are being significantly impacted by an AI system, and can find out when an AI system is engaging with them.* ***Contestability****: When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or output of the AI system.* ***Accountability****: Those responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled.*

With reference to the three broad ethics categories, workshop participants suggested that, when using the AI WHS Scorecard to explore, assess or ensure the ethical application of AI in the workplace, consideration should be given to the following:

Human condition:

- The psychological, personal and familial, as well as collegial contexts and relationships that may be affecting or be affected by AI in the workplace (AI Canvas: Prediction).
- The risk of positive intentions of AI (possibly merely replicating already existing processes) entailing unintended side effects (Prediction, Judgement).
- The risk of inequitable, discriminatory effects (Outcome).
- Secondary impacts of AI use beyond those intended initially (e.g., health impacts resulting from AI-facilitated intensification of production processes) (Outcome).
- The ultimate unpredictability of some events that AI may seek to predict (Outcome).

Worker Safety:

- The presence of conflating factors that may affect AI reliability and safety, depending, for instance, on variable environmental conditions (Prediction).
- The potential conflict between (and contradiction of) the public analysis of data for scenario testing and the privacy protection awarded to (training) data used in AI (Judgement).
- The risk that AI systems are not immune to gaming but may give a wrongful impression they are immune (Outcome).
- The potentially very personal data required for some AI systems, for instance, when measuring time and motion (especially in real-time) that poses added risk of data abuse (Training).
- The uncertainty about whether what may be considered a safe AI application for one may not prove safe for another person (Training).

Oversight:

- The capacity for human overwrite and (offline) validation of AI systems (Judgement).
- The extent to which employee-employer relationships may shape AI implementation and the information that is being shared (Outcome).
- The potentially onerous nature of AI system contestation (Outcome).
- The impact on third parties (Outcome).
- The transparency of the AI tool, especially where real-time data are used; use and risk of abuse of private data with and without the knowledge of data owners (Training). The extent to which the AI system is set up to grow "organically", responding to changing circumstances (Input).
- The need for continuous monitoring to ensure the validity of prediction and associated actions (Feedback).

Phase 2: understanding AI in workplaces

The Phase 2 consultation consisted of interviews with representatives of commercial and public sector organisations, and AI experts, and sought feedback on the utility of the further revised AI WHS Scorecard. At this stage, the scorecard draft closely resembled the final AI WHS Scorecard (version 2.0) as shown in Appendix F, except for the alignment of AI risks with Safe Work Australia WHS characteristics of work and associated hazards and risks, and the risk rating system (Column F "Characteristics of Work" to Column J "Risk Level"). Also added later as a recommendation from participants of this Phase 2 consultation was Column A, "Main Stages of Development".

Participants specifically provided information about organisations' processes and consultations during the introduction of AI technology in the workplace, the risk factors affecting WHS and their consideration during AI implementation, and the principles and practices that champion workplace health and safety.

Systematic approaches to considering the specific WHS hazards and risks associated with the introduction and the use of AI in the workplace were found to be largely missing from organisations' processes and consultations. Participants generally welcomed the proposal of a tool to assist with identifying and assessing AI-related WHS hazards and risks in the workplace. Participants also made helpful suggestions for further improving the AI WHS Scorecard.

Tracking the introduction of an AI technology in a workplace

The initial conversations with Phase 2 participants focussed on understanding the development and current status of their AI projects. The four organisations (a Local Government Council, a federal government agency, a disability service provider, and a manufacturing business) were at different stages of their AI development and use. The Local Government Council (hereafter: the Council) had started exploring opportunities for using AI to improve service delivery about two years earlier and had since progressed to full use of AI, expanding it into additional service areas. None of the other three organisations had proceeded to use AI fully. The manufacturing business planned to streamline its production process by using AI to propose and assess production schedules; and had commenced examining data and process requirements using AI consultancy services and reviewed implications for its WHS practices. The federal government agency had intensified the use of data analytics to process and match financial data. This data matching had been a task previously undertaken by staff who were now free to undertake different responsibilities, which included – and in this context were reduced to – assessing the resulting predictions and

initiating appropriate follow-on actions. The disability service provider explored AI's use to understand better their clients' satisfaction levels and potential unmet needs. They had yet to determine how exactly to utilise AI.

People and processes

The four organisations not only were at different stages of introducing AI in their workplaces (ideation, development, use), for three of them, it was also clear that they had adopted different approaches to initiating processes. For one organisation, participants could not describe the approach.

In the first approach, two organisations (the manufacturing business and the disability service provider) had adopted a group-based policy that brought together senior personnel from critical units of the organisation to explore and brainstorm if and how AI could be used. In both instances, the process was led at the chief executive level and centred on arranging a workshop for developing AI knowledge in the organisation using external consultants.

In contrast, in the second approach, the introduction of AI in the Council was driven by senior personnel in two of its functional units with a shared concern for improving the efficiency of service delivery. It included personnel with data analytical backgrounds working with the Council's chief digital officer. The process of introducing AI to enhance local service delivery was helped by the Council's participation in the federal government's Smart City Initiative. The Smart City initiative brought together relevant actors from within the Council and introduced them to technological experts through the initiative's broader network of contacts. The early initiation of AI was driven by functional units that would later gather and use the AI-generated data. However, it had soon become apparent that other functional units within the Council would also be affected, notably those charged with record keeping and document management, property rating, and human resources and payroll matters. These functional units needed to be briefed and brought together, and then provide access to and share records and data held typically only for their own tasks and purposes. It also required new ways of thinking about the units' roles and responsibilities in the Council as resources were shared across functional boundaries. The ease with which this was achieved was attributed to the Councils' recent integration of diverse IT systems across the organisation.

Identifying risk factors affecting WHS and their consideration during AI implementation

Interviewees provided little evidence of organisations taking strategic approaches to anticipate the impacts of AI on workplaces beyond the intended process or product change. However, the Council, the manufacturing business and the disability service provider all had applied or were in the process of applying risk assessment strategies using existing WHS policy frameworks. The federal government agency also engaged a unit within the organisation charged with overseeing ethical aspects of AI applications as set out by the DISER principles.

Existing WHS processes were adopted in the absence of an AI WHS Scorecard to identify AI WHS risks and hazards, because the organisations followed WHS public guidelines and regulations with great care. This deliberate, risk-aware approach was especially noticeable in the manufacturing business as it already and routinely needed to pre-empt, reduce and remove physical hazards in the production process. It was thus conditioned to be sensitive to the potential of AI-driven risks affecting the workplace:

"We strive to prevent all accidents and incidents. ZERO work-related injuries and illnesses is our core objective. We integrate WHSIM considerations into our business planning and decision making in all our daily activities." (Manufacturer, health and safety manager)

Overall, the organisations supported the development and use of an AI risk-specific assessment tool since AI risks were seen as new and challenging to anticipate without further guidance.

At present, the organisations adopted different strategies to assess and manage workplace risks. The Council did not have a "high-risk appetite" (senior manager/road management), and its principal orientation was to take a "conservative approach" (ibid), which minimised the need for change and change management as a result of the introduction of AI. Risk assessment was focused on community impacts as the AI involved monitoring and assessing public spaces, which would inevitably include capturing public activities and data on residents. In this organisation, AI was treated as an accepted tool ("everyone knew its power" [Council – senior manager/road maintenance]), with initially less thought given to how its operations might be experienced or perceived in the workforce. As one of the Council interviewees remarked, the approach was to "consider the risk therein", i.e., risks, typically privacy risks, resulting from and emerging due to the use of AI in public spaces. The Council was also concerned to retain "human oversight" (Council - data scientist), albeit primarily to validate AI outputs: whilst AI was used to scan environments for road defects, the system allowed visual inspection of records by humans. However, resulting work orders were scheduled according to an AI-generated urgency score, which was used to allocate work teams to tasks.

Despite its efforts to minimise organisational change, the Council's new work model encountered "push back" (Council – senior manager/road maintenance) from the workforce. The new AI systems meant that the work unit's work schedule "database is filling up more quickly" (ibid.), generating an increased and steadier flow of incident reports that allowed the Council to "bundle work" (ibid). Whereas the unit's workforce had previously been able to prepare its work schedules independently, these were now pre-programmed for them, causing dissatisfaction among some employees. Mediation efforts helped to alleviate concerns and objections, although they appeared not to have removed them entirely.

A different approach had been taken by the disability service provider, whose risk management process sought to take into account generic risks of potential, new AI-informed processes to budgets, workloads, and staffing and training. Although WHS was not explicitly considered in this context, the lead instigator was aware of the need for in-depth consultation within the organisation to ensure the AI project's success:

"The real risk here is that you don't get buy-in from everybody. You end up with a failed project because the passive resistance has manifested itself in certain things you thought being done not being done. Then everyone says << See I told you so, it is a failure >>." (CEO, disability service provider)

Information about the federal government agency's use of AI also indicated limited reflection of potential workforce impacts and the consideration of ethical principles only at the late AI deployment stage. Participants indicated that while the organisation might have considered earlier reflection on such matters as desirable, existing workloads had made this effectively unrealistic, if not impossible to achieve.

Employees of these four organisations suggested that only rudimentary consideration was given to the workforce implications of AI. However, Interviews with experts, i.e., individuals with large experience in the introduction of AI technologies in organisations, indicated that this was not necessarily the norm. Experts further argued that there were exceptions, notably organisations that commenced their AI reflections in consultation with their wider workforce, reaching beyond senior management levels.

Experts also warned against the risk of hype surrounding AI applications and an almost blinding trust in their potential. To their mind, this encouraged potential risks to be underestimated if not dismissed. They also noted that cautious attitudes towards AI were seen to relate to "the personality of the person" (Council – senior manager/road management) and on people "not understanding" (Expert 5) AI and being unreasonably fearful of it. Experts warned about overlooking the broader social consequences of an unbalanced AI application, the potential emergence of a dichotomy of winners and losers of AI that may result from a "disassociation of the workforce" (Expert 1) from the AI operating around it.

Identifying principles and practices that champion workplace health and safety

Phase 2 consultations found scant evidence of active harm reduction strategies specific to AI applications. Where it was evident, it focussed on the adoption of existing WHS principles to AI innovation processes. All four organisations let final oversight of AI applications rest with a (typically senior) member of staff who may assess and, if needed, overrule AI-recommended actions as a strategy that would ensure potentially harmful AI recommendations were stopped.

The "conservative approach" (Council - senior manager/road management) to introducing AI in the workplace, adopted by the Council, sought to minimise the need for change management by leaving the human in charge and avoiding major disruptions to established work processes. The risk of a disruption was assessed against a hypothetical no-change scenario that sought to anticipate associated costs, which would manifest even in the absence of AI.

The Council acknowledged that far-reaching organisational changes due to AI, for instance new data-sharing arrangements, new job descriptions and the creation of new positions. However, potentially harmful implications of AI for WHS were late considerations, coming in at the stage of AI use (rather than at design). Potential harm was handled on a "case by case basis" (Council - senior manager/road management) as workers expressed unease about changes affecting their roles and responsibilities.

Employees suggested that none of the four organisations had specific measures in place to mitigate the WHS risks related to the introduction and use of AI in the workplace. This was largely because they had no existing knowledge of AI WHS risks. However, some organisations actively sought guidance on approaching WHS in the AI context and welcomed the development of the AI WHS Scorecard.

Where the acquisition of AI is motivated by commercial objectives, in the words of one AI expert, it is important to understand the nature of this "appetite for AI" (Expert 1) as it may signal likely beneficiaries of the AI innovation as well as others who may be losing out. In managing AI risk and WHS, it would, therefore, be necessary to identify those likely to be affected by AI in an organisation ("ring-fencing" them, in the words of Expert 4), and identify whether the

effect is positive and beneficial, or adverse and possibly damaging. The AI WHS Scorecard assessment would then help recognise that introducing AI to drive cost-saving may positively and negatively impact workers.

Validating and revising the AI WHS Scorecard

The AI WHS Scorecard received a mix of positive and critical feedback in the Phase 2 consultation. On the positive side, interviewees supported the scorecard's concept of bringing together the AI canvas and the AI ethical framework:

> *"The principle of defining the AI canvas and forcing people to apply the ethical risk lens over it makes sense." (CEO, disability service provider)*

Others welcomed the AI WHS Scorecard's *stepwise* approach to risk assessment, which goes through different stages of the AI implementation process as described by the AI Canvas. This view was particularly shared by interviewees familiar with AI processes or the AI Canvas itself. Those with less AI knowledge found the AI Canvas more challenging to understand.

The AI WHS Scorecard was commended for its focus on "unintended consequences" (Expert 1). This emphasis was seen as appropriate, especially in light of the novelty of AI and its applications and the uncertainty with regards to direct and indirect outcomes and secondary effects that inevitably accompany innovation. One suggestion was to structure the AI WHS Scorecard to facilitate distinguishing between risks that may affect users of AI in the workplace (e.g., those that use AI to predict and then direct workflows) and those subject to its use (e.g., those required to follow and accept AI-predicted work schedules).

One expert compared the AI WHS Scorecard to a "training course" (Expert 3), intended to raise awareness and develop a better understanding of the WHS workplace challenges of AI. In several experts' opinions, financial objectives currently dominate AI applications and, specifically, the "top row of the AI Scorecard" (Expert 4), the early stages that explore the nature of the *prediction* that AI is expected to deliver. In their view, it was essential to re-direct this "monotone" (Expert 4) focus to capture how AI application may change workplaces entirely. The approach taken by the AI WHS Scorecard, of encouraging users to reflect on unintended, unforeseen, perhaps unforeseeable impacts, was expected to help with that process.

Participants acknowledge that the detail of ethical principles and associated risks (and examples of such risks appended to the scorecard) may appear overwhelming to users. However, they also held the view that, without such detailed description of AI risks, it would be challenging to identify and reflect on AI risks sufficiently broadly and comprehensively. The examples given in the AI WHS Scorecard allowed the contextualising of risks and relating them to an organisation's own AI use context.

Against this, one view expressed was that the AI WHS Scorecard may be relevant to "profit-driven organisations" (Council – senior manager/road maintenance) that wanted to use AI to "reduce resources" (ibid.) and that were "cost driven" (ibid.) rather than concerned with improving services and service delivery or increasing the range of services provided. In a similar tone, it was argued that the AI WHS Scorecard might not entirely correspond with the AI development process adopted and experienced by organisations but that an exact correspondence would be hard to

achieve given the diversity of contexts and potential applications for AI use. Instead, the scorecard appropriately sought to capture the variety of such contexts, expecting users to identify those most relevant to their own AI projects.

These contrasting understandings of the AI WHS Scorecard underlined the need to consult and involve a diverse group of workers in planning AI innovations within an organisation, and monitoring and evaluating AI, from the outset. Workers have the most detailed knowledge of the tasks and content associated with their roles. The feedback from case study participants highlighted that best practice implementation of the scorecard ought to involve workers beyond just management or IT. Workforce consultation was already used to help to minimise and manage potential WHS risks associated with AI:

> *"The change management process is designed to ensure all system elements are considered as part of the program planning phase and workers are consulted from concept to completion." (Manufacturer, health and safety manager)*

As also demonstrated by the Council, an exercise in collective brainstorming to anticipate potential impacts of AI on third parties can be a first, pragmatic step for organisations to take as part of their early AI risk assessment.

Feedback on AI WHS Scorecard design

Participants made suggestions to improve the presentation and utility of the AI WHS Scorecard, mostly centred on including risk ratings in its design.

Most participants expressed that rating AI WHS risks was challenging due to the range of variables to be considered, e.g., levels of risks, costs of failure or non-compliance, the relevance of individual risks; foresee-ability of risks. While the rating exercise was seen as possibly time- and resource-demanding, it was nonetheless deemed critical to obtain a comprehensive AI WHS Scorecard. A risk scoring scale would help users of the scorecard to focus on their workforce's core risks. Our initial scoring system used single-item scoring (identifying risk levels as high, medium, low or not applicable). The suggested improvement involved distinguishing between the possible *consequences* for the workforce of violating ethical principles, and the *likelihood* of such violations occurring. The combined scores from these two would then determine an overall *risk level* rating, which guides users to prioritising actions to reduce or remove identifiable risks.

Other suggestions included that:

- The AI WHS Scorecard identified "best practice" (Expert 4) examples to guide users to potential solutions, although it was acknowledged that these practices might not be applicable to, and implementable in, all working environments.
- The Protocol accompanying the AI WHS Scorecard (Appendix G) highlighted its objective *to stimulate reflection* on how ethical principles may guide AI applications (Experts 3,4). This emphasis would acknowledge that additions and amendments to adjust the AI WHS Scorecard were invited when using it.
- The AI WHS Scorecard avoided technical language, which some, especially those less familiar with AI, found difficult to follow. It was suggested that a concept that started clearly with "ideation" (Expert 3) might be more

accessible to users than the current initial stages of Prediction and Judgement adopted from the AI Canvas. Likewise, the final stage in the AI Canvas, Feedback, may be better captured as the final stage "Gone-live" (Expert 3).

- The AI WHS Scorecard included a glossary of terms to help to clarify key concepts referred to in the scorecard (CEO, disability service provider).
- Risks identified in the scorecard be re-written in the form of questions that scorecard users may wish to ask themselves to identify AI risks. Questions might help to encourage reflection about the potential of violating ethical principles and, by extension, WHS principles. Such a reformulation might query, "How do we design (AI) for reducing/avoiding/replacing [ethics risk]?" (Expert 4). This reformulation might give the scorecard a more positive language. In its current format, it appeared to imply "negativity" (Expert 3, 4).

Phase 3: incorporating the WHS practitioner perspective

The Phase 3 consultations sought feedback on the scorecard from WHS inspectors (version 2.0, see Appendix F). The scorecard now included the alignment of AI risks with Safe Work Australia's characteristics of work, and associated hazards and risks (Appendix F, columns F to J). A risk rating framework was also suggested.

None of the inspectors had come across AI-related queries or risks in businesses whose workplaces they inspected or visited. However, they made valuable comments around four discussion topics from their WHS experience and knowledge.

Impressions of the AI WHS Scorecard

The AI WHS Scorecard was perceived as comprehensive by most WHS inspectors who considered it as a potentially useful tool to assist in their consultation with businesses, similarly to other supporting material they use. However, the AI WHS Scorecard was described as more complex, due to its many dimensions making it quite "busy". Inspectors also preferred the scorecard in a format other than MS-Excel, which is used to accommodate the risk rating calculation derived from the likelihood and consequence variables.

Based on their experience of auditing businesses, inspectors believed that the scorecard would have more acceptance amongst larger businesses with greater financial resources and was not readily applicable to small and medium sized enterprises. The reason was the time and the resources that would be required to identify and address the multiplicity of hazards and risks noted on the scorecard (i.e., smaller businesses were "time-poor businesses"). The inspectors noted that when large businesses considered rolling out something new, they often sought WHS inspector advice on the processes, equipment or machinery involved. Large businesses may imitate this behaviour when adopting AI solutions in the workplace.

The WHS Inspectors expected that specific industries, such as manufacturing or construction, were also less likely to find the scorecard applicable. Inspectors felt they had insufficient experience to see how to action or relate AI and the AI WHS Scorecard to these industries. However, they felt the AI WHS Scorecard would be useful in sectors such as AI-based food delivery services around rider instructions. These systems took the decision-making capability out of the rider's hands, without necessarily considering environmental factors (e.g., road quality, suggesting tunnels to cyclists

where they were not permitted), equipment (e.g., bicycle condition), or personal (e.g., level of fitness, hurt ankle); and enforced strict time limits irrespective of these limiting factors. Thus, the proposed AI WHS Scorecard could help identify WHS hazards and risks associated with this type of AI application.

Mapping AI Ethics Principles against WHS hazards and risks

When asked about the process of identifying WHS hazards and risks in practice, the WHS Inspectors' response was "mainly by experience" (e.g., reading hazard reports, seeing or hearing about actual incidents). Inspectors would assess businesses on having demonstrated what was "reasonably practicable" in addressing "foreseeable" hazards. One inspector noted that risk assessment and risk identification depended on the monitoring systems in place (e.g., whether audits were conducted). With AI, risk identification could be more difficult to achieve since the risk may not only emanate from the AI tool directly, but also from the circumstances in which it is applied. For example, risk identification with AI in the case of food delivery riders would need to include consideration of changes in weather conditions or the accuracy of directions provided by map (or GPS) services, to foresee hazards.

Inspectors found that the AI Ethics Principles were well mapped against WHS hazards and risks. Inspectors from the psychological health and safety team at SafeWork NSW found that the mapping of psychological risks was particularly relevant as they are crucial features they look for when addressing situations where an organisation is changing work systems without consulting workers. Although the scorecard mapping was seen to be most helpful in terms of psychological risks, it might be less appropriate for detecting physical harms, such as found in the construction and manufacturing industries. Despite this, one inspector imagined the scorecard could be helpful in a situation in which AI was used to plan a construction project, which, hypothetically, resulted in on-site bottlenecks because the construction process in its entirety and the workforce in particular had not been prepared for the change.

Rating AI risks

The risk rating model used in the AI WHS Scorecard, where individuals could assign either a low, medium or high level of risk to all WHS risks, was found to be overly simplistic. Assigning a level of risk for psychological hazards and risks was considered challenging, for instance, because of the range of potential causes and individual assessments (e.g., there was rarely just one hazard of concern). Inspectors felt that at least four levels of risk should be considered, based on the likelihood (e.g., very low, low, high, very high) and the potential consequence of the risk (e.g., very low-impact consequences to very high-impact consequences). The resulting two-dimensional rating system would combine individual likelihood and consequence ratings to produce a final risk score allowing actions or interventions to be prioritised accordingly.

Aggregation of AI Ethics Principles

Inspectors noted that businesses tended to approach WHS risk assessment from the perspective of legislative and regulatory compliance and costs. They would not recognise or specifically be concerned with "human conditions". Suggestions were made on labelling "oversight" as "governance", which would be closely associated with the legislative reasoning. The label "worker safety" was found appropriate.

Discussion: the AI WHS Scorecard

The three consultation phases generated a variety of suggestions that shaped the AI WHS Scorecard. Insightful information was also collected about the current understanding of workplace hazards and risks associated with the use of AI amongst AI and WHS experts, business and government professionals. A number of factors were considered to determine which suggestions to adopt and incorporate in the construction and progressive improvement of the AI WHS Scorecard, namely: the feasibility of proposed amendments within the scope of this work, primary and secondary evidence that might be available to support a proposed change, and the impact of including or not including a proposed change on the design of the scorecard.

We started the refinement process for the scorecard by consulting literature that touched on the issues raised by our participants in the Phase 1 consultation. Specifically, we grouped the ethical principles into three categories (human condition, worker safety and oversight) and mapped AI risks raised in literature that were relevant to a workplace setting against the seven stages of the AI Canvas (see Table 5). This exercise drew on a diverse literature and the two online workshops. The key contributors are listed in the legend below Table 5 and cross-referenced in the table to the specific risks that they identified or helped to conceptualise, using their numbers in square brackets.

Table 5: AI WHS Scorecard (v1.1) with examples of AI WHS risks identified in the literature and the workshops.

AI Canvas	AI Ethics Principles							
	Human condition			***Worker safety***		***Oversight***		
	Human, social and environmental wellbeing	**Human-centred values**	**Fairness**	**Privacy protection and security**	**Reliability and safety**	**Transparency and explainability**	**Contestability**	**Accountability**
Prediction*: Identify the key uncertainty that you would like to resolve.*	• *Risk of using AI when an alternative solution may be more appropriate or humane. [5,12]* • *Risk of the system displacing rather than augmenting human decisions. [3]* • *Risk of augmenting or displacing human decisions with differential impact on workers who are directly or indirectly affected. [7,9,13]* • *Risk of the resolution of uncertainty affecting ethical, moral or social principles. [9,11,14]*			• *Risk of overconfidence in or overreliance on AI system, resulting in loss of/diminished due diligence. [3,7]*		• *Risk of inadequate or no specification and/or communication of purpose for AI use/an identified AI solution. [2,7,9,15,16]*		
Judgement*: Determine the payoffs to being right versus being wrong. Consider both false positives and false negatives.*	• *Risk of (insufficient consideration given to) unintended consequences of false negatives and false positive. [2,4,11,12]* • *Risk of AI being used out of scope. [3,4,7]* • *Risk of AI undermining company core values and societal expectations. [5,14]* • *Risk of AI system undermining human capabilities. [5]* • *Risk of trading off the personal flourishing (intrinsic value) in favour of organisational gain (instrumental good). [14]*			• *Risk of technical failure, human error, financial failure, security breach, data loss, injury, industrial accident/disaster. [1,7,16]* • *Risk of impacting on other processes or essential services affecting workflow or working conditions. [1,13]*		• *Risk of insufficient/ineffective transparency, contestability and accountability at the design stage and throughout the development process. [12,16]*		
Action*: What are the actions that can be chosen?*	• *Risk of inequitable or burdensome treatment of workers. [1,10]* • *Risk of gaming (reward hacking) of AI system undermining workplace relations. [4,16]*			• *Risk of adversely affecting worker or general rights (to a safe workplace/physical integrity, pay at right rate/EA, adherence to National Employment Standards, privacy). [1,7]*		• *Risk of inadequate or closed chain of accountability, reporting and governance structure for AI ethics within the organisation, with limited or no scope for review. [7,10,14]*		

AI Canvas	AI Ethics Principles							
	Human condition			***Worker safety***		***Oversight***		
	Human, social and environmental wellbeing	**Human-centred values**	**Fairness**	**Privacy protection and security**	**Reliability and safety**	**Transparency and explainability**	**Contestability**	**Accountability**
	• *Risk of worker attributing intelligence or empathy to AI system greater than appropriate.[3]* • *Risk of context stripping from communication between employees.[3]* • *Risk of worker manipulation or exploitation. [5,7]* • *Risk of undue reliance on AI decisions. [3,7]*			• *Risk of unnecessary harm, avoidable death or disabling injury/ergonomics. [1,7,8,16]* • *Risk of physical and psychosocial hazards. [3,16]*		• *Risk of (lack of process) for triggering human oversight or checks and balances, so that algorithmic decisions cannot be challenged, contested, or improved. [3,9]* • *Risk of AI shifting responsibility outside existing managerial or company protocols, and channels of internal accountability (via out- or sub-contracting). [13]*		
***Outcome**: Choose the measure of performance that you want to use to judge whether you are achieving your outcomes.*	• *Risk of chosen outcome measure not aligning with healthy/collegial workplace dynamics. [1,7]* • *Risk of outcome measure resulting in worker-AI interface adversely affecting the status of a worker/workers in the workplace. [3]*			• *Risk of performance measures differentially and/or adversely affecting work tasks and processes. [2,6,10]*		• *Risk of workers (not) able to access and/or modify factors driving the outcomes of decisions. [2,3,9,16]*		
***Training**: What data do you need on past inputs, actions and outcomes in order to train your AI to generate better predictions?*	• *Risk of training data not representing the target domain in the workplace. [7,15]* • *Risk of acquisition, collection and analysis of data revealing (confidential) information out of scope of the project. [7]* • *Risk of data not being fit for purpose [5,8,11,16].*			• *Risk of cyber security vulnerability. [1,11]* • *Risk of (in)sufficient consideration given to interconnectivity/ interoperability of AI systems. [2,9]*		• *Risk of inadequate data logs (inputs/outputs of the AI) or data narratives (mapping origins and lineage of data), adversely affecting ability to conduct data audits or routine M&E. [7,9,10,12]* • *Risk of (rapid AI introduction resulting in) inadequate testing of AI in a production environment and/or for impact on different (target) populations. [2,4]*		

AI Canvas	AI Ethics Principles							
	Human condition			***Worker safety***		***Oversight***		
	Human, social and environmental wellbeing	**Human-centred values**	**Fairness**	**Privacy protection and security**	**Reliability and safety**	**Transparency and explainability**	**Contestability**	**Accountability**
Input*: What data do you need to generate predictions once you have an AI algorithm trained?*	• *Risk of discontinuity of service. [1,13]* • *Risk of worker unable or unwilling to provide or permit data to be used as input to the AI. [9,15]*			• *Risk of impacting on physical workplace (lay out, design, environmental conditions: temperature, humidity). [10,15]* • *Risk of (in)secure data storage and cyber security vulnerability. [1,2,7,10,16]* • *Risk of worker competences and skills (not) meeting AI requirements. [13]* • *Risk of boundary creep: data collection (not) ceasing outside the workplace. [8,15]*		• *Risk of insufficient worker understanding of safety culture and safe behaviours applied to data and data processes within AI. [8,13]* • *Risk of partial disclosure or audit of data uses (e.g. due to commercial considerations, proprietary knowledge). [14,15]*		
Feedback*: How can you use the outcomes to improve the algorithm?*				• *Risk of assessment processes requiring review due to new approach or tool. [9]* • *Risk of identifiable personal data retained longer than necessary for the purpose it was collected and/or processed. [10]*		• *Risk of inadequate integration of AI operational management into routine M&E maintenance ensuring AI continues to work as initially specified. [3,4,8,16]* • *Risk of no offline systems or processes in place to test and review veracity of AI predictions/decisions. [9]*		

Legend: Numbered citations refer to the following sources:

1. ADAPT Centre et al. (2017):	4. Beard and Longstaff (2018)	7. ODI (2019)	10. van de Poel (2016)	13. Wikipedia. (2020)
2. AiGlobal (undated)	5. IEEE (undated)	8. TNO (undated)	11. Walmsley (2020)	14. Online Workshop (Phase 1)
3. Amodei et al. (2016)	6. Matsumoto and Ema (2020)	9. UK Cabinet Office (2020)	12. WEF (2020)	

Upon completing revisions of the AI WHS Scorecard (version 1.1, Table 5) and in time for Phases 2 and 3 of consultation, the AI Ethics Principles identified in columns were transposed into rows to increase readability. The transposition also allowed for additional space to demonstrate the link between the AI Ethics Principles and WHS workplace hazards. This link was added by including the "Key Characteristics of Work" identified by Safe Work Australia in its "Principles of Good Work Design" (see Literature Review, Figure 1) and by aligning each AI Ethics Principle and associated AI risks to a SafeWork Australia identified workplace "hazard or risk". We proceeded to revise the scorecard systematically and iteratively. The final design thus incorporated the following key suggestions from our participants:

- Inclusion of a range of risks and hazards (and examples for these).
- Modifications suggested about the presentation of AI Ethics principles (3 higher level categories).
- Retention of specific details including but not exclusively:
    - privacy and contestability as WHS and AI ethics concerns,
    - independent oversight as an AI ethics as well as risk management principle,
    - the role of communication within organisations using or intending to use AI,
    - the importance of explainability of AI.
- Simplification of the AI Canvas to a smaller number of higher-level stages.
- Linking of AI ethics principles, and associated AI hazards, to the WHS concept of Characteristics of Work, and their hazards and risks.
- Inclusion of a risk rating system to assist users in determining the possible consequences of AI risks alongside the likelihood of AI risk events occurring.
- Inclusion of a list of examples of AI hazards and risks, each corresponding to their more broadly captured risk in the AI WHS Scorecard.

The final AI WHS Scorecard

The final AI WHS Scorecard (version 2.0, Appendix F) is accompanied by a Protocol explaining its context and recommending how it may be used (Appendix G).

The AI WHS Scorecard incorporates the Australian Government endorsed AI Ethics Principles, which are used to identify and understand potential WHS risks of AI. It adopts the AI Canvas, which identifies the stages through which organisations transition as they conceive, develop, and use AI. Within these dimensions, AI-related WHS risks are described and linked to specific hazards and risks that Safe Work Australia has defined as part of Principles of Good Work Design.

The AI WHS Scorecard, available in MS Excel format, is equipped with a risk rating matrix. The risk matrix deconstructs a limited number of risk categories (low, low medium, medium, medium high, high; visualised by different colours) as a combination of two dimensions: the *consequence* of an adverse event, and the *likelihood* of that event. The risk matrix is a simple tool that one can use (i) to identify a risk and decide if it can be tolerated, and (ii) to prioritise which risks need to be addressed first. The approach taken by the AI WHS Scorecard is to formulate the potential

consequences of AI use for WHS from both the perspective of workers (see row labelled “Worker”, Table 6) and that of an organisation and its ability to perform its core service (see row labelled “Organisation”, Table 6). This design draws on the NSW Government tip-sheet “Overview of work-related stress” (SafeWork NSW, undated) which explains how increased stress levels of workers in an organisation can lead to diminished organisational performance.

The magnitude of effects on workers and organisations is measured using a five-point scaled rating, ranging from insignificant or negligible, moderate or extensive, to significant (see Table 6).

The combination of the consequence and likelihood scales results in a gradient of low to high risk levels. The gradient used in the AI WHS Scorecard builds on Julian Talbot’s discussion of the use of risk matrices (Talbot, 2018).

Appendix H provides an illustration of how the AI WHS Scorecard may be used, based on a fictitious example of a manufacturer seeking to adopt AI for improved machine maintenance.

Table 6: Risk rating system of the AI WHS Scorecard. Adapted from Safework NSW, undated, and Talbot, 2018.

			Consequence				
		Worker	Negative impact on mood. Staff may be irritated and inconvenienced.	Temporary reduction in productivity and efficiency	Decline in job satisfaction, morale, cohesion, and productivity.	Increase in absenteeism and conflicts at work.	Increase in staff turnover, health care expenditure and worker's compensation claims.
		Organisation	Minimal impact on non-core business operations. The impact can be dealt with by routine operations.	Some impact on business areas in terms of delays and quality. Can be addressed at the operational level.	Reduced performance such as not meeting targets, but organisation's existence is not threatened.	Breakdown of key activities leading to substantial reduced performance. Survival of organisation threatened.	Critical failure preventing core activities from being performed. Survival of organisation threatened.
	Qualitative Likelihood		**Insignificant**	**Negligible**	**Moderate**	**Extensive**	**Significant**
Likelihood	Is expected to occur in most circumstances	**Almost Certain**	Medium	Medium High	High	High	High
	Will probably occur in most circumstances	**Likely**	Low Medium	Medium	Medium High	High	High
	Might occur at some time	**Possible**	Low Medium	Low Medium	Medium	Medium High	Medium High
	Could occur at some time	**Unlikely**	Low	Low	Low Medium	Medium	Medium High
	May occur only in exceptional circumstances	**Rare**	Low	Low	Low	Low Medium	Medium

Conclusion

The research findings demonstrated high levels of concern about, and interest in better understanding, the potential effect that use of AI in the workplace may have on workers' health and safety. There is a lack of information and evidence concerning workplace effects of AI use. The knowledge gaps contrasted with the anticipated impact of AI from an ethical and WHS perspective, as commercial and other organisations utilise it for accelerating production processes and for improving products and services. Even though there was common consensus on the importance of understanding the impacts of AI on workers and managing any associated potential risks, the lack of resources to take actions was also acknowledged.

A review of the literature confirmed the limited resources currently available for analysing and describing the effects of AI on workers and workplaces. A closer inspection of the general literature on AI implementation strategies, ethics principles, and WHS practices identified concepts and examples of AI related risks to workers, which helped to inform the development of the AI WHS Scorecard, which was refined and improved throughout the research.

Organisations and AI experts using, preparing and planning to use AI in the workplace, as well as WHS practitioners were supportive of the overall intention and proposed design of the AI WHS Scorecard. Evidence suggests that when an organisation develops or uses AI, the new technology's impact on the workforce may only become of concern during later stages of this process, although organisations may consider implications for conventional WHS rules and regulations. At a late implementation stage, it may not be feasible to add AI features or make technical changes that ensure protection for workers. Thus, the proposed AI WHS Scorecard was seen to be helpful in guiding organisations in the process of adopting and using AI, while being cautious about the various ways in which AI may affect workplaces, the workers, and WHS.

The final version of the AI WHS Scorecard is ready for use. It uniquely unites three different modes of thinking regarding AI adoption. By incorporating the AI Canvas, the development of the scorecard facilitates a pragmatic stepwise approach that helps organisations identify and conceptualise AI opportunities. In integrating the AI Canvas with the AI Ethics Principles in a structured manner, the present research has sought to overcome current shortcomings of existing AI ethics frameworks lacking a distinct workplace focus, and of existing business scorecards lacking specific consideration of AI ethics in the workplace. By linking AI Ethics Principles to WHS hazards and risks, the research has also sought to fill a gap in WHS by providing a tool for inspecting AI use to those charged with ensuring workplaces are safe.

For organisations using or planning to use AI, the AI WHS Scorecard is intended to raise awareness and stimulate reflection on the effects that AI may have on workplaces. The scorecard is expected to further evolve with the continued use of AI in organisations and through its adoption by these organisations. Whilst the research reached out and was informed by a diverse population of AI experts, WHS managers and inspectors, and public and private sector managers and employees (with and without AI use experience), there is scope for further strengthening the

consultation with AI users across the workplace spectrum. Future work may focus on consulting more users of AI in the workplace, offer more detailed observations of AI use in practice, and trial the use of the AI WHS Scorecard in organisations.

Acknowledgements

We would like to acknowledge the contributions of the interview and workshop participants who made available their time and offered their thoughts to this study, the SafeWork NSW operational managers and inspectors and, especially, the businesses and their representatives who took part in our study. Without each and everyone's feedback and suggestions, this project would not have been possible.

This work was funded by the NSW Government's Centre for Work Health and Safety who also oversaw the work, reviewed the report and approved its publication.

References

ADAPT Centre, Trinity College Dublin & Dublin City University. (2017) *Ethics Canvas v1.8*. Retrieved from www.ethicscanvas.org

Agrawal, A., Gans, J., & Goldfarb, A. (2018a) A Simple Tool to Start Making Decisions with the Help of AI. *Harvard Business Review,* 17 April. Retrieved from https://hbr.org/2018/04/a-simple-tool-to-start-making-decisions-with-the-help-of-ai

Agrawal, A., Gans, J., & Goldfarb, A. (2018b) *Prediction Machines: The Simple Economics of Artificial Intelligence*. Boston: Harvard Business Review Press.

AiGlobal. (undated) *Responsible AI Design Report Card*. Retrieved from https://ai-global.org/

AIHLEG. (2019) *Ethics Guidelines for Trustworthy AI*. High-Level Expert Group on Artificial Intelligence. Brussels, European Commission.

Amodei, D., C. Olah, J. Steinhardt, P. Christiano, J. Schulman & D. Mané. (2016) *Concrete Problems in AI Safety*. arXiv:1606.06565v2.

Australian Human Rights Commission. (2019) *Human Rights and Technology*. Discussion paper. Sydney.

Autor, D., Mindell, D., & Reynolds, E. (2020) *The Work of the Future: building better jobs in an age of intelligent machines*. Massachusetts Institute of Technology, Cambridge, MA.

Beard, M. and Longstaff, S. (2018) *Ethical Principles for Technology*. Sydney, The Ethics Centre.

Calvo, R. A., Peters, D., Vold, K., Ryan, R. M., Burr, C., & Floridi, L. (2020). Supporting human autonomy in AI systems: A framework for ethical enquiry. In: *Ethics of Digital Well-Being: A Multidisciplinary Approach*. Springer.

Commission on Workers and Technology. (2019) *Background Briefing Note*. June 2019. Fabian Society, London.

Dawson D. and Schleiger E., Horton J., McLaughlin J., Robinson C., Quezada G., Scowcroft J., & Hajkowicz S. (2019) *Artificial Intelligence: Australia's Ethics Framework*. Data61 CSIRO, Australia.

Devitt, K., Gan, M., Scholz, J., & Bolia, R. (2020) *A Method for Ethical AI in Defence*. Department of Defence, Canberra.

DISER. (undated) *AI Ethics Framework*. Department of Industry, Science, Energy and Resources. Retrieved from https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework

Donati, P. (2020) *How to promote the dignity of work in the face of its hybridization in the digital economy*. Proceedings of the Workshop, Dignity and the Future of Work in the Age of the Fourth Industrial Revolution. 14-15 October 2019, Studia Selecta. Vatican City. Retrieved from http://www.pass.va/content/scienzesociali/en/publications/studiaselecta/dignity_of_work/donati.html

Government of Australia. (2011) *Work Health and Safety Act 2011*. Retrieved from https://www.legislation.gov.au/Details/C2018C00293

Government of Canada. (undated) *Algorithmic Impact Assessment Tool*. Retrieved from https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

Griffin, M., Chapman, M., Hosszu, K., Orchard, M., Parker, S., Jorritsma, K, Gagne, M., & Dunlop, P. (2019) *MAPNet: Rethinking Work Skills for the Future*. White Paper for the Future of Work Institute, Curtin University, Perth.

Hagendorff, T. (2020) The Ethics of AI Ethics: An Evaluation of Guidelines. *Mind and Machine*, 30: 99-120.

Hajkowicz, S.A, Karimi, S., Wark, T., Chen, C., Evans, M., Rens, N., Dawson, D., Charlton, A., Brennan, T., Moffatt, C., Srikumar, S., & Tong, K.J. (2019) *Artificial intelligence: Solving problems, growing the economy and improving our quality of life*. CSIRO Data61, Australia.

Horton, J., Cameron, A., Devaraj, D., Hanson, R.T., & Hajkowicz, S.A. (2018) *Workplace Safety Futures: The impact of emerging technologies and platforms on work health and safety and workers' compensation over the next 20 years*. CSIRO, Canberra.

IEEE. (undated) *A Call to Action for Businesses Using AI. Ethically Aligned Design for Business*. New Jersey, IEEE Standards Association.

IEEE. (2016) *Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems (AI/AS)*. The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. Retrieved from https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v1.pdf

Kellogg, K. C., Valentine, M. A., & Christin, A. (2019) Algorithms at Work: The New Contested Terrain of Control. *Academy of Management Annals*, 14(1): 366–410. https://doi.org/10.5465/annals.2018.0174.

Mantelero, A. (2018) AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law and Security Review*, 34 (4): 754-772.

Matsumoto, T. and Ema, A. (2020) *Proposal of the Model Identifying Risk Controls for AI Services*. The 34th Annual Conference of the Japanese Society for Artificial Intelligence, Kumamoto, Japan, June 12, 2020. Retrieved from https://confit.atlas.jp/guide/event/jsai2020/subject/4N2-OS-26a-02/tables?cryptoId=

McKinsey Digital. (2020) *Global survey: The state of AI in 2020*. Retrieved from https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Analytics/Our%20Insights/Global%20survey%20The%20state%20of%20AI%20in%202020/Global-survey-The-state-of-AI-in-2020.pdf

Moore, P. V. (2018) *The Threat of Physical and Psychosocial Violence and Harassment in Digitalized Work*. Geneva, Switzerland: International Labour Office.

NSW Government. (2019) *NSW AI Ethics Framework*. Retrieved from: https://www.digital.nsw.gov.au/transformation/policy-lab/artificial-intelligence-ai/nsw-ai-ethics-framework

ODI. (2019) *Data Ethics Canvas*. Open Data Institute. Retrieved from theodi.org/tools

OECD. (2019) *Digital Innovation: Seizing Policy Opportunities*. OECD Publishing, Paris.

O'Neil, C. (2016) *Weapons of Math Destruction. How big data increases inequality and threatens democracy*. Allen Lane, UK.

Paschen, U., Pitt, C., & Kietzmann, J. (2020) Artificial intelligence: Building blocks and an innovation typology. *Business Horizons*, 63: 147-155.

Perrault, R., Shoham, Y., Brynjolfsson, E., Clark, J., Etchemendy, J., Grosz, B., Lyons, T., Manyika, J., Mishra, S., & Niebles, J.C. (2019) *The AI Index 2019 Annual Report*. AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA.

Pew Research Center. (2018) *Artificial Intelligence and the Future of Humans*. Retrieved from https://www.pewresearch.org/internet/2018/12/10/artificial-intelligence-and-the-future-of-humans/

Safe Work Australia. (2018) *How to manage work health and safety risks Code of Practice*. Retrieved from https://www.safeworkaustralia.gov.au/system/files/documents/1901/code_of_practice_-_how_to_manage_work_health_and_safety_risks_1.pdf

Safe Work Australia. (undated) *Principles of good work design. A work health and safety handbook*. Retrieved from https://www.safeworkaustralia.gov.au/doc/handbook-principles-good-work-design

SafeWork NSW. (undated) *Overview of Work-related stress: Tip Sheet 1*. Retrieved from https://www.safework.nsw.gov.au/resource-library/mental-health/mental-health-strategy-research/stress-tip-sheets/overview-of-work-related-stress

SmartDubai. (2020) *AI Ethics Self-Assessment Tool Report*. Retrieved from https://www.smartdubai.ae/self-assessment

Talbot, J. (2018) *What's right with risk matrices?* Retrieved from https://www.juliantalbot.com/post/2018/07/31/whats-right-with-risk-matrices

TNO. (undated) *Safe application of robots in the work place - safety chart*. Retrieved from https://www.tno.nl/en/focus-areas/healthy-living/roadmaps/work/healthy-safe-and-productive-working/safe-working/

UK Cabinet Office. (2020) *Data Ethics Framework*. London, UK Cabinet Office. Retrieved from https://www.gov.uk/government/publications/data-ethics-framework

van de Poel, I. (2016) An Ethical Framework for Evaluating Experimental Technology. *Science and Engineering Ethics,* 22: 667-686.

Walmsley, C. (2020) The Impact Canvas: An Ethical Design Experiment. *Design Management Review*, 31: 20-25.

WEF. (2020) *Companion to the Model AI Governance Framework. Implementation and Self-Assessment Guide for Organizations*. Geneva, Switzerland, World Economic Forum.

Wikipedia. (2020). *Workplace impact of artificial intelligence*. Retrieved from https://en.wikipedia.org/w/index.php?title=Workplace_impact_of_artificial_intelligence&oldid=977454706

Appendices

A: Expert and stakeholder interview guide.

Preamble – introduce and summarise the project

Thank you for your time today. We have been commissioned by the NSW Centre for Work Health and Safety to conduct research into the potential impacts and risks of adopting Artificial Intelligence (AI) technology in business (and other organisations). Our focus is on understanding the impact of AI on occupational health and safety. We are interviewing experts in AI or otherwise familiar with its challenges and opportunities. Our specific interest is the ethical application of AI in workplaces so to reduce any risk to occupational health and safety. Later in this study we also hope to speak with some businesses/organisations that have or are about to implement AI technologies, and gather some insight about the processes they have adopted in doing so.

With the information we gather, we will develop a scorecard with protocol to assist businesses/organisations in adopting AI technology in ways that promote occupational health and safety.

We would like to talk with you about some of our initial ideas we have about what the scorecard and protocol should look like. We are looking for commentary, corrections and other suggestions.

[Confirm receipt of Information Sheet. Collect Consent Form, as appropriate. Confirm consent.]

Our conversation today should last approximately 45 minutes but may take longer if you wish.

Introduction

- Please introduce yourself and tell us about your role.
- What is your relationship to AI? For instance, do you work in this area directly?
- If working in an organisation that produces or has adopted AI, please tell us a little about it.
    - What kind of business/organisation do you work for?
    - What stage is your organisation at in terms of introducing, using or producing AI?
    - Within your organisation, what is your responsibility in that regard?

AI ethics guidelines

- We are interested in your view of ethics in the application of AI in workplaces. Are you familiar with the AI ethics guidelines produced by CSIRO?

  [Showcard: AI ethics].

  [Provide context as appropriate.]
- Just thinking about the implementation or application of AI in workplaces,
- Which of these ethics criteria do you think are most relevant?
- Is the list complete – or anything missing?
- At what stages of the process should they be considered?

- How should this be done? Are there any examples? [Prompt: for instance: how would one assess "fairness" of AI technology in the workplace?]
- Who should be involved?

AI Canvas

- Are you familiar with the "AI Canvas" that was produced by a group of researchers and academics in Toronto, Canada, and is now often used to understand the stages of introducing AI technology?
  [Showcard: AI Canvas]
- Again, just thinking about the implementation or application of AI in workplaces,
- At which point in this canvas do ethical concerns come into play?
- Which kind of ethical concerns?
- Are there any examples?

Scorecard development

- Thinking about the AI Canvas and the AI ethics guidelines, how helpful would it be to combine these two? That is, helpful to organisations that are concerned about the impact of AI on their workplace.
  [Showcard: Canvas/Ethics matrix]
- Would a combination in the form of a matrix (as shown) be a useful and practical tool for a business/an organisation to use?
- What would be its strength/weaknesses?
- Are there key areas in this matrix that a scorecard concerned with workplace health and safety should focus on?
- Are there areas of lesser importance? (Why? How can we identify these?)
- We are nearing the end of the time allocated for our discussion. Before we finish, is there anything else you would like to mention?
- May we contact you with any follow up questions or points for clarification arising from our discussion today? If so, what is the best way to contact you?
- Would you like to receive a copy of our final project report, sharing our insights from this research?

Thank you again for your time today.

B: Online workshop guide.

Breakout Room: AI WHS Canvas

- Facilitator introduction
- With conference link participants should have received:
    - A case study we want to work with in this session [share "Case Study Showcard"] prepared to align with the AI Canvas (designed by Agarwal et al. in Toronto) which we have cross-tabulated with the CSIRO's AI Ethics Principles.
    - The result is this AI Matrix [share]
- Seven AI Canvas items and eight AI ethic principles: call for simplification?
    - We have subdivided the CSIRO's 8 ethics principles into 3 broad categories: Human condition, Safety and Oversight
    - We also want to focus on just 3 AI Canvas items: Judgement, Outcome and Training.
- In a business, decisions re these are typically based on cost-benefit calculations, for instance:
    - Judgement: the cost of the AI tool getting it wrong
    - Outcome: another cost-benefit indicator, here: revenue
    - Training: what data are needed to program the AI tool. For business, a question of availability and cost of making data available.
- We want to explore with you if beyond cost-benefit considerations, ethical principles that we describe as human dimension, safety and oversight also might or ought to be considered? If so,
    - Which one?
    - How should we measure whether ethics principles are met?
    - Who are the stakeholders to be involved to answer these questions?
- IMPORTANT reminder: the focus is on workplaces:
    - NOT (external) customers
    - NOT society at large.
- We have about 10 minutes per broad category of ethics principle.

C: Case study interview guide – CEO.

Preamble

Thank you for your time today. We have been commissioned by the NSW Centre for Work Health and Safety to conduct research into the potential impacts and risks of adopting Artificial Intelligence (AI) in business (and other organisations). Our focus is on understanding the impact of AI on occupational health and safety. We have already interviewed experts in AI or otherwise familiar with its challenges and opportunities. Our specific interest is in the ethical application of AI technology in workplaces so to reduce any risk to occupational health and safety. We are now speaking with senior managers and employees of businesses/organisations that have implemented, or are about to implement, AI technologies, and gather insights about the processes they have adopted in doing so.

With the information we gather, we will develop a scorecard with accompanying protocol to assist organisations in adopting AI technology in ethical ways that promote occupational health and safety.

We would like to talk with you about your business's/organisation's AI use (or planned use), the rationale for this innovation and the processes involved. We would also like to discuss with you the utility of the scorecard we have prepared to date. Specifically, we would like to explore whether it might be helpful in the context of your business/organisation configuring its AI project.

The scorecard sets typical AI development processes (although your business/organisation may not have followed these in any detail) against a set of ethical principles, which were originally developed by the government agency CSIRO/data61, and endorsed by the Australian Federal Government. Our own research to date has suggested slight modifications to those principles, which are reflected in our scorecard. We will explain this further during our conversation.

[Confirm receipt of Information Sheet. Collect Consent Form, as appropriate. Confirm consent.]

Our conversation today should last approximately one hour but may take longer if you wish.

Interview questions

About Yourself – and the Business/Organisation

- Please introduce yourself and your business/organisation.
- What does your business/organisation produce or provide?
- How many employees does it have?
- Who are your main clients or customers?
- What is your role in the business?
- How is management structured?
- What experience does your business/organisation have in using AI in the workplace?
- Has your business/organisation another other experience with AI? [PROMPT: as producer?]

- Does your business/organisation have any AI experienced employees, that is, employees who have previously work on AI-related projects within or outside your organisation?

About the AI innovation

- What is the AI technology that your business/organisation has adopted or is in the process of adopting?
- What stage is the AI project in terms of development and use?

The beginnings

- When was the AI project idea conceived?
    - What triggered this? (PROMPT: internal business consideration, competition, something else?)
    - What is/was the objective that the innovation sought to achieve or help to achieve?
    - Are or were there alternatives to AI for achieving the same objective?

  [PROMPT: What are/were they?]

Planning and early implementation process

Please tell us how the AI project was developed and, if appropriate, rolled out/put to work.

- Were there identifiable stages?
- What was explored at each of these stages? And how long did it take to conclude that stage?
- Who was involved in these stages?
- What is your own relationship to the AI project in your business/organisation?
- Beyond those directly involved, did you engage or consult any others in the business/organisation?
- Did you identify anyone with responsibility for delivering the AI project?
- Did you start with a clear plan for implementing the project? Or was it more likely evolving?
- Did you engage outside contractors? Who? To do what?

Outcomes

- How has the AI application changed your business/organisational practices?
- How about business/organisational performance?

Impacts of workplace

- What has it meant for your workforce?
- What processes or products are affected?
- Are there any effects on how the business organised its workflows?
- Are any employees affected? Are job rolls affected?

[Only ask if participant is/was directly involved in AI implementation, that is other than and in addition to executive oversight. Ask if there is time for some more questions about our scorecard and if we may come back at a later stage.]

- Are you familiar with the “AI Canvas” that was produced by a group of researchers and academics in Toronto and is now often used to understand the stages in introducing AI/machine learning technology? [Showcard: AI canvas]
- Do you recognise the stages identified in this AI canvas amongst your own stages of AI implementation?

- If no, which aspects are different? How easy or hard would it be to match these stages onto your business's/organisation's conceptualisation of implementation stages?

- We would like to test the utility of our scorecard. In the following, we would like to use this chart to explore your experience of the AI implementation process. If you find that the implementation stages depicted in this scorecard do not match your understanding of these steps and sequences, we can use your own reference points instead. [Determine preference].
- To begin with, could you tell us whether, as far as you are aware, at each stage of the AI development process any of the following ethical principles were considered?
  - If so, how and when, and who was involved? And what exactly was reflected upon?
  - Were there other issues that may be of an ethical nature such as those described here considered? If so, what were they?
  - How was it determined that ethical principles were met?

End

- We are nearing the end of the time allocated for our discussion. Before we finish, is there anything else you would like to mention?
- May we contact you with any follow up questions or points for clarification arising from our discussion today? If so, what is the best way to contact you?

Thank you again for your time today.

D: Case study interview guide – employee.

Interview questions and prompts

Preamble

Thank you for your time today. We have been commissioned by the NSW Centre for Work Health and Safety to conduct research into the potential impacts and risks of adopting Artificial Intelligence (AI) in business/in an organisation. Our focus is on understanding the impact of AI on occupational health and safety. We have already interviewed experts in AI or otherwise familiar with its challenges and opportunities. Our specific interest is in the ethical application of AI technology in workplaces so to reduce any risk to occupational health and safety. We are now speaking with senior managers and employees of businesses/organisations that have implemented, or are about to implement, AI technologies, and gather insights about the processes they have adopted in doing so.

With the information we gather, we will develop a scorecard with accompanying protocol to assist businesses/organisations in adopting AI in ethical ways that promote occupational health and safety.

We would like to talk with you about your business's/organisation's AI use, the rationale for this innovation and the processes involved. We would also like to discuss with you the utility of the scorecard we have prepared to date. Specifically, we would like to explore whether it might be helpful in the context of your business/organisation configuring its AI project.

The scorecard sets typical AI development processes (although your business/organisation may not have followed these in any detail) against a set of ethical principles, which were originally developed by the government agency CSIRO/data61, and endorsed by the Australian Federal Government. Our own research to date has suggested slight modifications to those principles, which are reflected in our scorecard. We will explain this further during our conversation.

[Confirm receipt of Information Sheet. Collect Consent Form, as appropriate. Confirm consent.]

Our conversation today should last approximately one hour but may take longer if you wish.

About Yourself – and the Business/Organisation

- Please introduce yourself and your business/organisation.
    - What is your job/role in the business/organisation?
    - Who do you report to / how many employees directly report to you?

We are interested in exploring with you the introduction, implementation and, insofar as relevant, current use of [name/describe AI project].

- Can you please tell me, what was or has been your role with respect to that project?

Planning and Implementation process

Please tell us how the AI project was developed and, if appropriate, rolled out/put to work.

- Were there identifiable stages?

- At which of these stages were/are you directly involved? [PROMPT: In what capacity? What tasks?]
- Are you familiar with the "AI canvas" that was produced by a group of researchers and academics in Toronto and is now often used to understand the stages in introducing AI/machine learning technology? [Showcard: AI canvas]
- Do you recognise the stages identified in this AI canvas amongst your own stages of AI implementation?
    - If no, which aspects are (most) different? How easy or hard would it be to match these stages onto your business's/organisation's own conceptualisation of implementation stages?

[Using AI Canvas or, if participant prefers, using self-identified stages – focus on areas with direct involvement]

- We would like to test the utility of our scorecard. In the following, we would like to use this chart to explore your experience of the AI implementation process. If you find that the implementation stages depicted in this scorecard do not match your understanding of these steps and sequences, we can use your own reference points instead.

[Determine preference].

- To begin with, could you tell us the extent to which at stages of the AI development process to which you contributed, any of the following ethical principles were considered? If so, how and when, and who was involved? And what exactly was reflected upon?
    - Were there other issues that may be of an ethical nature such as those described here considered? If so, what were they?
    - How was it determined that ethical principles were met, if at all?
    - What exactly was explored at each of these stages?
    - How long did it take to conclude that stage?
    - Who (else) was involved in these implementation stages?

Impacts on workplace

- What has the AI project meant for the workforce/your colleagues?
- What processes or product (has) does it replace(d), remove(d) or add(ed) to?
- What processes or products are affected?
- Are there any effects on how the business/organisation manages its workflows?
- Are any employees affected? Are job rolls affected?

End

- We are nearing the end of the time allocated for our discussion. Before we finish, is there anything else you would like to mention?
- May I [or another project team member] contact you with any follow up questions or points for clarification arising from our discussion today? If so, what is the best way to contact you?

Thank you again for your time today.

E: WHS Inspector Advisory Group consultation.

Interview questions and prompts

Thank you for agreeing to participate in this workshop. The objective of this workshop is to get feedback on our draft AI WHS scorecard.

- Please tell us what you think of the scorecard.
    - How useful might it be to your work?
    - Any instant suggestions for improvements?
- The scorecard seeks to map AI ethics principles against WHS hazards and risks. Do you agree with our current mapping? How useful is this?
- We are specifically interested in your opinion of how AI ethics risks identified in the scorecard may be rated. We currently invite users to rate risks subjectively as 'low', 'medium' or 'high'. What is your opinion on this way of rating risks? What are the alternatives?
- We currently aggregate AI ethics principles into three broad groups, which we name "human condition', "worker safety" and "oversight". Do you think these labels 'work'? Can you suggest better alternatives?

Thank you again for your time.

F: Final AI WHS Scorecard (version 2.0).

Below is a static version of the interactive scorecard. It has been completed using examples. The consequences, likelihoods and risk level are all for demonstration purposes.

A	B	C	D	E	F	G	H	I	J
Main Stages of Development	**AI Canvas**	**Ethics Domains**	**Ethics Risks to WHS**	**Examples – Potential WHS Related Harms**	**Characteristics of Work**	**WHS Hazards and Risks**	**Consequence**	**Likelihood**	**Risk Level**
Ideation	**Prediction:** Identify the key uncertainty that you would like to resolve.	Human condition	Risk of using AI when an alternative solution may be more appropriate or humane.	Predicting a worker's physical or mental exhaustion levels for monitoring purposes without instituting strategies to prevent exhaustion in the future.	Psychological	Work demands	Insignificant	Rare	
		Human condition	Risk of the system displacing rather than augmenting human decisions.	Prediction tool changes allocation of roles and responsibilities, with some worker assigned higher status roles, others relegated to lower status roles, or facing redundancy.	Psychological	Organisation justice	Insignificant	Unlikely	
		Human condition	Risk of augmenting or displacing human decisions with differential impact on workers who are directly or indirectly affected.	A warehouse manager for a toy company ignores feedback from order fulfilment staff that a popular toy is about to sell out during the pre-Christmas period, because the AI stock control tool predicted adequate stock levels. Staff are disempowered and demotivated.	Biomechanical	Job control	Insignificant	Possible	
		Human condition	Risk of the resolution of uncertainty affecting ethical, moral or social principles.	Predicting the health/health trajectory of an employee, such as likelihood of pregnancy, may contravene right to privacy or social/moral convention.	Psychological	Organisation justice	Insignificant	Likely	
		Worker safety	Risk of overconfidence in or overreliance on AI system, resulting in loss of/diminished due diligence.	After a six-month 'break-in' period without incidents at a new AI-enabled plant, preventive safety measures are no longer prioritised; new employees are no longer trained in PPE requirements.	Cognitive, Physical	Physical hazards, Information processing load, Complexity and duration	Insignificant	Almost Certain	

A	B	C	D	E	F	G	H	I	J
Main Stages of Development	**AI Canvas**	**Ethics Domains**	**Ethics Risks to WHS**	**Examples – Potential WHS Related Harms**	**Characteristics of Work**	**WHS Hazards and Risks**	**Consequence**	**Likelihood**	**Risk Level**
		Oversight	Risk of inadequate or no specification and/or communication of purpose for AI use/an identified AI solution.	(i) Planned use of AI is presented as a means for improving efficiency of business, whilst impact on workforce is not noted or explored, resulting in new uncertainty and sense of insecurity among workforce. (ii) A workflow is intended for change to accommodate an AI system, but employees do not see the benefits, but anticipate a threat and resent the change.	Psychological	Management of change	Negligible	Rare	
	Judgement: Determine the payoffs to being right versus being wrong. Consider both false positives and false negatives.	Human condition	Risk of (insufficient consideration given to) unintended consequences of false negatives and false positive.	False negatives or false positive disadvantage or victimise a worker, causing stress, overwork, ergonomic risks, anxiety, boredom, fatigue and burnout, potentially building barriers between people, facilitating harassment or bullying.	Psychological	Work demands	Negligible	Unlikely	
		Human condition	Risk of AI being used out of scope.	A productivity assessment tool designed to improve workflow efficiency is used for penalising or firing people.	Psychological	Organisation justice	Negligible	Possible	
		Human condition	Risk of AI undermining company core values and societal expectations.	A prediction tool improves working conditions of some workers, when impact on remaining workforce is unclear or adverse, undermining the company inclusion and diversity policy.	Psychological	Organisation justice	Negligible	Likely	
		Human condition	Risk of AI system undermining human capabilities.	AI system automates processes, assigning workers to undertake remaining tasks resulting in progressive de-skilling.	Psychological	Role variety	Negligible	Almost Certain	
		Human condition	Risk of trading off the personal flourishing (intrinsic value) in favour of organisational gain (instrumental good).	A workflow management system requires workers to follow machine directions, restricting personal autonomy (time planning, task sequence, speed) in order to prioritise company efficiency.	Psychological	Job control	Moderate	Rare	

A	B	C	D	E	F	G	H	I	J
Main Stages of Development	**AI Canvas**	**Ethics Domains**	**Ethics Risks to WHS**	**Examples – Potential WHS Related Harms**	**Characteristics of Work**	**WHS Hazards and Risks**	**Consequence**	**Likelihood**	**Risk Level**
		Worker safety	Risk of technical failure, human error, financial failure, security breach, data loss, injury, industrial accident/disaster.	Random manual human inspections on machinery are no longer conducted because the predictive maintenance AI didn't foresee a problem (false negative). Consequently, the machine breaks down and results in injury.	Physical, Biomechanical	Physical hazards, Force, Movement, Posture	Moderate	Unlikely	
		Worker safety	Risk of impacting on other processes or essential services affecting workflow or working conditions.	An employee responsible for IT security is inundated with alerts by an AI network intrusion detection system. The false alarm rate is very high, and the bulk of their time is spent manually overriding false positive alerts.	Biomechanical, Cognitive, Psychological	Movement, Information processing load, Complexity and duration, Work demands	Moderate	Possible	
		Oversight	Risk of insufficient/ineffective transparency, contestability and accountability at the design stage and throughout the development process.	Selective workforce consultation fails to record specific concerns not otherwise observed, recognised or shared by those consulted.	Psychological	Managing relationships, Management of change	Moderate	Likely	
	Action: What are the actions that can be chosen?	Human condition	Risk of inequitable or burdensome treatment of workers.	A workflow management system disproportionately, repeatedly or persistently assigns some workers to challenging tasks that others with principally identical roles can thus avoid.	Cognitive	Complexity and duration	Moderate	Almost Certain	
		Human condition	Risk of gaming (reward hacking) of AI system undermining workplace relations.	An automated customer satisfaction survey system encourages repeated feedback on an internal department's performance by splitting support services into multiple tasks with associated case opening and closing tickets.	Psychological	Organisation justice	Extensive	Rare	
		Human condition	Risk of worker attributing intelligence or empathy to AI system greater than appropriate.	A chatbot fails to indicate when the service is automated or undertaking by a human, implying equal capacity to provide effective and conclusive service.	Not applicable		Extensive	Unlikely	
		Human condition	Risk of context stripping from communication between employees.	A productivity tool fails to recognise and is not adjusted in a	Psychological	Supervisor/peer support	Extensive	Possible	

A	B	C	D	E	F	G	H	I	J
Main Stages of Development	**AI Canvas**	**Ethics Domains**	**Ethics Risks to WHS**	**Examples – Potential WHS Related Harms**	**Characteristics of Work**	**WHS Hazards and Risks**	**Consequence**	**Likelihood**	**Risk Level**
				timely fashion to account for, [change in] worker circumstances that affect performance or workplace presence, whilst continuing to provide feedback or directions. An employee's childcare commitment is an example of constraints on workplace presence.					
		Human condition	Risk of worker manipulation or exploitation.	Workers are pitched against another by publicly displaying performance indicators, presenting internal competition as a game whilst seeking to increase output.	Psychological	Managing relationships	Extensive	Likely	
		Human condition	Risk of undue reliance on AI decisions.	A set of quantifiable performance indicators replaces face-to-face worker-supervisor performance reviews, substituting for dialogue and review of challenges and opportunities. Managerial autonomy is replaced by machine authority, and decisions and their impacts are not considered or are not reversible.	Psychological	Organisation justice	Extensive	Almost Certain	
		Worker safety	Risk of adversely affecting worker or general rights (to a safe workplace/physical integrity, pay at right rate/EA, adherence to National Employment Standards, privacy)	An AI analyses the content of emails to determine employee satisfaction and engagement levels. Another AI uses audio analytics to determine stress levels in voices when staff speak to each other in the office.	Psychological	Job control, Supervisor/peer support, Managing relationships, Management of change	Significant	Rare	
		Worker safety	Risk of unnecessary harm, avoidable death or disabling injury/ergonomics.	An AI assigns staff to a roster to ensure all gaps are filled. In achieving this, staff are allocated slots in a fragmented way that is inconvenient to them and increases stress levels.	Physical, Psychological	Physical hazards, Work demands, Job control	Significant	Unlikely	

A	B	C	D	E	F	G	H	I	J
Main Stages of Development	**AI Canvas**	**Ethics Domains**	**Ethics Risks to WHS**	**Examples – Potential WHS Related Harms**	**Characteristics of Work**	**WHS Hazards and Risks**	**Consequence**	**Likelihood**	**Risk Level**
		Worker safety	Risk of physical and psychosocial hazards.	AI causing intensity of work/workload to increase or closer physical proximity of machine tools and worker (e.g. cobots), requiring workspace adjustments to avoid injury. An AI assigns a task to a person without the necessary experience or skill to perform it, because it has not considered the need to acquire new skills.	Physical, Psychological	Physical hazards, Job control, Work demands	Significant	Possible	
		Oversight	Risk of inadequate or closed chain of accountability, reporting and governance structure for AI ethics within the organisation, with limited or no scope for review.	(i) A company CEO fails to appoint a champion for AI ethics and safety. Frequency of WHS incidents increases because AI is not incorporated into WHS. (ii) An employee cannot change a forecast that an AI system has made even if they know it is unlikely to be correct. This may cause stress and resentment because they could be held accountable for something beyond their control.	Cognitive, Psychological	Complexity and duration, Work demands, Job control, Supervisor/peer support	Significant	Likely	
		Oversight	Risk of (lack of process) for triggering human oversight or checks and balances, so that algorithmic decisions cannot be challenged, contested, or improved.	A mid-level manager takes extended stress leave after they are unable to explain to senior management why the AI system keeps wrongly predicting inventory increase because customers are calculated to replace products when, in fact, they are booking repair services.	Psychological	Work demands, Supervisor/peer support	Significant	Almost Certain	

A	B	C	D	E	F	G	H	I	J
Main Stages of Development	**AI Canvas**	**Ethics Domains**	**Ethics Risks to WHS**	**Examples – Potential WHS Related Harms**	**Characteristics of Work**	**WHS Hazards and Risks**	**Consequence**	**Likelihood**	**Risk Level**
		Oversight	Risk of AI shifting responsibility outside existing managerial or company protocols, and channels of internal accountability (via out- or sub-contracting).	Off-the-shelf acquisition of AI leaves user with limited understanding of its utility, condition for reliability, maintenance requirements.	Cognitive, Psychological	Information processing load, Job control	Negligible	Unlikely	
Development	**Outcome**: Choose the measure of performance that you want to use to judge whether you are achieving your outcomes.	Human condition	Risk of chosen outcome measure not aligning with healthy/collegial workplace dynamics.	Efficiency improvements have differential effects across the workforce, improving conditions for some, but not others, or creating or promoting competitive behaviours, undermining collaborations or collegial relations.	Psychological	Organisation justice	Negligible	Rare	
		Human condition	Risk of outcome measure resulting in worker-AI interface adversely affecting the status of a worker/workers in the workplace.	Workers gain exclusive additional benefits or rewards unavailable to others, such as training or earning increases/bonuses (as operators of AI, also to match their greater responsibilities and new core functions to the efficiency and reputation of the business).	Psychological	Organisation justice	Moderate	Rare	
		Worker safety	Risk of performance measures differentially and/or adversely affecting work tasks and processes.	AI tool leads to faster and more precise processing of test samples in a medical lab, also requiring improved storage capacity and speedier throughput-management.	Biomechanical, Psychological	Force, Movement, Posture, Job control	Extensive	Possible	
		Oversight	Risk of workers (not) able to access and/or modify factors driving the outcomes of decisions.	An HR department uses a chatbot which is supposed to answer employees' questions in plain language. An employee feels the answer provided by the chatbot is insufficient, but no one in HR is willing to engage in a dialogue because they see the question as falling inside the domain of the chatbot.	Psychological	Managing relationships, Management of change	Extensive	Possible	

A	B	C	D	E	F	G	H	I	J
Main Stages of Development	**AI Canvas**	**Ethics Domains**	**Ethics Risks to WHS**	**Examples – Potential WHS Related Harms**	**Characteristics of Work**	**WHS Hazards and Risks**	**Consequence**	**Likelihood**	**Risk Level**
	Training: What data do you need on past inputs, actions and outcomes in order to train your AI to generate better predictions?	Human condition	Risk of training data not representing the target domain in the workplace.	Training data for a new system of leave and sick leave projections include only more recent workplace recruits with shorter tenure for whom better contextual data are available.	Psychological	Organisation justice	Moderate	Likely	
		Human condition	Risk of acquisition, collection and analysis of data revealing (confidential) information out of scope of the project.	Training data includes personal (e.g. health) or contextual (e.g. ethnicity) unrelated to the workflow allocation algorithm.	Psychological	Organisation justice	Moderate	Almost Certain	
		Human condition	Risk of data not being fit for purpose.	Training data for a job performance algorithm uses past performance reviews as the outcome measure, which it wants to replace with a more robust and objective assessment tool. The use of an untrusted past performance indicator indicates the data source is possibly unsuitable.	Psychological	Organisation justice	Extensive	Unlikely	
		Worker safety	Risk of cyber security vulnerability.	AI uses staff email and instant messaging data, along with microphone-equipped name badges, to gather data on employee interactions. The business, new to this data collection method, considers insecure storage options for this very personal information.	Psychological	Organisation justice	Moderate	Possible	
		Worker safety	Risk of (in)sufficient consideration given to interconnectivity/interoperability of AI systems.	Multiple data sources need integrating, each quality assessed and assured.	Cognitive, Psychological	Information processing load, Complexity and duration, Work demands	Negligible	Likely	
		Oversight	Risk of inadequate logging of the inputs and outputs of the AI, or incomplete mapping of data origins and lineage, adversely affecting ability to conduct data audits or routine monitoring and evaluation.	A production planning team ends up scheduling work that the production team cannot execute; missing or inadequate documentation means that systemic flaws cannot be	Cognitive, Psychological	Complexity and duration, Work demands, Management of change	Insignificant	Almost Certain	

A	B	C	D	E	F	G	H	I	J
Main Stages of Development	**AI Canvas**	**Ethics Domains**	**Ethics Risks to WHS**	**Examples – Potential WHS Related Harms**	**Characteristics of Work**	**WHS Hazards and Risks**	**Consequence**	**Likelihood**	**Risk Level**
				identified. Blame is shifted onto the AI system and the organisation's procurement department.					
		Oversight	Risk of inadequate testing of AI in a production environment and/or for impact on different (target) populations.	A chatbot copies unacceptable language; an HR recruitment tool rules out women applicants.	Psychological	Organisation justice	Insignificant	Almost Certain	
	Input: What data do you need to generate predictions once you have an AI algorithm trained?	Human condition	Risk of discontinuity of service.	A workforce planning tool omits timely correction for seasonal factors, trends or shocks, leading to a shortage of staff or produce at key times.	Cognitive	Complexity and duration	Negligible	Almost Certain	
		Human condition	Risk of worker unable or unwilling to provide or permit data to be used as input to the AI.	Data training suggests that work injury data could enhance the predictive capability of the algorithm but would require all workers to agree for their injury records to be linked to the model. Some workers fear this may disadvantage them and decline.	Psychological	Management of change	Moderate	Likely	
		Worker safety	Risk of impacting on physical workplace (lay out, design, environmental conditions: temperature, humidity).	New or changing human-machine interface (e.g. cobots) requiring movement-distance control and monitoring.	Physical, Biomechanical	Physical hazards, Force, Movement, Posture	Negligible	Almost Certain	
		Worker safety	Risk of (in)secure data storage and cyber security vulnerability.	Connectedness and size of personal data collection requiring transition from offline to online/cloud data storage, increasing vulnerability during and after transition. Efficiency gain through AI reliant on sustained synchronised data flow from multiple sources to avoid bottlenecks, service disruption or bias.	Cognitive, Psychological	Information processing load, Work demands, Management of change	Insignificant	Likely	

A	B	C	D	E	F	G	H	I	J
Main Stages of Development	**AI Canvas**	**Ethics Domains**	**Ethics Risks to WHS**	**Examples – Potential WHS Related Harms**	**Characteristics of Work**	**WHS Hazards and Risks**	**Consequence**	**Likelihood**	**Risk Level**
		Worker safety	Risk of worker competences and skills (not) meeting AI requirements.	An AI-trained eye-screening unit used to monitor changes in workers' vision resulting from Computer Vision Syndrome is sensitive to light changes. The health assistant, previously using conventional tools of optometry, is aware of the risk of invalid eye scans, but has not been instructed in setting up the instrument to meet the correct lighting conditions.	Cognitive, Psychological	Information processing load, Work demands, Job control	Insignificant	Likely	
		Worker safety	Risk of boundary creep: data collection (not) ceasing outside the workplace.	Employees continuing (or indeed incentivised) to wear Fitbits outside working hours, enabling organisation to gather additional data beyond that originally intended for collection.	Psychological	Organisation justice	Insignificant	Unlikely	
		Oversight	Risk of insufficient worker understanding of safety culture and safe behaviours applied to data and data processes within AI.	(i) Use of multiple data sources increases frequency and pathways of data transmission, with added risks of safety failures; (ii) an AI tool is used to accelerate analytical processes, requiring also increased capacity of safe storage.	Cognitive, Psychological	Information processing load, Management of change	Insignificant	Rare	
		Oversight	Risk of partial disclosure or audit of data uses (e.g. due to commercial considerations, proprietary knowledge).	A worker is asked to incorporate an AI prediction into their decision-making process, but the prediction contradicts their intuition. Because they do not understand how the AI arrived at its prediction the worker chooses to ignore it.	Psychological	Work demands, Job control	Insignificant	Unlikely	

A	B	C	D	E	F	G	H	I	J
Main Stages of Development	**AI Canvas**	**Ethics Domains**	**Ethics Risks to WHS**	**Examples – Potential WHS Related Harms**	**Characteristics of Work**	**WHS Hazards and Risks**	**Consequence**	**Likelihood**	**Risk Level**
Application	**Feedback:** How can you use the outcomes to improve the algorithm?	Human condition	Risk of impacts (not) being reversible.	Workers' on-the-job responsibilities and autonomy are permanently reduced, adversely affecting skills utilisation, on the job satisfaction, workplace status.	Psychological	Role variety	Insignificant	Unlikely	
		Worker safety	Risk of assessment processes requiring review due to new approach or tool.	A new HR recruitment process using AI achieves a more gender-balanced intake of new staff. Do the data input or algorithm require review to maintain this outcome?	Cognitive, Psychological	Information processing load, Complexity and duration, Organisation justice	Insignificant	Unlikely	
		Worker safety	Risk of identifiable personal data retained longer than necessary for the purpose it was collected and/or processed.	Training data retained beyond full AI application, including information used in training but not in final model.	Psychological	Organisation justice	Insignificant	Possible	
		Oversight	Risk of inadequate integration of AI operational management into routine Mechanical & Electrical (M&E) maintenance ensuring AI continues to work as initially specified.	AI operations management requires specialist skills different and in addition to conventional operational process management skills; joint operability required.	Psychological	Role variety	Insignificant	Possible	
		Oversight	Risk of no offline systems or processes in place to test and review veracity of AI predictions/decisions.	An AI tool is used to triage incoming calls to an organisation, but the tool provides incomplete answers unable to resolve the query; dissatisfied client complains.	Psychological	Work demands	Insignificant	Possible	

G: AI WHS Protocol.

AI Ethics Protocol

This protocol accompanies the AI Ethics Scorecard.

Objectives.

The scorecard is intended as a guide for organisations using, planning to use, or exploring the use of Artificial Intelligence (AI) in a workplace. It is designed to assist in identifying contexts or actions that may affect the ethical application of AI. It is based on a set of AI ethics principles endorsed by the Australian Government Department of Industry, Science, Energy and Resources (DISER).

Format.

The scorecard maps steps in the ideation, testing and application of AI (Column A: "Main Stages of Development" and, in more detail, Column B: "AI Canvas") against AI ethics principles (Column C: "Ethics Domains").

For each of these steps and ethics domains, the scorecard identifies potential risks that AI may pose when used in a workplace, potentially affecting workers' health and safety (Column D: "Ethics Risks to WHS"). Examples of such risks are also shown (Column E: "Examples - Potential WHS Related Harms").

The AI Canvas was originally proposed by Ajay Agrawal, Joshua Gans, and Avi Goldfarb.

A link to the AI Canvas can be found here: https://www.predictionmachines.ai/.

The "Risk Domains" are aggregates of originally eight AI Ethics Principles endorsed by DISER, namely:

Human Condition	Worker Safety	Oversight
Human, social and environmental wellbeing Human-centred values Fairness	Privacy protection and security Reliability and safety	Transparency and explainability Contestability Accountability

For definitions, see https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles.

The scorecard also cross-references AI risks with "characteristics of work" (Column F) and "hazards or risks" (Column G) identified in the Safe Work Australia "Principles of Good Work Design Handbook".

Use.

Users may consult the scorecard at and for the relevant steps as identified in the “Main Stages of Development” and the “AI Canvas”. Columns H and I allow users to assess the “Consequence” (Column H) and “Likelihood” (Column I) of each “Ethics Risk to WHS”, using dropdown menus. Assessments result is the display of a “Risk Level” (Column J), using a colour scheme, indicative of the need for active consideration to be given to preventative measures (see back of pages for details).

It is recommended that risk levels be assessed in collaboration with those involved in or likely to be affected by the AI ideation, testing and application.

Listed risks are suggestions for consideration. Not all risks will be relevant in all instances. Not all risks will necessarily capture adverse effects but may indicate instances in which responsible action can avoid or compensate for potential harms. The risk level ratings are suggestive only.

Disclaimer.

As AI develops, risks are likely to change. The scorecard is a generic guide that does not and cannot claim to be comprehensive.

Risk Level Scoring

Traffic Light Legend
Low
Low Medium
Medium
Medium High
High

			Consequence				
		Worker	Negative impact on mood. Staff may be irritated and inconvenienced.	Temporary reduction in productivity and efficiency	Decline in job satisfaction, morale, cohesion, and productivity.	Increase in absenteeism and conflicts at work.	Increase in staff turnover, health care expenditure and worker's compensation claims.
		Organisation	Minimal impact on non-core business operations. The impact can be dealt with by routine operations.	Some impact on business areas in terms of delays and quality. Can be addressed at the operational level.	Reduced performance such as not meeting targets, but organisation's existence is not threatened.	Breakdown of key activities leading to substantial reduced performance. Survival of organisation threatened.	Critical failure preventing core activities from being performed. Survival of organisation threatened.
	Qualitative Likelihood		**Insignificant**	**Negligible**	**Moderate**	**Extensive**	**Significant**
Likelihood	Is expected to occur in most circumstances	**Almost Certain**	Medium	Medium High	High	High	High
	Will probably occur in most circumstances	**Likely**	Low Medium	Medium	Medium High	High	High
	Might occur at some time	**Possible**	Low Medium	Low Medium	Medium	Medium High	Medium High
	Could occur at some time	**Unlikely**	Low	Low	Low Medium	Medium	Medium High
	May occur only in exceptional circumstances	**Rare**	Low	Low	Low	Low Medium	Medium

H: AI WHS Scorecard use example.

An organisation uses various machinery and equipment while delivering its service to customers. It struggles with unplanned downtime costs due to sporadic equipment failure. The interruptions result in revenue loss, component replacement costs, and occasionally even fines for not delivering its service. Currently, the organisation uses a time-based maintenance schedule where a piece of equipment gets maintained and serviced at fixed time intervals whether it needs it or not. The time-based maintenance is labour intensive and ineffective in identifying problems that develop between the scheduled inspections. The organisation wants to address this problem by adopting AI for predictive maintenance. Predictive maintenance involves instrumenting the machinery with sensors to facilitate continuous equipment condition monitoring and predicting future wear and tear. The purpose is to schedule maintenance activity when it is most cost-effective and before the equipment loses performance below a specified threshold. The organisation will use the AI tool to automatically trigger maintenance planning, work order execution, and reporting.

An example AI Canvas that outlines key conceptual dimensions for the predictive maintenance scenario is shown in Table H.1. For each conceptual dimension of the AI Canvas, the organisation would reflect on the ethics risk and assess the workplace hazard risk level.

Table H.1: AI Canvas for a predictive maintenance scenario.

Prediction: Identify the key uncertainty that you would like to resolve. *Does this equipment need to be serviced?*
Judgement: Determine the payoffs to being right versus being wrong. Consider both false positives and false negatives. *If the prediction is correct. then unplanned maintenance can be avoided. Unplanned maintenance is usually costly and disruptive. False positives (incorrectly predicting that, yes, the equipment needs to be serviced, when in fact it does not) will result in spending unnecessary resources, whereas false negatives (incorrectly predicting that, no, the equipment does not need servicing, when in fact it does) will result in the unplanned maintenance that had been hoped to be avoided in the first place.*
Action: What are the actions that can be chosen? *Devote resources to servicing the equipment.*
Outcome: Choose the measure of performance that you want to use to judge whether you are achieving your outcomes. *A reduction in the number of unplanned maintenance events.*
Training: What data do you need on past inputs, actions and outcomes in order to train your AI to generate better predictions? - *Continual observational data for the equipment (e.g. vibration measurements, temperature measurements etc.)* - *Equipment utilization data.* - *Historical maintenance records.* - *Historical equipment failure data.*
Input: What data do you need to generate predictions once you have an AI algorithm trained? *Equipment utilization and measurement data.*
Feedback *Whenever equipment is serviced, or fails, the status of the wear and tear together with all other data is used as new training data for the AI.*

Prediction Dimension

Focusing on the prediction dimension of the AI canvas (“Does this equipment need to be serviced?”), one might identify three high-risk hazards that affect workers in a physical or psychosocial manner. The risks are linked with organisational justice, job-control and physical hazards. For example, if the AI tool is used to trigger maintenance planning, then it is likely to change the allocation of roles. The system may not know (or care) that certain employees have more experience and know-how for maintaining specific equipment. Employees may resent being automatically asked to service less familiar equipment because they may feel that their years of experience are no longer recognised or valued. If the maintenance schedule is solely determined by an AI tool (i.e. there is no room for a technician to schedule a spontaneous maintenance outside of the AI’s schedule) then technicians may also lose control over their job. The loss in job control can lead to an overreliance on the AI system, resulting in diminished due diligence. For example, when the organisation was using a time-based maintenance schedule, even though they were encountering unplanned downtime, the risk of catastrophic failure was low. A prediction-based maintenance schedule may result in catastrophic equipment failure if the AI tools fails to detect a fault and fails to schedule maintenance for a prolonged time.

Table H.2: Scorecard applied to the prediction dimension of the AI Canvas which is associated with the ideation phase.

AI Canvas	Ethics Risks to WHS	Examples – Potential WHS Related Harms	Work Characteristics	Hazard or risk	Consequence	Likelihood	Risk Level
Prediction: Identify the key uncertainty that you would like to resolve.	Risk of using AI when an alternative solution may be more appropriate or humane.	Predicting a worker's physical or mental exhaustion levels for monitoring purposes without instituting strategies to prevent exhaustion in the future.	Psychological	Work demands	Insignificant	Rare	
	Risk of the system displacing rather than augmenting human decisions.	Prediction tool changes allocation of roles and responsibilities, with some worker assigned higher status roles, others relegated to lower status roles, or facing redundancy.	Psychological	Organisation justice	Moderate	Likely	
	Risk of augmenting or displacing human decisions with differential impact on workers who are directly or indirectly affected.	A warehouse manager for a toy company ignores feedback from order fulfilment staff that a popular toy is about to sell out during the pre-Christmas period, because the AI stock control tool predicted adequate stock levels. Staff are disempowered and demotivated.	Biomechanical	Job control	Moderate	Likely	
	Risk of the resolution of uncertainty affecting ethical, moral or social principles.	Predicting the health/health trajectory of an employee, such as likelihood of pregnancy, may contravene right to privacy or social/moral convention.	Psychological	Organisation justice	Insignificant	Rare	
	Risk of overconfidence in or overreliance on AI system, resulting in loss of/diminished due diligence.	After a six-month 'break-in' period without incidents at a new AI-enabled plant, preventive safety measures are no longer prioritised; new employees are no longer trained in PPE requirements.	Cognitive, Physical	Physical hazards, Information processing load, Complexity and duration	Significant	Likely	
	Risk of inadequate or no specification and/or communication of purpose for AI use/an identified AI solution.	(i) Planned use of AI is presented as a means for improving efficiency of business, whilst impact on workforce is not noted or explored, resulting in new uncertainty and sense of insecurity among workforce. (ii) A workflow is intended for change to accommodate an AI system, but employees do not see the benefits, but anticipate a threat and resent the change.	Psychological	Management of change	Moderate	Likely	

Judgement Dimension

On the judgement dimension of the AI Canvas, the likely risks are similar to the hazards identified on the prediction dimension. If there is no oversight or review of how the AI tool assigns workers to service equipment one may fail to notice that employees are not given the opportunity to service diverse machinery and equipment. They may be inadvertently constrained to work with a subset of equipment and may experience progressive deskilling. These risks are linked with role variety and job control. There is also a risk of physical harm because scheduled human inspections might not be conducted on machines that the AI tool incorrectly considers as operating normally. Conversely, there is a risk of biomechanical harm if the AI system makes substantial false positive predictions and overburdens technicians with service assignments.

Table H.3: Scorecard applied to the judgement dimension of the AI Canvas which is associated with the ideation phase.

AI Canvas	Ethics Risks to WHS	Examples – Potential WHS Related Harms	Work Characteristics	Hazard or Risk	Consequence	Likelihood	Risk Level
Judgement: Determine the payoffs to being right versus being wrong. Consider both false positives and false negatives.	Risk of (insufficient consideration given to) unintended consequences of false negatives and false positive.	False negatives or false positive disadvantage or victimise a worker, causing stress, overwork, ergonomic risks, anxiety, boredom, fatigue and burnout, potentially building barriers between people, facilitating harassment or bullying.	Psychological	Work demands	Moderate	Likely	
	Risk of AI being used out of scope.	A productivity assessment tool designed to improve workflow efficiency is used for penalising or firing people.	Psychological	Organisation justice	Negligible	Rare	
	Risk of AI undermining company core values and societal expectations.	A prediction tool improves working conditions of some workers, when impact on remaining workforce is unclear or adverse, undermining the company inclusion and diversity policy.	Psychological	Organisation justice	Insignificant	Rare	
	Risk of AI system undermining human capabilities.	AI system automates processes, assigning workers to undertake remaining tasks resulting in progressive de-skilling.	Psychological	Role variety	Extensive	Possible	
	Risk of trading off the personal flourishing (intrinsic value) in favour of organisational gain (instrumental good).	A workflow management system requires workers to follow machine directions, restricting personal autonomy (time planning, task sequence, speed) in order to prioritise company efficiency.	Psychological	Job control	Moderate	Possible	
	Risk of technical failure, human error, financial failure, security breach, data loss, injury, industrial accident/disaster.	Random manual human inspections on machinery are no longer conducted because the predictive maintenance AI didn't foresee a problem (false negative). Consequently, the machine breaks down and results in injury.	Physical, Biomechanical	Physical hazards, Force, Movement, Posture	Extensive	Possible	
	Risk of impacting on other processes or essential services affecting workflow or working conditions.	An employee responsible for IT security is inundated with alerts by an AI network intrusion detection system. The false alarm rate is very high, and the bulk of their time is spent manually overriding false positive alerts.	Biomechanical, Cognitive, Psychological	Movement, Information processing load, Complexity and duration, Work demands	Moderate	Possible	
	Risk of insufficient/ineffective transparency, contestability and accountability at the design stage and throughout the development process.	Selective workforce consultation fails to record specific concerns not otherwise observed, recognised or shared by those consulted.	Psychological	Managing relationships, Management of change	Negligible	Possible	

Action Dimension

An appraisal of the action dimension of the AI Canvas may reveal that the AI tool could negatively impact on the complexity and duration of work. If the way the AI tool schedules maintenance jobs is not clearly communicated and reviewed, some employees may be required to do a disproportionate amount of work. For example, some equipment may require more frequent servicing and the technicians the system associated with that equipment will be required to work more than technicians associated with equipment that rarely breaks down. The AI tool may also assign a task to a person without the necessary experience or skill to perform it, because it has not considered the need to acquire new skills. In general, there is a substantial risk that the algorithmic decisions cannot be challenged, and that the organisation fails to introduce an explicit mechanism for triggering human oversight.

Table H.4: Scorecard applied to the action dimension of the AI Canvas which is associated with the ideation phase.

AI Canvas	Ethics Risks to WHS	Examples – Potential WHS Related Harms	Work Characteristics	Hazard or risk	Consequence	Likelihood	Risk Level
Action: What are the actions that can be chosen?	Risk of inequitable or burdensome treatment of workers.	A workflow management system disproportionately, repeatedly or persistently assigns some workers to challenging tasks that others with principally identical roles can thus avoid.	Cognitive	Complexity and duration	Extensive	Likely	
	Risk of gaming (reward hacking) of AI system undermining workplace relations.	An automated customer satisfaction survey system encourages repeated feedback on an internal department's performance by splitting support services into multiple tasks with associated case opening and closing tickets.	Psychological	Organisation justice	Negligible	Rare	
	Risk of worker attributing intelligence or empathy to AI system greater than appropriate.	A chatbot fails to indicate when the service is automated or undertaking by a human, implying equal capacity to provide effective and conclusive service.	Not applicable		Insignificant	Rare	
	Risk of context stripping from communication between employees.	A productivity tool fails to recognise and is not adjusted in a timely fashion to account for, [change in] worker circumstances that affect performance or workplace presence, whilst continuing to provide feedback or directions. An employee's childcare commitment is an example of constraints on workplace presence.	Psychological	Supervisor/peer support	Moderate	Unlikely	
	Risk of worker manipulation or exploitation.	Workers are pitched against another by publicly displaying performance indicators, presenting internal competition as a game whilst seeking to increase output.	Psychological	Managing relationships	Moderate	Rare	
	Risk of undue reliance on AI decisions.	A set of quantifiable performance indicators replaces face-to-face worker-supervisor performance reviews, substituting for dialogue and review of challenges and opportunities. Managerial autonomy is replaced by machine authority, and decisions and their impacts are not considered or are not reversible.	Psychological	Organisation justice	Moderate	Likely	
	Risk of adversely affecting worker or general rights (to a safe workplace/physical integrity, pay at right rate/EA, adherence to National Employment Standards, privacy)	An AI analyses the content of emails to determine employee satisfaction and engagement levels. Another AI uses audio analytics to determine stress levels in voices when staff speak to each other in the office.	Psychological	Job control, Supervisor/peer support, Managing relationships, Management of change	Negligible	Rare	

AI Canvas	Ethics Risks to WHS	Examples – Potential WHS Related Harms	Work Characteristics	Hazard or risk	Consequence	Likelihood	Risk Level
	Risk of unnecessary harm, avoidable death or disabling injury/ergonomics.	An AI assigns staff to a roster to ensure all gaps are filled. In achieving this, staff are allocated slots in a fragmented way that is inconvenient to them and increases stress levels.	Physical, Psychological	Physical hazards, Work demands, Job control	Extensive	Unlikely	
	Risk of physical and psychosocial hazards.	AI causing intensity of work/workload to increase or closer physical proximity of machine tools and worker (e.g. cobots), requiring workspace adjustments to avoid injury. An AI assigns a task to a person without the necessary experience or skill to perform it, because it has not considered the need to acquire new skills.	Physical, Psychological	Physical hazards, Job control, Work demands	Significant	Possible	
	Risk of inadequate or closed chain of accountability, reporting and governance structure for AI ethics within the organisation, with limited or no scope for review.	(i) A company CEO fails to appoint a champion for AI ethics and safety. Frequency of WHS incidents increases because AI is not incorporated into WHS. (ii) An employee cannot change a forecast that an AI system has made even if they know it is unlikely to be correct. This may cause stress and resentment because they could be held accountable for something beyond their control.	Cognitive, Psychological	Complexity and duration, Work demands, Job control, Supervisor/peer support	Significant	Likely	
	Risk of (lack of process) for triggering human oversight or checks and balances, so that algorithmic decisions cannot be challenged, contested, or improved.	A mid-level manager takes extended stress leave after they are unable to explain to senior management why the AI system keeps wrongly predicting inventory increase because customers are calculated to replace products when, in fact, they are booking repair services.	Psychological	Work demands, Supervisor/peer support	Significant	Likely	
	Risk of AI shifting responsibility outside existing managerial or company protocols, and channels of internal accountability (via out- or sub-contracting).	Off-the-shelf acquisition of AI leaves user with limited understanding of its utility, condition for reliability, maintenance requirements.	Cognitive, Psychological	Information processing load, Job control	Negligible	Rare	

Outcome Dimension

After studying the outcome dimension of the AI Canvas, one might discover no high-impact hazards. The main concern is that technicians may want to understand how the AI tool is making its predictions and constructing its schedule and may not have access to that information. Failure to address this issue may complicate the change management process.

TableH.5: Scorecard applied to the outcome dimension of the AI Canvas which is associated with the development phase.

AI Canvas	Ethics Risks to WHS	Examples – Potential WHS Related Harms	Work Characteristics	Hazard or Risk	Consequence	Likelihood	Risk Level
Outcome: Choose the measure of performance that you want to use to judge whether you are achieving your outcomes.	Risk of chosen outcome measure not aligning with healthy/collegial workplace dynamics.	Efficiency improvements have differential effects across the workforce, improving conditions for some, but not others, or creating or promoting competitive behaviours, undermining collaborations or collegial relations.	Psychological	Organisation justice	Negligible	Rare	
	Risk of outcome measure resulting in worker-AI interface adversely affecting the status of a worker/workers in the workplace.	Workers gain exclusive additional benefits or rewards unavailable to others, such as training or earning increases/bonuses (as operators of AI, also to match their greater responsibilities and new core functions to the efficiency and reputation of the business).	Psychological	Organisation justice	Insignificant	Rare	
	Risk of performance measures differentially and/or adversely affecting work tasks and processes.	AI tool leads to faster and more precise processing of test samples in a medical lab, also requiring improved storage capacity and speedier throughput-management.	Biomechanical, Psychological	Force, Movement, Posture, Job control	Negligible	Rare	
	Risk of workers (not) able to access and/or modify factors driving the outcomes of decisions.	An HR department uses a chatbot which is supposed to answer employees' questions in plain language. An employee feels the answer provided by the chatbot is insufficient, but no one in HR is willing to engage in a dialogue because they see the question as falling inside the domain of the chatbot.	Psychological	Managing relationships, Management of change	Negligible	Possible	

Training Data Dimension

Thinking about the training dimension of the AI Canvas, the principal risk is that the data collected for training the fault prediction is inadequate. Substantial effort will be required to instrument all the equipment, to create the data pipelines necessary to amass the training data and to verify the veracity and completeness of the acquired data. The performance of the system hinges upon the data quality.

Table H.6: Scorecard applied to the outcome dimension of the AI Canvas which is associated with the development phase.

AI Canvas	Ethics Risks to WHS	Examples – Potential WHS Related Harms	Work Characteristics	Hazard or Risk	Consequence	Likelihood	Risk Level
Training: What data do you need on past inputs, actions and outcomes in order to train your AI to generate better predictions?	Risk of training data not representing the target domain in the workplace.	Training data for a new system of leave and sick leave projections include only more recent workplace recruits with shorter tenure for whom better contextual data are available.	Psychological	Organisation justice	Moderate	Possible	
	Risk of acquisition, collection and analysis of data revealing (confidential) information out of scope of the project.	Training data includes personal (e.g. health) or contextual (e.g. ethnicity) unrelated to the workflow allocation algorithm.	Psychological	Organisation justice	Insignificant	Rare	
	Risk of data not being fit for purpose.	Training data for a job performance algorithm uses past performance reviews as the outcome measure, which it wants to replace with a more robust and objective assessment tool. The use of an untrusted past performance indicator indicates the data source is possibly unsuitable.	Psychological	Organisation justice	Moderate	Possible	
	Risk of cyber security vulnerability.	AI uses staff email and instant messaging data, along with microphone-equipped name badges, to gather data on employee interactions. The business, new to this data collection method, considers insecure storage options for this very personal information.	Psychological	Organisation justice	Moderate	Unlikely	
	Risk of (in)sufficient consideration given to interconnectivity/ interoperability of AI systems.	Multiple data sources need integrating, each quality assessed and assured.	Cognitive, Psychological	Information processing load, Complexity and duration, Work demands	Significant	Likely	
	Risk of inadequate logging of the inputs and outputs of the AI, or incomplete mapping of data origins and lineage, adversely affecting ability to conduct data audits or routine monitoring and evaluation.	A production planning team ends up scheduling work that the production team cannot execute; missing or inadequate documentation means that systemic flaws cannot be identified. Blame is shifted onto the AI system and the organisation's procurement department.	Cognitive, Psychological	Complexity and duration, Work demands, Management of change	Moderate	Likely	
	Risk of inadequate testing of AI in a production environment and/or for impact on different (target) populations.	(i) A chatbot copies unacceptable language. (ii) An HR recruitment tool rules out women applicants.	Psychological	Organisation justice	Moderate	Possible	

Input dimension

Whenever one is deploying a suite of interconnected (Internet of Things, or IoT) devices one must consider the cybersecurity implications. Since the AI tool will base its predictions on the data provided by the various sensors, if a hacker manages to compromise a sensing device, they can indirectly take control of the organisation's maintenance schedule. A hacker could manipulate the data stream and make the AI tool predict a fault when none occurred, or vice-versa. Either way, they can cause substantial financial losses and even potential physical harm if they allow machinery to reach catastrophic failure.

Table H.7: Scorecard applied to the input dimension of the AI Canvas which is associated with the development phase.

AI Canvas	Ethics Risks to WHS	Examples – Potential WHS Related Harms	Work Characteristics	Hazard or Risk	Consequence	Likelihood	Risk Level
Input: What data do you need to generate predictions once you have an AI algorithm trained?	Risk of discontinuity of service.	A workforce planning tool omits timely correction for seasonal factors, trends or shocks, leading to a shortage of staff or produce at key times.	Cognitive	Complexity and duration	Insignificant	Rare	
	Risk of worker unable or unwilling to provide or permit data to be used as input to the AI.	Data training suggests that work injury data could enhance the predictive capability of the algorithm but would require all workers to agree for their injury records to be linked to the model. Some workers fear this may disadvantage them and decline.	Psychological	Management of change	Insignificant	Rare	
	Risk of impacting on physical workplace (lay out, design, environmental conditions: temperature, humidity).	New or changing human-machine interface (e.g. cobots) requiring movement-distance control and monitoring.	Physical, Biomechanical	Physical hazards, Force, Movement, Posture	Insignificant	Rare	
	Risk of (in)secure data storage and cyber security vulnerability.	Connectedness and size of personal data collection requiring transition from offline to online/cloud data storage, increasing vulnerability during and after transition. Efficiency gain through AI reliant on sustained synchronised data flow from multiple sources to avoid bottlenecks, service disruption or bias.	Cognitive, Psychological	Information processing load, Work demands, Management of change	Extensive	Likely	
	Risk of worker competences and skills (not) meeting AI requirements.	An AI-trained eye-screening unit used to monitor changes in workers' vision resulting from Computer Vision Syndrome is sensitive to light changes. The health assistant, previously using conventional tools of optometry, is aware of the risk of invalid eye scans, but has not been instructed in setting up the instrument to meet the correct lighting conditions.	Cognitive, Psychological	Information processing load, Work demands, Job control	Insignificant	Rare	
	Risk of boundary creep: data collection (not) ceasing outside the workplace.	Employees continuing (or indeed incentivised) to wear Fitbits outside working hours, enabling organisation to gather additional data beyond that originally intended for collection.	Psychological	Organisation justice	Insignificant	Rare	
	Risk of insufficient worker understanding of safety culture and safe behaviours applied to data and data processes within AI.	(i) Use of multiple data sources increases frequency and pathways of data transmission, with added risks of safety failures; (ii) an AI tool is used to accelerate analytical processes, requiring also increased capacity of safe storage.	Cognitive, Psychological	Information processing load, Management of change	Negligible	Rare	
	Risk of partial disclosure or audit of data uses (e.g. due to commercial considerations, proprietary knowledge).	A worker is asked to incorporate an AI prediction into their decision-making process, but the prediction contradicts their intuition. Because they do not understand how the AI arrived at its prediction the worker chooses to ignore it.	Psychological	Work demands, Job control	Insignificant	Rare	

Feedback

Upon contemplating the feedback dimension of the AI Canvas, one might realise that there is a significant risk that equipment maintenance could grind to a halt if the AI tool went offline for whatever reason. Therefore, the organisation may want to have a backup plan for managing the maintenance schedule. Unless the organisation has a process in place to test and review the veracity of the AI predictions, there is a danger that the performance of the system may stagnate without anyone noticing. Another risk is that the sensors used to monitor the machinery may themselves fail. One will need to ensure that all sensors are replaced with the same model and version that was used to train the system. If the data the AI tool ingests are not of the same kind that it was trained on, its prediction accuracy is likely to be poor.

Table H.8: Scorecard applied to the feedack dimension of the AI Canvas which is associated with the application phase.

AI Canvas	Ethics Risks to WHS	Examples - Potential WHS Related Harms	Work Characteristics	Hazard or Risk	Consequence	Likelihood	Risk Level
Feedback: How can you use the outcomes to improve the algorithm?	Risk of impacts (not) being reversible.	Workers' on-the-job responsibilities and autonomy are permanently reduced, adversely affecting skills utilisation, on the job satisfaction, workplace status.	Psychological	Role variety	Insignificant	Rare	
	Risk of assessment processes requiring review due to new approach or tool.	A new HR recruitment process using AI achieves a more gender-balanced intake of new staff. Do the data input or algorithm require review to maintain this outcome?	Cognitive, Psychological	Information processing load, Complexity and duration, Organisation justice	Insignificant	Rare	
	Risk of identifiable personal data retained longer than necessary for the purpose it was collected and/or processed.	Training data retained beyond full AI application, including information used in training but not in final model.	Psychological	Organisation justice	Insignificant	Rare	
	Risk of inadequate integration of AI operational management into routine Mechanical & Electrical (M&E) maintenance ensuring AI continues to work as initially specified.	AI operations management requires specialist skills different and in addition to conventional operational process management skills; joint operability required.	Psychological	Role variety	Extensive	Possible	
	Risk of no offline systems or processes in place to test and review veracity of AI predictions/decisions.	An AI tool is used to triage incoming calls to an organisation, but the tool provides incomplete answers unable to resolve the query; dissatisfied client complains.	Psychological	Work demands	Significant	Likely	