NSW GOVERNMENT

# AI WHS Scorecard

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.

Prepared by:

Dr Andreas Cebulla[1]
Dr Zygmunt Szpak[2]
Dr Genevieve Knight[3]
Dr Catherine Howell[4]
Dr Sazzad Hussain[5]

May, 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

# About the AI WHS Scorecard

The AI WHS Scorecard has been developed wiith the Centre for Work Health and Safety as part of the research project, "Ethical Use of Artificial Intelligence in the Workplace" (2020-2021).

The Scorecard is intended to help organisations to address new potential work health and safety risks to workers in a workplace using or exploring the use of Artificial Intelligence (AI).

The AI WHS Scorecard helps businesses manage workplace health and safety risks when introducing and using emerging AI technologies, by establishing an AI implementation risk assessment process that is consistent with the Australian Government endorsed AI Ethics Principles and Safe Work Australia standards for managing workplace health and safety hazards.

The AI WHS Scorecard is accompanied by a Protocol that explains its design features, provides an explanation of how it was developed, and explains its intended use. The Protocol is included in the research report that, together with the evidence-based research insights in this report, establishes the actions businesses can take to consider the safety of their workforce when introducing AI processes. Read the Protocol here – Protocol accompanying the AI WHS Scorecard (Appendix G).

# Final AI WHS Scorecard

*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 | |

| 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 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 | |
| | | 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 | 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 | Psychological | Organisation justice | Negligible | 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 |
| | | | | workers, when impact on remaining workforce is unclear or adverse, undermining the company inclusion and diversity policy. | | | | | |
| | | 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 | |
| | | 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 | Selective workforce consultation fails to record specific concerns not otherwise observed, | Psychological | Managing relationships, Management of change | Moderate | 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 |
| | | | design stage and throughout the development process. | recognised or shared by those consulted. | | | | | |
| | **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 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 | 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 |
| | | | | 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 | 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 |
| | | | | them and increases stress levels. | | | | | |
| | | 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.<br>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.<br>(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 | 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 |
| | | | | wrongly predicting inventory increase because customers are calculated to replace products when, in fact, they are booking repair services. | | | | | |
| | | 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 | |

| 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 |
| 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 | |

| 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 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 | |
| | **Training**: What data do you need on past inputs, actions and outcomes in | 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 | Training data includes personal (e.g. health) or | Psychological | Organisation justice | Moderate | 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 |
| | | | (confidential) information out of scope of the project. | contextual (e.g. ethnicity) unrelated to the workflow allocation algorithm. | | | | | |
| | | 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 | |

| 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 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 | Insignificant | Almost Certain | |
| | | 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 | Psychological | Management of change | Moderate | 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 |
| | | | | may disadvantage them and decline. | | | | | |
| | | 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 | |
| | | 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 | Cognitive, Psychological | Information processing load, Work demands, Job control | 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 |
| | | | | up the instrument to meet the correct lighting conditions. | | | | | |
| | | 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 | 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 |
| | | | | worker chooses to ignore it. | | | | | |

| 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 | |