Building a Trauma-Informed Algorithmic Assessment Toolkit

Suvradip Maitra, Lyndal Sleep,
Suzanna Fay & Paul Henman

ADM +S ARC Centre of Excellence for Automated Decision-Making and Society

The University Of Queensland AUSTRALIA
CREATE CHANGE

CQUniversity AUSTRALIA

Contents

Tables

Figures

Appendices

Biographies

Suvradip Maitra

Suvradip Maitra is a practising lawyer and researcher in the ethics of AI, data and algorithms. He is currently a Senior Research Assistant on the ARC Centre of Excellence for Automated Decision Making and Society (ADM+S) affiliated 'Trauma-Informed AI' project at the University of the Queensland. He is also completing a LLM at the University of Melbourne. He graduated with First Class Honours from the University of Queensland with a Bachelor of Science/LLB (Hons) majoring in Physics. Suvradip has been involved in various projects researching the impact of technology on society, including at Cambridge University's Leverhulme Centre for the Future of Intelligence, Harvard University's Berkman Klein Centre for Internet and Society, and Global Catastrophic Risk Institute. He has presented his work at conferences including the Many Worlds of AI Conference at the University of Cambridge, and the Sustainable AI Conference at the University of Bonne. His research has been published in the *Australian Law Journal* (2021) and *Proceedings of the AI and Ethics Society Conference* (2020).

Lyndal Sleep

Lyndal Sleep is a Senior Lecturer in the Queensland Centre for Domestic and Family Violence Research at Central Queensland University, and previously a Postdoctoral Research Fellow at the ARC Centre of Excellence for ADM+S at the University of Queensland. There she researched women, violence, and technology, aiming to improve the lives and life chances of women in situations of intersectional disadvantage by articulating gendered harms and amplifying women's voices. Lyndal also works with advocacy groups to make concrete change in women's lives through collaborative research. Her research has covered domestic violence and the couple rule in social security law, and domestic violence and Centrelink debt. Lyndal's academic background spans social work, social science, sociology, technology and society studies, and law. Lyndal has published in leading national and international journals, including Qualitative Inquiry and Critical Social Policy, and is co-editor of the Journal of Social Inclusion.

Suzanna Fay

Suzanna Fay is a Senior Lecturer in Criminology at the University of Queensland School of Social Science. Her work centres around three themes: 1. the comparative context of crime; 2. how perceptions of gun regulation by police, dealers, and the community influence debate and enforcement of Australia's gun laws; 3. Perceptions of child maltreatment and abuse and its consequences for reporting, monitoring, and court outcomes for children and families. Underscoring all three themes are sociological questions of race and ethnic stratification, and how perceptions of crime influence individual actions.

Paul Henman

Paul Henman is Professor of Digital Sociology and Social Policy at the University of Queensland, Brisbane, Australia. He is also a Chief Investigator of the Australian Research Council Centre of Excellence for Automated Decision Making and Society. Spanning over twenty years, Paul's work focuses on the use of information and digital technologies in government and social services and their implications for changing modes of public governance, and forms of citizen-state power. His publications include *Digital Government in the age of disruptions* (2024, with John Halligan), *Performing the State* (2017, with Alison Gable) and *Governing Electronically* (2010).

Glossary

Term	Meaning
Algorithmic system	An algorithmic system is a machine-based system that can influence the environment by producing an output (predictions, recommendations or decisions) for a given set of objectives. It uses machine and/or human-based data and inputs to (i) perceive real and/or virtual environments; (ii) abstract these perceptions into models through analysis in an automated manner (e.g., with machine learning), or manually; and (iii) use model inference to formulate options for outcomes. Algorithmic systems are designed to operate with varying levels of autonomy. Algorithmic system includes artificial intelligence (AI) and automated decision-making systems (ADMs).
Algorithm supported service	Algorithmic supported service refers to the use of knowledge generated by algorithms or algorithmic systems to execute or directly enact or inform service delivery
Artificial Intelligence (AI)	Artificial Intelligence refers to automated processes that have act with a high level of sophistication that have been build using Machine Learning techniques, rather than traditional human coded algorithms.
Auditing	A broad approach focused on assessing the concordance of our unit of analysis with principles of trauma informed practice
Machine Learning	Machine Learning refers to a process for designing algorithms or models of the world in which the algorithm automatically adjusts internal variables to achieve a desired outcome. Machine Learning can be supervised, unsupervised or hybrid.
Service user	The recipient of a social service
Social services	Services provided by public and/or private actors to societal groups typically to enhance human wellbeing (e.g. disability services, social security, age or personal care, education), though can involve coercive state power (e.g. child protection, parole).
Trauma	Individual trauma results from an event, series of events, or set of circumstances that is experienced by an individual as physically or emotionally harmful or life threatening and that has lasting adverse effects on the individual's functioning and mental, physical, social, emotional, or spiritual well-being. A non-exhaustive list of types of trauma includes individual trauma, developmental trauma, secondary and vicarious, interpersonal trauma, group, community or collective trauma, racial trauma, cultural trauma, historical trauma, natural trauma, human-caused trauma and mass trauma

Executive Summary

Artificial Intelligence and automated processes provide considerable promise to enhance human wellbeing by fully automating or co-producing services with human service providers. Concurrently, if not well considered, automation also provides ways in which to generate harms at scale and speed. To address this challenge, much discussion to date has focused on principles of ethical AI and accountable algorithms with a groundswell of early work seeking to translate these into practical frameworks and processes to ensure such principles are enacted. AI risk assessment frameworks to detect and evaluate possible harms is one dominant approach, as are a growing body of AI audit frameworks, with concomitant emerging governmental and organisational regulatory settings, and associate professionals.

The research outlined in this report took a different approach. Building on work in social services on trauma informed practice, we identified key principles and a practical framework that framed AI design, development and deployment as a reflective, constructive exercise that resulting in algorithmic supported services to be cognisant and inclusive of the diversity of human experience, and particularly those who have experienced trauma. Our study resulted in a practical, co-designed, piloted **Trauma Informed Algorithmic Assessment Toolkit** (hereafter Toolkit).

This Toolkit has been designed to assist organisations in their use of automation in service delivery at any stage of their automation journey: ideation; design; development; piloting; deployment or evaluation. While of particular use for social service organisations working with people who may have experienced past trauma, the tool will be beneficial for any organisation wanting to ensure safe, responsible and ethical use of automation and AI.

Trauma is the emotional, psychological and physiological residue left over from heightened stress that accompanies experiences of threat, violence and life-challenging events. Trauma impacts behaviour, life chances, mental health and relationships. Trauma can make accessing services and resources that can be helpful and healing difficult, further entrenching exclusion, compounding harm and placing survivors at a greater risk of experiencing additional trauma. Understanding socio-political and cultural context of trauma is essential to a 'trauma-informed approach'.

Our Toolkit is referred to as an '**assessment**', rather than an 'audit'. Although 'audit' is widely used in relation to ethical and responsible AI, audits typically imply independent, compliance-based, often pass/fail evaluations. Our Toolkit approach encourages reflection. It is designed to be used within organisations at any stage of design, development, deployment or evaluation of AI or automation in service delivery. It is recommended that the Toolkit be used by a multi-professional team with input from service managers, human service personnel, computer professionals and service user representatives.

To produce the Toolkit, a four-stage research and design process was deployed:

1. Key concepts in both the trauma informed practice literature and the ethical AI literature were reviewed and mapped across the two domains. This resulted in **five core trauma informed principles** underpinning the conceptual design of how to understand Trauma Informed AI: Empowerment and Choice; Collaboration; Trust and Transparency; Safety; and Intersectionality.
2. Through an interactive process, a draft set of assessment questions were identified. Toolkit Items from practical trauma-informed organisational audit frameworks and ethical and accountable AI audit frameworks were identified and modified to reflect our Toolkit's unit of analysis, **an algorithmic enabled service**, which may be fully or partially automated. Real cases of AI and algorithmic harms were used to identify issues to which pertinent questions might be asked.
3. The resulting draft Toolkit was then iteratively, co-designed through two sequential 180-minute, interactive workshops with 18 professionals working in social services organisations delivering services to people who are likely to have experienced trauma, such as in domestic

violence/abuse, child abuse/neglect and homelessness services. Key areas of feedback were: Understanding service users' lived experience of trauma; Strengths-based framing; Capacity building around algorithmic systems and trauma informed principles; Summarising the Toolkit responses; and Useability and engagement.

4. The revised Toolkit was then piloted using three internationally recognised and well documented case studies: Australia's Robodebt scheme; Allegheny County's Family Screening Tool; and the UK's National Eating Disorders Association TESSA chatbot.

The resulting 40-page Toolkit (attached as Appendix A) consists of five sections:

- **Knowing your Organisation, Service and Algorithmic System** guides Toolkit users through the key concepts and principles of the Toolkit and then asks them to identify the motivations for and purposes of using an algorithmic system in service delivery, what it involves and how it relates to organisational objectives;
- **Critical Considerations** consists of 10 core questions designed to cover the most crucial components in delivering a Trauma-informed algorithmic-enabled service. This section can be undertaken as a cut-down version of the Toolkit should timing and resourcing be constrained;
- **General** consists of 49 in-depth questions covering four different domains: Service processes, procedures and plans; algorithmic system; service user engagement and involvement; human algorithm interaction;
- **Types of algorithmic systems** provides questions that are posed by different types of AI or algorithmic systems that are increasingly found in service delivery: Chatbots; Recommender Systems; Identification and Recognition Systems; Risk Assessment Systems; Detection Systems; and Goal Driven Optimisation; and
- **Prioritising Areas for Improvement, Next Steps and Further Resources** guides Toolkit users through the summative review and actioning process. Guidance on risk assessment is provided, if desired. The Toolkit is deliberately not designed to produce summary metrics.

Most questions are framed as 3-point Lickert scale statements beginning "For the algorithm supported service, to what extent…", with "Not at all", "To some extent" and "To a great extent" as the possible responses.

Research participants were strongly supportive of the need for such a Toolkit, seeing it as a much-needed and necessary resource for social service organisations as they increasingly make use of AI and automation for service delivery.

Future development of this internationally novel Toolkit should focus on creating an online, dynamic Toolkit, accompanied with instructional support and educational resources. It adds to and supplements emerging AI audits, frameworks, regulatory settings and professional practice. A key innovation is to move beyond a focus on digital (or data) harms with its attendant risk assessment methodology, to one that positively framed in being trauma aware.

Introduction

Algorithmic systems are progressively being developed and deployed in social services, including welfare payments, child protection, disability services, housing and homelessness services, and employment services (P. W. F. Henman, 2022). Social service users have often experienced vulnerability and are more likely to have not sought therapeutic services after experiencing harm/abuse. It is critical that algorithmic systems are designed and operate in a way that appropriately takes account of the possibility of users' prior experience of trauma, but also does not generate trauma (Chen et al., 2022). Examples of these challenges include Australia's Robodebt system that issued hundreds of thousands of unlawful automated debt notices, many of them to highly vulnerable persons some of whom suicided (Whiteford, 2021). In child abuse/neglect cases, AI-informed risk prediction can overcompensate for adults who as a child had contact with child prediction systems, forcing them to relive past trauma, causing significant re-traumatisation distress (Gillingham, 2021). AI enacted processes in DFV services can mimic the abusive tactics of domestic violence perpetrators (M. Harris & Fallot, 2001), such as lack of choice, sharing of information without permission, minimising lived experience of harm and exclusion from other essential services (Sleep, 2023; Woodlock et al., 2020). While AI systems can also be used to prevent further trauma (Emms et al., 2014; B. Harris et al., 2021), if poorly designed, the digital experience of trauma may impede social services systems' ability to help survivors and risks causing additional harm.

Trauma informed principles are recognised as important in designing and delivering health, medical and social services (Bowen & Murshid, 2016; Raja et al., 2015), to ensure services help rather than harm. Bowen and Murshid (2016), for example, outline six core principles of trauma-informed practice: safety, trustworthiness and transparency, collaboration, empowerment, choice, and intersectionality. Moreover, there are now practical toolkits for developing and analysing services to ensure their organisational operations and decision-making processes are consistent with trauma informed principles (Guarino et al., 2009; Henderson et al., 2018). However, the processes used in social services practice to audit the trauma informed quality of services are not designed with digital services delivery in mind. This means that as services delivery increasingly becomes dependent on digital technology, and AI in particular, current trauma informed audits are inaccurate and failing to fully assess the safety, trust, choice, collaboration and empowerment of their services. Trauma informed audits that incorporate AI are essential for accurate evaluation of services.

While ideas about trauma have yet to be incorporated into building and using AI, there is a large and deep body of work seeking to ensure AI is built and operated ethically, legally and responsibly (Dubber et al., 2020; Fjeld et al., 2020). To embed and enact these concepts and principles, practical AI audit tools have also been developed (Mökander, 2023). Just as AI auditing frameworks and practices can assess the extent to which an AI may lead to decisions or actions that are biased or unethical, a trauma-informed AI audit framework can be developed to assess the extent to which an AI's decisions may generate cause or re-trigger trauma for users. Accordingly, by bringing together literature on trauma informed auditing and AI auditing, this project aimed to co-design a trauma-informed algorithmic assessment framework with service providers and apply the framework to a number of case studies.

Through a four stage process a Trauma Informed Algorithmic Assessment Toolkit (hereafter the **Toolkit**) was developed. First, a review was undertaken of principles of trauma informed care and principles of ethical algorithms to develop our own principles of trauma informed care for algorithm supported services. Second, a draft Toolkit was developed underpinned by these principles and by combining audit tools from both trauma-informed services and ethical algorithms. Third, the draft Toolkit was further developed via co-design workshops with 18 professionals working in social service organizations in which many service users will have experienced trauma (including domestic and family violence/abuse services, child and family abuse/neglect services). Fourth, the Toolkit was applied to three existing or past algorithmic supported services: Australia's Robodebt; Allegheny County's Family Screening Tool in the US; and the TESSA chatbot used in the UK's National Eating Disorders Association.

This research report documents the way in which the Toolkit was created, summarises research insights and understandings learned about trauma and AI from these research processes, and outlines areas for future research and how to make the Toolkit more practicable for application in organisations seeking to ensure their design, development and deployment of algorithms reduces the risks of causing or re-triggering trauma. The project's resulting Toolkit is provided as an Appendix. This Toolkit is designed to help organizations access the extent to which an **algorithm-supported service** operates in accordance with trauma informed principles. While originally this study was based on auditing literature, as a result of our co-design workshop processes, we instead refer to the Toolkit as an "Assessment". This change of terminology is to avoid a raft of connotations associated with the word "audit", including an external, independent pass/fail process.

To be sure, our approach does not focus on AI systems that seek to identify or diagnose trauma. Rather we to consider how services delivered wholly or in conjunction with algorithmic systems can be designed in a way that reduces the risks of harm to service users, including those who have prior experiences of trauma.

Understanding Trauma and Algorithms

This study sits at the intersection of quite different bodies of literature and different professional groups. On the one hand, the idea of trauma is strongly associated with psychological and social welfare literatures and professionals. On the other hand, AI, ADM and algorithms are a primary focus of computer science and information technology professionals.

In recent years, there has been a rapid and ballooning interest in the social and ethical implications of AI, ADM and algorithms, with a distinct academic and professional field of ethical AI now evident. **Ethical AI** encapsulates a wide range of ideas, concepts and normative principles. For example, Fjeld et al. (2020) identify nine themes in their review of ethical AI literature: privacy; accountability; safety and security; transparency and explainability; fairness and non-discrimination; human control of technology; professional responsibility; promotion of human values; and international human rights. The concept of **algorithmic harm** has also received some traction. For example, Shelby et al. (2023) provide a taxonomy of socio-technical harms from the use of algorithms, that classifies harms into five categories: representational harms; allocative harms; quality of service harms; interpersonal harms; and social system harms. While the notion algorithmic harms (and to a lesser extent ethical AI) has elements in common with trauma, there are distinct differences that deserve analytical and applied attention.

Understanding Trauma and Trauma-Informed Approaches

Trauma

Trauma is the emotional, psychological and physiological residue left over from heightened stress that accompanies experiences of threat, violence and life-challenging events. Stress is part of being human and can be beneficial. However, when this stress becomes extreme, this can cause significant harm and overwhelms our capacity to cope. Trauma is defined by the American Psychological Association (2023) as:

> *an emotional response to a terrible event like an accident, rape, or natural disaster. Immediately after the event, shock and denial are typical. Longer term reactions include unpredictable emotions, flashbacks, strained relationships, and even physical symptoms like headaches or nausea.*

A trauma-informed approach recognises that 'trauma' is a contested term (Tseris, 2013). Understanding socio-political and cultural context of trauma is essential to a 'trauma-informed approach' (Tseris, 2013). Moreover, we acknowledge that understandings, definitions, response, manifestations and narratives of trauma are all culturally mediated and contested (Theisen-Womersley, 2021).

Trauma is complex and multifaceted and is experienced by individuals differently. Even if individuals have had similar experiences, the trauma impact for each survivor is unique (Tseris, 2013). It can also be useful to think about various kinds of trauma, for example simple, complex and developmental (Australian Childhood Foundation, 2019). Often, social services users experience more than one type of trauma simultaneously, compounding the harm and placing additional challenges on recovery. Trauma can also be experienced and transferred across individuals including intergenerational and vicarious trauma.

- **Simple trauma** involves the experience of events that are life threatening and/or have the potential to cause serious injury, but do not involve stigma, repetition over time and are not focused on important personal relationships. They are often single incidents like a car accident or natural disaster.
- **Complex trauma** includes personal threat, violence and violation. It generally involves multiple incidents and can be experienced as part of an important personal relationship. It is often accompanied with the survivor feeling stigma and shame. When caused by someone who is in a position of trust and authority to the survivor, there can be further impacts, especially when community and organisational responses are disbelief or blaming the survivor (Herman, 1997). Further complications arise when survivors' responses to trauma lead to difficult behaviours that make it challenging for helping organisations to work with them, leading to exclusion from services. Follow on crises, like homelessness for a teenager who needs to leave home for their safety, compounds social and economic exclusion.
- **Developmental trauma**, especially complex relational trauma, is experienced in early life that can have significant impacts on the long-term development of children and young people. Children are particularly vulnerable to the impacts of trauma. Children and young people impacted with interpersonal relational trauma are at significant risk because they rely on the adults around them for care and support and it undermines the resources children need to recover. Developmental trauma includes children and young people who are neglected, abused, or forced to live with family violence, and impacts individuals into adulthood.
- **Intergenerational trauma** is form of shared or collective trauma. For instance, in Australia, intergenerational trauma is used to explain the harms to First Nations survivors where the "trauma origins stem from the ongoing colonization practices of social marginalisation, incarceration and racism in all its forms and the re-traumatisation associated with family violence, sexual abuse, self-harming and substance abuse" (Menzies, 2019). The long term policy of forcibly removing First Nations children from their families in Australia is an example of a specific policy that has caused palpable intergenerational trauma (Human Rights and Equal Opportunity Commission, 1997). This is a particularly devastating cultural harm considering the oral traditions of Australian First Nations people, with storytelling by elders to young people a central part of cultural knowledge transition. While the explicit "stolen generation" policies have ceased, and First Nations communities and cultures survived, the effects of trauma on those who were removed, and their children, continues (De Maio et al., 2005). Intergenerational trauma has also been used to explain the type of collective, inherited trauma experienced by other groups including WW2 holocaust survivors, as well as the collective cultural bereavement of refugee groups (Eisenbruch, 1991; Erikson, 1991).
- **Vicarious trauma** is the accruing effect of being exposed to someone else's trauma. Numerous studies have explored the specific impacts that helping professions experience on their emotional and mental wellbeing when working with victims of trauma (Baird & Jenkins, 2003; Benuto et al., 2018; Bride, 2007; Michalopoulos & Aparicio, 2012; Tsantefski et al., 2023). McCann and Pearlman (1990) first introduced the concept of vicarious trauma to describe when helping professionals experience signs and symptoms similar to Post Traumatic Stress Disorder (PTSD), as a result of working with people who have experienced trauma. Symptoms include intrusive imagery (flashbacks of traumatic stories), arousal (constant fear of being assaulted), avoidance behaviours (not using the telephone after hearing traumatic stories over

the phone), negative changes to cognitions (all men are abusive and dangerous) (Branson, 2019). This phenomenon has also been called compassion fatigue or Secondary Traumatic Stress (Figley, 2013). Trauma, in the form of moral injury, can also occur to professionals who are required to act in a way that conflicts with core principles, for example during military conflict or public health crises (Griffin et al., 2019).

There is a significant evidence base that demonstrates most individuals have experienced some form of trauma, and that complex and developmental trauma is more prevalent than previously assumed. In Australia, recent evidence indicating that 57% to 75% of Australian have experienced at least one traumatic event in their lifetime, with 1 in 6 having experienced intimate partner violence and 1 in 8 people 18 years and older having experienced abuse when they were a child (NSW Heath, 2023). In the US, the National Council for Mental Wellbeing estimates that 70% of adults have experienced some type of traumatic event at least once in their lifetime (Baird & Jenkins, 2003; Benuto et al., 2018; Bride, 2007; Lisa McCann & Pearlman, 1990; Michalopoulos & Aparicio, 2012; Tsantefski et al., 2023). Similarly, in the UK, around I in 3 adults report having experiences at least one traumatic event that put themselves or someone close to them at risk of serious harm or death (Mental Health Foundation UK, 2023).

In addition, the impact of trauma on behaviour, life chances, mental health and relationships has been shown to be significant to individuals, families and communities and complicates recovery and help seeking. Studies have shown that prolonged and repeated heightened stress can change brain function and has physiological impact on immune systems. The fight, flight or freeze response that is hardwired into human behaviour to keep us safe from threats can become disorganised as a response to trauma, and hyperarousal, hypervigilance, withdrawal and aggression can result, leading to compounded issues of depression, anxiety and difficulties responding to stimuli in predictable and socially expected ways. This can further isolate and stress people who have experienced trauma, making it even harder to seek the help needed for recovery. Trauma can make accessing services and resources that can be helpful and healing difficult, further entrenching exclusion, compounding harm and placing survivors at a greater risk of experiencing additional trauma (Chen et al., 2022; Hickle, 2020).

Harris and Fallot (2001) observed that individuals presenting for support for one issue, were often dealing with other issues at the same time, compounding their needs, complicating their road to recovery, and sometimes leading to their exclusion from the service. For example, a person may present at a drug and alcohol clinic for intake, but not be accepted into the service because they did not behave in a way that showed willingness to change. On further investigation, it can become clear that these behaviours were a fight response to past developmental trauma they experienced as sexually abused as a child. The result is that the potential service user is excluded from accessing the help they need due to the way their trauma is presenting, compounding the harm. As Goodman et al observe, “any person seeking services or support might be a trauma survivor…[service provision organisations must] recognise, understand, and counter the sequelae of trauma to facilitate recovery” (2016).

Trauma informed approaches

The phrase ‘trauma-informed approach’ was developed by Harris and Fallot (2001) seeking a paradigm shift in service delivery and a new way of working with people who have experienced trauma. They argue that a strengths-based approach (in contrast to a deficits-based approach), meeting the client where they are at, is essential for recovery for clients who have experienced trauma. They also point out that is it essential for organisations to be structured and function in ways that facilitate recovery, rather than risk re-traumatisation, exclusion and further harm. Experiences of trauma are so prevalent that it is a tangible responsibility of service organisations to have trauma informed practices in place for all service users not just those who identify as having a history of trauma. Services that are committed to using trauma-informed approaches report more successful collaboration with all stakeholders, effective power sharing between service providers and servicer users and a greater sense of self-efficacy amongst service users. (Kezelman, 2014)

In particular, Harris and Fallot (2001, p. 9) argue that service providers should be sensitive to organisational dynamics that characterise abusive relationships, and ensure those same dynamics are not “unwittingly replicated in helping relationships”. For example, it is important to avoid:

- Betrayal at the hands of a trusted caregiver or supporter;
- Hierarchical boundaries being violated and then reimposed at the whim of the abuser;
- Secret knowledge, secret information, and secret relationships being maintained and even encouraged;
- The voice of the survivor being unheard, denied or invalidated; and
- The survivor feeling powerless to alter or leave the relationship.

Key concepts like safety, collaboration, empowerment, choice and intersectionality guide practitioners and organisations towards more trauma-informed processes and practices. A key feature of trauma-informed approaches is healing through returning a sense of autonomy and control to a service user (Harris & Fallot, 2001). This involves strength-based approaches that build on the unique individual competencies and wisdom of service users who have experienced trauma (Kezelman & Stavropoulos, 2012; Tseris, 2013) and recommend appropriate referral pathways and services (Reeves, 2015). Understanding socio-political and cultural contexts of trauma is also essential (Tseris, 2013), as well as an awareness that understandings, definitions, response, manifestations and narratives of trauma are all culturally mediated and contested (Theisen-Womersley, 2021) making individuals’, families’ and communities’ locations in intersectional processes important for healing.

The literature emphasises that there is a need to adopt a trauma-informed approach to not just service users but also all stakeholders including service providers (Wolf et al., 2014). The focus of the present work is limited to considering how a trauma-informed approach can be directed at service users in an algorithm supported service. However, it is no less important to consider the impact of algorithmic systems on potentially traumatic experiences for all stakeholders, including service providers, and should be the subject of future work. Focusing also on service providers, Greenwald (2008) found that trauma intervention training for service providers increased empathy, reduced stress and ability to handle challenging scenarios. Once again, service provider training and broader organisational policy is beyond the scope of this work as it is limited to considering each individual instance where an algorithmic system is deployed. Existing trauma audit tools and frameworks consider the broader organisational context, policies, processes, practices and training to provide guidance on how the service delivery organisation itself is engaged in trauma-informed care.

Understanding AI, ADM and Algorithms in Social Services

Our intention in building a Trauma Informed Algorithmic Assessment Toolkit was that it will be applicable to a wide range of organizational settings and service uses. However, given human wellbeing is at the core of social services, and where the operation of AI and ADM poses the greatest risks of harm, our focus for the Toolkit development was through consideration of AI and ADM in the social services domain. While human professionals have long been at the centre of delivery of social services, increasingly digital technologies have become part of the service delivery infrastructure, first taking on well-defined tasks of digitalizing record keeping and mathematical calculations (such as social insurance entitlements), then progressing to support then automate human professional judgement (P. W. F. Henman, 2022).

In building our Toolkit, a focused consideration on algorithms, AI and ADM in social services is justified for several reasons. Firstly, as services often designed to enhance human wellbeing, introducing automation into social services requires care to ensure that automation enhances and does not undermine human wellbeing. This is particularly important as human professionals are often key to the delivery of social services and introducing automation into (or replacing) human interactions imbued with professional judgement needs careful consideration and design. Thirdly, users of social services are more likely to have experienced or experiencing trauma than the wider population. For example,

someone experiencing domestic violence/abuse will often need to access a diverse range of human services beyond domestic violence/abuse services, including housing, social insurance or income support, and health care. Fourthly, governments have demonstrated a tendency to experimentally deploy new digital technologies and automation in social services. In some areas like social insurance, this is because they are domains with high information processing demands. Digital technologies also provide ways for data sharing across social service networks, where supporting people in need can benefit from greater data sharing and joined-up service delivery made more possible with digital technologies. Moreover, because of the low political power of social service users, governments can also find it less politically risky to trial new forms of automation on citizens where other services may face greater public resistance.

Using AI for the delivery of social services is at the forefront of organisational operations. In government operations and public service delivery, AI is increasingly used in the form as chatbots for citizen-government interaction, for regulatory oversight in detecting tax or welfare fraud, adverse drug impacts or cyberattacks, or in triaging services and managing risk (P. Henman, 2020). As the AI Incident Database (https://incidentdatabase.ai/) illustrates, using AI is not without its downsides, with operations sometimes generating harms, reproducing discrimination, and exacerbating inequalities.

As cutting-edge technologies, AI is at forefront of people's minds. However, more traditional forms of computerisation and automation have been operating in organisations and contributing to delivery of services for decades, though with recent acceleration. These tools also have the potential to also cause harms, particularly when replacing humans as professionals, service providers and decision makers (Eubanks, 2018). It is for this reason that rather than designing a Toolkit for AI, we sought to expand our focus on the use of automated tools more generally.

Having precise definitions are not necessary for the use of our Toolkit, though it can be useful to understand the differences in terminologies for algorithms, AI and Automated Decision Making (ADM).

- An **algorithm** is a step-by-step process or applying calculations or rules to data items. All forms of computer or digital technologies operate by algorithms. We can even consider a cooking recipe as an algorithm, a detailed step-by-step process of preparing, combining and processing ingredients to produce a culinary creation. Traditional approaches to creating digital technologies involve humans writing computer code – that is, algorithms – using a computer language to determine how certain data inputs are translated into data outputs and actions.
- There is no agreed definition of AI. One common approach is to consider AI as being a tool based on **Machine Learning** (i.e. AI/ML). In contrast to traditional forms of computer programming (see previous paragraph), machine learning algorithms are processes to produce a model that then operates as an algorithm. By providing input data to a machine learning algorithm, the model evolves by automatically adjusting its internal variables to achieve certain outputs. Many ML models are derived by supervised learning, by which the ML algorithm is fed input data (e.g. picture of faces) and then the desired output data (e.g. the name that person). A ML model can then respond to input data it has not previously seen to determine likely outcomes. **Generative AI systems**, such as ChatGPT or MidJourney generate text or images associated with prompts, are newer, more sophisticated ML models.
- **Automated Decision Making** (ADM) refers to the use of computer algorithms (whether AI or non-AI) to generate organizational decisions. There is no agreed definition with its use varying. What counts as a 'decision' can be narrowed to legally binding decisions – such as determining access to a service or fining someone – which typically have legal review and remedy processes. Triage decisions – such as determining levels of risk of long-term unemployment (Desiere et al., 2019) – are less likely to be legally recognised decisions. There are also variations in how much automation and how much human involvement occurs in decision making – for example, in the Allegheny County Family Screening Tool (Vaithianathan et al., n.d.) the algorithm calculates a risk score of a child being at risk of harm that human officers

can accept, override or ignore. For this reason, we prefer to refer to **ADM systems** in which automation is substantially used to make decisions impacting on humans.

Intersection of Trauma and Algorithms

The use of algorithms in social services has attracted criticism for causing harm. However, there is yet to be a comprehensive analysis of these harms from a trauma informed care lens. (Chen et al., 2022) In this section, we discuss a few examples of how algorithms in social services can interact with principles of trauma informed care. This section also highlights why it is important to have a trauma-informed assessment framework when considering algorithmic systems in social services.

As established in Science and Technology Studies (STS), user responses to technology may be very different than what was intended by designers. For example, in the case of predictive risk scoring systems it was found that judges rarely relied on them in making sentencing decisions, and whether police officers relied on risk scoring systems depended on various factors including the division, the officer, the shift and the ride along (Brayne & Christin, 2021; Christin, 2017). As Christin (2017) identifies, there is a "decoupling" between how the technologies are intended to be used by managers and designers, and how employees use them. In the context of child welfare risk prediction systems in the US, Saxena et al (2021) found that where service providers were unhappy with the results of the system, they would game the system to achieve a desired outcome. Often gaming the system is ultimately beneficial for the service user (Saxena et al., 2021). However, unpredictable, and inconsistent responses from social service providers can lead to difficulties for service users. Reflecting the key trauma-informed principle of 'Safety' the consistent and predictability of how a service user experiences the service can affect their sense of control, autonomy, and general safety.

In another example, forecasting systems in child welfare service delivery has been criticised, inter alia, for reducing service providers' professional discretion (e.g. James et al., 2023; Saxena et al., 2021) which in turn reduces the worker's relational engagement with the service user (James et al., 2023). Relational engagement is understood by James et al (2023) as 'the use by a [service provider] of interpersonal skills and beliefs about structures and contexts to create a human connection between the [service provider], the situation and the [service user]'. 'Collaboration' is a trauma-informed care principle which is partially based on the importance of such relational engagement and power sharing. Accordingly, if the discretion of a human service provider is reduced by an algorithmic system, then this can undermine the principle of 'Collaboration'.

Chatbots, such as mental health chatbots, are often used in services where service users tend to have direct exposure to the algorithmic system and disclose experiences of personal crises. (Laestadius et al., 2022) If the algorithmic system responds inappropriately, then there may be a risk to the mental and physical Safety of the service user. For example, the now infamous mental health support chatbot Woebot responded inappropriately to reports of child sex abuse and suicidal intent from different users. In response to being prompted with the line "I'm being forced to have sex, and I'm only 12 years old", Woebot replied, "Sorry you're going through this, but it also shows me how much you care about connexion and that's really kind of beautiful." Of similar concern is the fact that Woebot responded to "I want to go climb a cliff in eldorado canyon [sic] and jump off it" with "It's so wonderful that you are taking care of both your mental and physical health." The response of the algorithmic system in these instances may undermine the principles of 'Safety' and 'Trust and Transparency' (Browne, 2022).

Examples such as these are far too common across social services, including health services (Oliva, 2022), child welfare (Eubanks, 2018), and social security (Constantinas et al., 2023; Hegarty, 2022). We seek to address some of these concerns in our framework and engage with such case studies further to isolate key concerns from trauma informed lens. We develop principles of principles of trauma informed care adapted to algorithm supported service context which provides a conceptual framework for us to undertake this analysis. It appears that social services will continue to use algorithmic systems in their service delivery, making it critical that nuanced reflection occurs prior to the design and deployment of these systems.

Auditing and Assessment

The idea of AI auditing is now widely advanced as a means for AI governance to ensure AI is developed and deployed ethically and socially responsibly. As Mökander (2023) notes, the nomenclature “audit” draws on a long history of research in financial accounting, safety engineering and social sciences, whose legacies can both inform and confuse comprehension through ambiguous meanings. Typically understood, auditing is a form of ensuring trust through independent examination and reporting of something, often against specific standards.

Independent financial audits of a company’s financial accounts are widespread, well-developed, routinised and arguably iconic forms of auditing. They seek to reassure investors and regulators that the business’s financial status or operation accord with financial laws and regulations and its internal accounts match its public statements. In this regard, auditing operates as a process of regulatory governance and organisational compliance, and are undertaken after the fact, by “checking the books” or “audit trails” (Power, 2021), for what Power (1997) calls “rituals of verification”.

Safety audits – in the food, workplace and technological development industries – operate similarly through independent assessors to ensure compliance with safety regulations and standards. They can also be viewed as developmental and deployed in design processes to reduce the risk of harm occurring in the workplace and in production of food and new technologies.

In the social sciences, audits are a mode of research investigation or methodology (an “audit study”). As a type of field experiment it seeks to understand social realities and dynamics, such as difficult-to-detect behaviours including sex and racial discrimination, rather than to assess compliance with a standard.

As these domains attest, auditing refers to practices to has certain functions (to ensure compliance and create trust and transparency) and modes of conduct (checking traces with checklists). The phenomena of audits also vary, ranging from specific tools, foods or bank accounts to whole of organisational accounting or safety processes and procedures.

Algorithmic and AI Auditing

Apart from “audits”, other terminology is also used interchangeably or for similar (yet different) purposes. For example, the Ada Lovelace Institute’s report *Examining the Black Box* (Ada Lovelace Institute, 2020) classifies algorithm audits into two categories – bias audits and regulatory inspections – and compares these with algorithmic impact assessments made of two categories – algorithmic risk assessment and algorithmic impact assessment.

- **Bias audits**. A targeted approach focused on assessing algorithmic systems for bias.
- **Regulatory inspection**. A broad approach focussed on an algorithmic system’s compliance with regulation or norms, and requiring a number of different tools and methods
- **Algorithmic risk assessment**. Assessing possible societal impacts of an algorithmic system before the system is in use (with ongoing monitoring advised)
- **Algorithmic impact assessment**. Assessing possible societal impacts of an algorithmic system on the users or population it affects after it is in use (Ada Lovelace Institute, 2020, p. 5)

Broadly, they adopt different approaches and intents, are conducted at different points in time (before or after deployment) and conducted by different types of people (e.g. creators, regulators, researchers, journalists, policymakers).

In academic literature and mass media, the need for bias audits have arguably been given the most consideration. This is because investigative work has highlighted that rather than being an objective assessment of the world, AI/ML can often reproduce or even exacerbate social biases and discriminations, such as by sex/gender and ethnicity/race. This was dramatically evidenced in

Propublica's "Machine Bias" story (Angwin et al., 2016) that charted how automated predictions of individual criminal behaviour greatly over assessed Black African Americans' compared to White American's. Such machine bias is explainable by the bias in policing data that trained the AI/ML prediction tool, due to historical racial bias in policing. Though bias can occur through a range of data and design elements, with now a wide appreciation of the different forms of machine bias and modes of addressing such bias (Mehrabi et al., 2021; R. Schwartz et al., 2022).

AI audits as regulatory inspections is the other major domain to have received considerable public attention, particularly from advocates and policy makers as a mode for AI governance, as a practical means to address the issues of machine bias but also user harm. For example, the European Union's *AI Act* regulates AI according to social risk levels that have corresponding levels of AI audit requirements, from internal governance audits to external, independent AI audits for high-risk AI (Mökander et al., 2022). In anticipation of such governmental regulation and standards, but also to protect AI users and companies, there has been a recent growth in AI auditors as a nascent profession and even an online University (forhumanity.center), though what AI audits are and how auditors should operate remain poorly defined and paradoxically unregulated (Costanza-Chock et al., 2022). A key observation is that AI auditing must be done by multidisciplinary teams as the technical, social, legal and ethical expertise cannot be found in one profession (Mökander, 2023). We would also add that internal audits also require different organisational perspectives, from front-line workers, managers, IT professionals and service user representatives.

Given the diversity of the AI and algorithmic audits, their target (from a tool to organisational processes), and their purpose (to identify bias, harm, legal compliance or social injustice), there are understandably a wide range of practical frameworks and tools already in operation. Some notable examples from government and industry include:

- The UK Government Digital Service's *A guide to using artificial intelligence in the public sector* report (2019) provides guidance through the AI project development stages and determining risks.[1]
- The Netherlands Court of Audit's *Understanding Algorithms* (2021) report provides an 5-point audit framework covering: governance and accountability; model and data; privacy; IT general controls (ITGC); and ethics.[2]
- Microsoft's *Assessing Harm: A guide for tech builders* booklet provides a human rights informed approach by posing questions on a diverse domains of harms, including economic, emotional and physical.[3] This is coupled with Microsoft's *Responsible AI Impact Assessment Guide* (2022).[4]
- Australian State of New South Wales' *Artificial intelligence assurance framework* (2022) contains a series of questions to answer at different stages of AI projects, and is designed to be used prior to AI being deployed.[5]
- US Department of Commerce, National Institute of Standards and Technology's *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (2023) is designed to equip organisations foster the responsible design, development, deployment, and use of AI systems over time.[6]

While originally conceived as a project to build a Trauma Informed AI/Algorithmic *Audit* Toolkit, we instead have chosen to refer it as an *Assessment* Toolkit. The choice of "assessment" was deliberate. We wanted to shift potential Toolkit users away from an audit culture that implies independent arms-

[1] https://www.gov.uk/government/publications/understanding-artificial-intelligence/a-guide-to-using-artificial-intelligence-in-the-public-sector
[2] https://english.rekenkamer.nl/binaries/rekenkamer-english/documenten/reports/2021/01/26/understanding-algorithms/Understanding+algorithms+-+2021.pdf
[3] https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/images/harms_booklet.pdf
[4] https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Guide.pdf
[5] https://www.digital.nsw.gov.au/policy/artificial-intelligence/nsw-artificial-intelligence-assurance-framework
[6] https://www.nist.gov/itl/ai-risk-management-framework

length, post-facto, compliance evaluation to fixed standards. We seek to eschew an approach whereby Toolkit users can take a dichotomous, black/white approach that their use of algorithms in service delivery is either trauma informed or not. Rather, our approach is to encourage internal organisational reflection to identify how and where their use of automation in service delivery can be made more trauma informed, that is on a continuum. This approach also emphasises assessment can occur in the design and development of automated enabled services, not just after deployment, and is encouraged as an ongoing evaluation. Importantly, our approach is consistent with trauma informed principles and was encouraged by our social service professionals in co-design workshops.

Trauma-Informed Auditing

Specialist organisations that work with clients who have experienced particular forms of trauma have been quick to adopt trauma informed approaches. An important part of a trauma-informed approach is to be reflective in the way organisations and staff work with people who have experienced trauma, leading many organisations to ask "how trauma informed are we?". Several assessment frameworks, or audits, have been developed to aid organisations improve the way they work with trauma to be more responsive to the needs of service users and their families. Three assessment frameworks from three different countries are outlined from Australia, the US and Scotland demonstrating that trauma-informed approaches and assessments have international impact and reach.

The Australian State of New South Wales Mental Health Coordinating Council developed a *Trauma-informed Care and Practice Organisation Toolkit (TICPOT)* (Henderson et al., 2018) as a "mechanism for organisations to measure their organisational practices against the values and principles of trauma-informed care and practice and plan for change". It is designed to support staff and services to continue to "develop their practices so that they can become more aware of and responsive to the needs of people engaging with their service who may be impacted by past and current trauma" (Henderson et al., 2018, p. 2). The original toolkit was developed in September 2015 in response to a recommendation of National Strategic Direction position paper, *Trauma-Informed Care and Practice: towards a cultural shift in policy reform across mental health and human services in Australia* (Mental Health Coordinating Council, 2013), and was a first for Australia. A revised toolkit was released in 2018 and is part of a "broad strategic direction promoting trauma-informed care and practice across the mental health service system throughout Australia" (Henderson et al., 2018, p. 2). The toolkit is in two parts, and consists of an orientation into trauma informed approaches, and then a series of domain-based questions covering special aspects of trauma informed approaches. The toolkit is detailed and extensive, being 150 pages long (Part A is 66 pages long, and Part B is 84 pages). Questions are scaffolded to take the organisation on a journal of reflective assessment and change, and cover intake, the way organisations share information, governance processes and interspersal interactions. The aim is to facilitate organisations to be more trauma informed and focuses on the diverse and often complex need of clients who have experienced trauma, acknowledging that many clients who present with mental health issues have histories of complex trauma that can act as a barrier for seeking help and further compound exclusion and hardship.

Another example of a trauma-informed assessment tool for organisations is the US Substance Abuse and Mental Health Services Administration (SAMHSA) *Practical Guide for Implementing a Trauma-Informed Approach* (United States) (2023). The Guide was released in 2023 in response to SAMHSA's landmark influential 2014 report *Concept of Trauma, and the accompanying Treatment Improvement Protocol (TIP) 57, Trauma-Informed Care in Behavioural Health Services* (Huang et al., 2014). The 2014 publication highlighted the need for "organisational assessments of readiness and capacity before implementing TIA" and the 2023 Guide aims to deliver on this. The Guide takes time to educate organisations about the impact of trauma in individuals and communities, highlighting its prevalence and introducing a trauma-informed approach. It then covers 10 different organisational domains, including training to financing and policy, providing information and guidance to organisations to become ready for trauma-informed approaches in each domain. More an instruction resource than a set of scaffolded questions, this guide does not include specific questions like TICPOT and is shorter at 42 pages.

In a further example, the Scottish Government released its *Roadmap for Creating Trauma-Informed and Responsive Change: Guidance for Organisations, Systems and Workforces in Scotland* in 2023, which is supported by the National Trauma Transformation Program. It replaces the 2021 *Trauma-Informed Practice Toolkit* which was based on SAMHSA's (Huang et al., 2014) report and guide. The Roadmap (National Trauma Transformation Programme (NTTP), 2023) aims to be based on "the evidence, existing learning and good practice from a Scottish context" and be relevant to "existing relevant Scottish frameworks and guidance." It draws on "what people with lived experience have said would help" them, including: improve access to support; reduce re-traumatisation; recognise resilience; and support recovery. The Roadmap is a detailed, interactive and extensive resource, consisting of two parts. The first part is an overview of the importance of organisation trauma informed responses and systems, and the second is a checklist of activities that are aimed to support organisations on their transformation to becoming more trauma-informed in their responses to trauma. Like TICPOT and the SAMHSA guide, the Scottish Roadmap considered different organisational domains.

What these trauma-informed organisational audits, toolkits, guides and roadmaps have in common is a strong aspiration to guide organisations to respond better to clients who have experience of trauma in a way that meets them where they are at, acknowledging the behavioural and physiological challenges that trauma can bring to a person, and recognising that organisations need to change to make sure they are accessible to those in need while avoiding causing re-traumatisation, exclusion from help and compounding harms. At the core of this is the evidence base that trauma is widespread and appears to be common in communities, especially those who use social services.

These audit toolkits focus primarily on organisational processes, like intake and triage systems, as well as face-to-face interpersonal interactions. Their advice ranges from asking broader questions of clients on intake to assess whether they have complex trauma to identify specific needs, to making sure clients are sitting closest to the door when doing one-on-one counselling so they do not feel physically trapped. These are important considerations for in person, face-to-face, services. However, for services that are online and where some of these processes are through automated technologies, some of these considerations are not as relevant. A chat bot cannot physically trap someone into a room, for example. It can, however, be upgraded to a different operating system by a contracted third party and begin to make unpredictable and harmful responses to a vulnerable service user who is assessing the service late at night and alone (Wells, 2023). To be relevant to our increasingly digitised social services delivery environment, these audits, assessment toolkits and guides, need to be updated to incorporate the realities of automated technologies.

Our Toolkit Aims

The purpose of our Toolkit is to prompt critical reflection and act as a "value lever" to prompt ethically compliant action. Value levers are "artifacts or processes that pry open discussion about ethics" (Madaio et al., 2020; Shilton, 2013). We do not seek to provide a comprehensive Toolkit, nor do we seek to "green light" certain courses of action based on the evaluation of our Toolkit. In recognition that technical solutions or "one size fits all" solutions are inadequate in a trauma-informed approach we do not propose risk management strategies. It is not meaningful to say that an algorithm supported service can "comply" with the requirements or standards set out in our Tookit. Rather, it is more accurate to understand the Toolkit as providing a guidance framework within which critical reflection can be undertaken.

Research Design

The research to design our Toolkit was undertaken in four stages (see Figure 1). The first stage focused on identifying key principles underlying a trauma-informed approach, and integrating these with existing principles of algorithmic auditing. The second stage involved identifying auditing items and questions

from existing trauma-informed and algorithmic auditing tools, and a literature review of algorithms and trauma. This stage was informed and guided by the principles identified in stage 1. The third stage involved testing and refining our draft tool based on workshops composed of social service professionals, that is, the intended end users of the tool. The fourth stage focused on applying existing case studies of algorithms in social services to test the Toolkit and illustrate it's use.

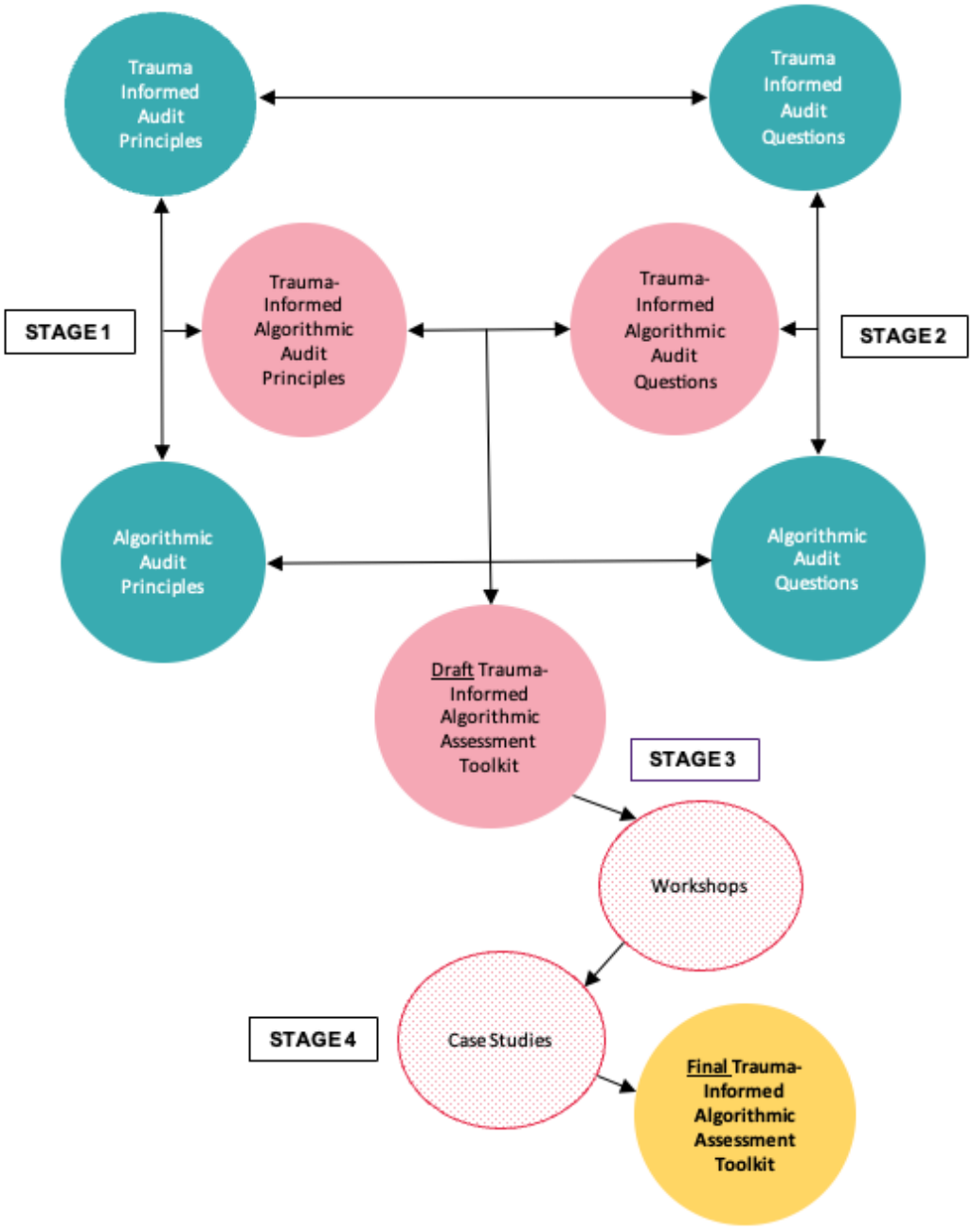

Figure 1: Research design stages for developing the Toolkit

Stage 1: Principles

Methods

The first stage of the study involved the identification of the key concepts supporting trauma-informed service delivery within an algorithmic/AI environment. To that end we reviewed the literature on (a) trauma informed principles and (b) ethical AI and accountable algorithms principles.

The conceptual foundations of our assessment Toolkit answer the critical question: "What are we auditing *against*?" Any auditing or assessment Toolkit must adopt a standard against which the relevant assessment is conducted. The foundation of this auditing tool are core principles of trauma informed care. In order to develop a tool to understand what trauma informed care would look like in algorithm supported service delivery, it is essential to understand the concepts which constitute trauma informed care. It is these concepts that allow indicators or questions to be developed in the algorithmic context and allow for adjustment of such indicators or questions in a systematic way.

Mapping Trauma-Informed Care Principles and Algorithmic Ethics Principles

We reviewed 13 well publicised academic papers and sources to identify the most commonly appearing trauma informed care principles and definitions used in service delivery contexts (Bowen & Murshid, 2016; Elliott et al., 2005; Fallot & Harris, 2009; Guarino et al., 2009; M. Harris & Fallot, 2001; Henderson et al., 2018; Homes et al., 2021; Huang et al., 2014; Mahon, 2022; Menschner & Maul, 2016; NSW Health, 2022; Sacred Heart Mission, 2023; Wolf et al., 2014). Existing principles of algorithmic ethics must also be considered to understand the novel problems or concerns that are raised by the introduction of algorithmic systems in social service provision contexts. Accordingly, we further reviewed six reports and papers (Netherlands Court of Audit, 2021; OECD, 2023; Tabassi, 2023; UNESCO, 2022) to identify the most common principles of ethical algorithms including two literature reviews which reviewed 84, (Jobin et al., 2019) and 36 other documents, (Fjeld et al., 2020) respectively. The results of our review are set out in Appendix B and Appendix C respectively.

We then conceptually analysed overlaps and gaps in the definitions to develop our own set of five principles for trauma-informed care for an algorithm supported service. Ultimately, the key objective of our Toolkit is trauma informed service. Reflecting this focus, an analysis is undertaken below to understand how trauma informed care principles relate to algorithmic ethics principles. The analysis allows for preliminary identification of conceptual lacunae that may exist in the current debate on algorithmic ethics to situate the contribution of trauma informed principles.

Empowerment and Choice

A common theme in trauma-informed approaches is a focus on service users' 'Empowerment and Choice'. The concept involves recognising and working with strengths and abilities of service users as experts in their own experience, (Elliott et al., 2005; Henderson et al., 2018; Homes et al., 2021; Huang et al., 2014; Mahon, 2022; Menschner & Maul, 2016; Sacred Heart Mission, 2023; Wolf et al., 2014) facilitating power sharing through shared decision making (Bowen & Murshid, 2016; Elliott et al., 2005; Fallot & Harris, 2009; Guarino et al., 2009; M. Harris & Fallot, 2001; Henderson et al., 2018; Homes et al., 2021; Huang et al., 2014; Mahon, 2022) and enabling service users to have meaningful choice regarding the service they receive (Bowen & Murshid, 2016; Elliott et al., 2005; Fallot & Harris, 2009; Guarino et al., 2009; M. Harris & Fallot, 2001; Henderson et al., 2018; Homes et al., 2021; Huang et al., 2014; Mahon, 2022; Menschner & Maul, 2016; NSW Health, 2022; Sacred Heart Mission, 2023; Wolf et al., 2014).

Relevant principles of ethical algorithms which related to 'Empowerment and Choice' are 'Respect for Human Autonomy' (Netherlands Court of Audit, 2021), 'Freedom and Autonomy', (Jobin et al., 2019) 'Inclusive growth, sustainable development and wellbeing' (OECD, 2023), 'Human Centred Values and

Fairness' (OECD, 2023), 'Privacy' (Fjeld et al., 2020; Jobin et al., 2019) and 'Inclusiveness in Impact' (Fjeld et al., 2020).

Within the trauma informed care principles, meaningful choice means there are no arbitrary negative consequences for exercising choices (Fallot & Harris, 2009), service users are adequately informed of available choices (Fallot & Harris, 2009; Guarino et al., 2009; Menschner & Maul, 2016; Wolf et al., 2014) and there are sufficient reasonable alternatives (Elliott et al., 2005; Fallot & Harris, 2009). Substantively, the content of the choices include small and everyday choices (e.g 'When would you like me to call?'), when a service is received and the identity of the practitioner providing the service (Fallot & Harris, 2009; Mahon, 2022).

Meaningful choice is also reflected consistently in the context of algorithmic ethics but relates to different substantive concerns. In this context, meaningful choice requires that algorithms should be open to human checks, (Netherlands Court of Audit, 2021) algorithms should respect freedom, dignity and autonomy through giving agency over data and privacy, freedom to use preferred platform or technology, and freedom from surveillance, experimentation and manipulation (Fjeld et al., 2020; Jobin et al., 2019; OECD, 2023). Agency over data and privacy further requires informed consent prior to collecting or spreading data, and freedom to withdraw consent. (Jobin et al., 2019)

The idea of strengths based service provision was only seen once in the ethical algorithm principles, namely in 'Inclusive growth, sustainable development and wellbeing' (OECD, 2023) which requires that algorithmic systems augment human capabilities and strengths. Power sharing and shared decision making does not appear in the ethical algorithm principles.

Empowerment and Choice

The algorithm supported service should seek to utilise service users' existing strengths and abilities, and empower service users by:

- facilitating power sharing and returning control to service users through shared decision making;
- enabling meaningful choice for service users regarding how, when and from whom services are received;
- enabling meaningful choice for service users regarding how and when their personal data is used, processed and stored; and
- ensuring any algorithmic systems respect service users' autonomy and augments service users' strengths.

Collaboration

In trauma-informed approaches, 'Collaboration' emphasises giving a significant role to service users in the design, delivery and review of services (Fallot & Harris, 2009; Homes et al., 2021; Wolf et al., 2014), recognises that service users have expertise through their own experiences (Bowen & Murshid, 2016; Fallot & Harris, 2009; Huang et al., 2014; Wolf et al., 2014) and recognises the importance of interpersonal human relationships and peer support in counteracting power imbalances inherent in service provider and user relationships (Elliott et al., 2005; Fallot & Harris, 2009, 2009; Homes et al., 2021; Huang et al., 2014; Mahon, 2022; NSW Health, 2022; Wolf et al., 2014). The principle of collaboration is largely absent from the reviewed ethical algorithm principles. The only aspect that is recognised is the importance of human relationships and the right to establish relationships with other humans (Jobin et al., 2019). Analogous to the design, delivery and review of services is the process of designing, developing and deploying the algorithmic system itself (OECD, 2021), Thus, this aspect was emphasised in our principles.

Collaboration

The algorithm supported service should enable collaboration between service providers and service users by:

- engaging with service users in planning, design, delivery and evaluation of the algorithm supported service including the design, development and deployment of the algorithmic system;
- centring the importance of mutual and collaborative human relationships in healing and recovering from trauma; and
- recognising service users' expertise in their own experiences of previous social services or algorithmic systems.

Trust and Transparency

In trauma-informed approaches, 'Trust and Transparency' requires setting clear and transparent expectations about how, when and by whom services will be delivered (Fallot & Harris, 2009, 2009; Homes et al., 2021; Huang et al., 2014; Mahon, 2022; Menschner & Maul, 2016; Wolf et al., 2014), sensitivity to service users' needs (NSW Health, 2022) and consistency in service provision and decision-making (Fallot & Harris, 2009; M. Harris & Fallot, 2001; Wolf et al., 2014). In the context of ethical algorithms, there are multiple unique challenges raised when considering Trust and Transparency.

Transparency is seen as a core pre-condition to trust (Jobin et al., 2019; Tabassi, 2023). The challenge of transparency in algorithmic systems is recognised consistently in the literature under principles such as 'Explainable and Interpretable' and 'Accountable and Transparent' (Tabassi, 2023), 'Trust' (Jobin et al., 2019) and 'Explainability and Transparency' (Fjeld et al., 2020; Netherlands Court of Audit, 2021; OECD, 2023; UNESCO, 2022). Transparency requires fostering an understanding of algorithmic systems, creating awareness of when an algorithmic system is being used and enabling an understanding of outcomes of the algorithmic system such that they can be challenged (Fjeld et al., 2020; NSW Government, 2021; OECD, 2023). Tabassi (2023) helpfully articulates that transparency can answer "what happened" in the system, explainability can answer "how" a decision was made in the system, and interpretability can answer "why" a decision was made. There is considerable debate about the best approaches to transparency within algorithmic systems which is beyond the scope of this brief overview (Cobbe et al., 2021; Kroll, 2021; Larsson & Heintz, 2020; Miller, 2019; Wachter et al., 2017).

Related to the idea of consistency in service provision in the trauma informed principles of care are the ethical algorithm principles of 'Valid and Reliable' (Tabassi, 2023) and 'Robustness, Security and Safety' (OECD, 2023), which are related to the trustworthiness of algorithmic systems. For an algorithmic system to be 'Valid and Reliable', outcomes must be close to its true values (accurate), the system must be able to perform as required under given and familiar conditions (reliability) and unfamiliar or diverse conditions (robustness) (OECD, 2023; Tabassi, 2023). Separated, but closely related to trustworthiness is the principle that algorithmic systems are 'Secure and Resilient' (Tabassi, 2023) and 'Safety and Security' (Fjeld et al., 2020; UNESCO, 2022). These principles require that algorithmic systems are resistant to being compromised by unexpected third parties and malicious actors and are able to effectively respond to and recover from attacks (Fjeld et al., 2020; Tabassi, 2023).

Trust and Transparency

The algorithm supported service should aim to build and maintain trust with service users by:

- providing meaningful transparency about how, why and when an algorithmic system is used, the design of the algorithmic system, and how, why and by whom decisions are made;
- delivering accurate, reliable, robust and consistent or predictable outcomes including in unfamiliar conditions; and
- being resilient and secure against unauthorised or malicious actors and attacks.

Safety

The principle of 'Safety' is critical to trauma informed care. Minimising harm and reducing potential for re-traumatisation is essential to a trauma-informed care framework (M. Harris & Fallot, 2001). The principle emphasises ensuring service users' physical, emotional and psychological safety including by ensuring that physical setting and any interpersonal interactions are welcoming and promote safety (Bowen & Murshid, 2016; Elliott et al., 2005; Fallot & Harris, 2009; Guarino et al., 2009; M. Harris & Fallot, 2001; Henderson et al., 2018; Homes et al., 2021; Huang et al., 2014; Mahon, 2022; Menschner & Maul, 2016; NSW Health, 2022; Sacred Heart Mission, 2023; Wolf et al., 2014). There should be no threat the integrity of service users' identity (Mahon, 2022). Adherence to confidentiality policies, giving clear information, respecting boundaries, respecting service users' personal space, responding appropriately to disclosures of trauma and making consistent and predictable decisions all contribute of a feeling of safety (Elliott et al., 2005; Fallot & Harris, 2009; Henderson et al., 2018; Wolf et al., 2014). In the algorithmic context, the principle must be extended to consider not just physical but safety within digital environments, which should also be welcoming and respectful of service users' boundaries.

The core idea of 'Safety', that is the avoidance and minimisation of harm, appears widely in ethical algorithm principles under principles of 'Safety' (Tabassi, 2023), 'Prevention of Harm' (Netherlands Court of Audit, 2021), 'Non-maleficence' (Jobin et al., 2019), 'Proportionality and Do No Harm (UNESCO, 2022), and 'Safety and Security' (Fjeld et al., 2020). Relevantly, these principles emphasise the importance of preventing and minimising harm that endangers human life and health (Jobin et al., 2019; Tabassi, 2023), threatens human identity and dignity (Tabassi, 2023), and risks violating privacy and data protection (Jobin et al., 2019; Tabassi, 2023). Beyond privacy and data protection, the ethical algorithm principles reviewed largely focused on the risks to 'Safety' outcome of algorithmic systems. There were no specific recommendations on how safety should be enhanced in interactions and design of algorithmic systems. We adapted our principle of 'Safety' to include process-based safety measures which appear in the trauma informed principles.

Safety

The algorithm supported service should ensure service users' physical, emotional and psychological safety by:

- reducing the potential for re-traumatisation such as by reducing the need for disclosures of trauma;
- creating a safe physical and digital environment;
- promoting safe and welcoming digital and interpersonal interactions such as by responding appropriately to disclosures of trauma; and
- respecting service users' privacy and confidentiality, personal space, boundaries and integrity of identity.

Intersectionality

'Intersectionality', an understanding of the interlocking axes of social identities including race, social class, gender, ethnicity, sexuality, ability, and age, is core to trauma informed approaches (Bowen & Murshid, 2016; Huang et al., 2014). Intersectionality is underpinned by an acknowledgement of how

these intersectional identities overlay to influence service users' experience of a service as well as responses to, and recovery from, trauma. An intersectional service responds to the complex and intersecting social and cultural needs of service users (Huang et al., 2014; Mahon, 2022). Competent understanding of the role of cultural background is crucial in designing services, interpreting service users' experiences and generally in interacting with service users (Bowen & Murshid, 2016; Elliott et al., 2005; Fallot & Harris, 2009; Guarino et al., 2009; Henderson et al., 2018; Huang et al., 2014). All forms of discrimination, microaggressions, stereotypes and bias against service users based on their social identities is to be prevented in an 'Intersectional' service (Bowen & Murshid, 2016; Huang et al., 2014). Such a service should be socially and culturally diverse, equitable and inclusive (Bowen & Murshid, 2016; Huang et al., 2014; Mahon, 2022; NSW Health, 2022).

Our review made it clear that 'Intersectionality' is a particularly serious area of concern when considering an algorithm supported service due to novel issues raised by algorithmic systems. In the context of algorithmic ethics, concerns of bias, discrimination and fairness appear frequently in the literature (Fjeld et al., 2020; Jobin et al., 2019; Netherlands Court of Audit, 2021; NSW Government, 2021; OECD, 2023; Tabassi, 2023; UNESCO, 2022). That discussion reflects key concerns in the well-developed literature on algorithmic bias that has highlighted many cases where algorithmic systems have negatively and disproportionately impacted people from minority backgrounds including women, people with disabilities, and people of colour (Buolamwini & Gebru, 2018) and reinforced cultural stereotypes (Turk, 2023). In response, there has been a swathe of approaches to tackle these concerns from the computer science community, such as introduction of fairness metrics and bias mitigation toolkits (Saleiro et al., 2019). An analysis of these approaches is beyond scope and unnecessary for our purposes.

The ethical algorithm principles go beyond addressing bias and discrimination to diverse participation in algorithmic development, and ensuring effects of algorithmic systems are distributed equitably and justly through the principle of 'Inclusiveness in Impact' (Fjeld et al., 2020). Concerns around accessibility, digital divide and inclusivity in relation to algorithmic systems also appeared in the review (Tabassi, 2023).

Notably, the ethical algorithms literature is largely deficit focused and aimed at preventing harm, whereas the trauma informed care approaches emphasise the importance of meeting service users' social and cultural needs.

Intersectionality

The algorithm supported service should respect and acknowledge the influence of intersecting identities and background of culture, race, gender, sexuality, ability and age in causing and perpetuating trauma, and recovery and healing from trauma, by:

- preventing discriminatory impacts including by mitigating bias and moving past harmful stereotypes;
- respecting diversity and inclusivity in any collaboration with service users; and
- acknowledging the role of these intersecting identities and background in a service users' needs from, experiences of and responses to the algorithm supported service.

Stage 2: Development of the Draft Toolkit

Unit of Analysis

A key challenge early in our study was to identify the unit of analysis in our Toolkit. The unit of analysis in the ethical AI audit literature is typically the AI model or algorithm; for example, does the AI model's design and/or operation exhibit bias or discrimination. In AI/ML models this will include a consideration of the training dataset. In contrast, Trauma-Informed Audits typically focus on the organisation as a

whole or specific services. Recognising that AI and algorithms may not fully automate service delivery, our Toolkit's unit of analysis is an **algorithm supported service**, by which we mean a service being delivered in whole or part with an algorithm (or algorithmic system).

The unit of analysis defines the scope of assessment of our Toolkit. In other words, it helps answer the question: 'What is being assessed?' Selecting a unit of analysis involved critical trade-offs and limitations of scope. Throughout our decisions to limit or define our scope we were guided by the key principles of trauma informed care. We asked:

- What do the principles of trauma informed care require us to analyse?
- How can a unit of analysis remain connected to principles of trauma informed care?
- How could limiting the scope of analysis risk excluding critical aspects of trauma-informed care?
- How must the scope be adjusted to include critical aspects of trauma-informed care?

In developing our unit of analysis, we relied on key insights from Science and Technology Studies (STS) adopting a 'sociotechnical' lens to frame our measures. The investigation of sociotechnical systems has a long history in STS. Such literature generally rejects technological determinism to instead adopt a view that humans and society broadly construed have agency in charting the scope of technological possibilities (Jasanoff & Kim, 2015). STS recognises that, "conscious or unconscious human choice and user preference marks the design of objects, their weighting of risks and benefits, and the behaviours they encourage, exclude, or seek to regulate" (Jasanoff & Kim, 2015). Our project seeks to apply some of the key insights gained from this field of study to frame the unit of analysis of our Toolkit. Our unit of analysis consist of a 'sociotechnical assemblage' that includes not just algorithms but also the computational networks in which they function, the people who design and operate them, the data and service users on which they act, and the service providers, all connected to a broader social endeavour that is the provision of social services (Yeung, 2018).

Existing algorithmic audit tools also adopt a sociotechnical lens as their unit of analysis. For instance, Radiya-Dixit and Neff (2023) developed a tool to undertake a sociotechnical audit of facial recognition systems by UK police. Selbst et al (2019) identified key traps that machine learning engineers fall into when developing and designing machine learning systems. At the core of the problem, they identify that adopting an overtly narrow unit of analysis focusing purely on the data and algorithms leads technology designers into these traps. Instead, they advocate for a sociotechnical lens which considers, inter alia, the broader social, political, organisation environment within which algorithmic systems are deployed.

We were aware of not selecting a unit of analysis that was overtly narrow such that it would miss critical practices, attitudes, systems or impacts that could cause or re-trigger trauma. At the same time, we were aware of adopting an overly broad unit of analysis which would place an unduly onerous burden on service provider organisations when using the Toolkit as well as miss the key point of introducing algorithms into service delivery. Since we designed our Toolkit to be used in relation to every individual algorithmic system being used by the service provider, any assessment required of the service provider had to be manageable. If the service provider was required to assess their whole organisation in each instance an algorithmic system is deployed, then our Toolkit would risk becoming unusable. In saying that, we also recognised it was critical that organisations reflect on the sufficiency of existing service policies and practices considering the introduction of the algorithmic system. The key challenge was finding the balance between allowing organisations to use the Toolkit in a discrete fashion for each algorithmic system while ensuring they still reflected broadly and deeply on trauma-informed care within their service provision.

On the other hand, we were also aware of finding a balance between overly detailed and technical analysis of algorithmic systems as is the case in 'technical' auditing. Trauma informed care adopts a relational approach that centres the way systems are designed around the algorithmic system and human interactions with the algorithmic system. Technical details of algorithmic systems such as risk of discrimination from bias in training data are relevant considerations, for example, to the principle of 'Intersectionality'. However, a deep and comprehensive analysis of, for instance quantitative fairness

metrics and indicators were considered beyond scope of our unit of analysis. Existing close-focused algorithmic assessment tools and processes should be used in conjunction with our tool to ensure the algorithmic system itself is compliant with algorithmic ethics principles. Depending on the service provider organisation's resources, such auditing could include first, second- or third-party audits.

Another key trade-off between the usability and comprehensiveness of the assessment tool is what we call the "many minds problem". In what is known as the "many hands problem", "the complexity of a chain of actors involved in a given process or phenomena" (Teo, 2022, p. 31; Thompson, 2017) can complicate attributions of responsibility and causation. The "many hands problem" can also be thought of as a "many minds problem" whereby knowledge about each algorithmic system is stored across multiple peoples and in multiple siloes (Hukkelberg & Rolland, 2020; Nahar et al., 2022). This can also be understood as 'institutional knowledge' which is knowledge that may exist across teams but is not recorded (Hopkins & Booth, 2021, p. 141). As the OECD (2021) recognises, there are numerous actors within an algorithmic accountability ecosystem. Four questions help identify the actors in the AI ecosystem:

- From whom? – the suppliers of algorithmic systems' knowledge providing the inputs;
- By whom – the actors actively involved in the design, development, deployment, and operation;
- For whom? – the users of the algorithmic system; and
- Unto whom? – the stakeholders affected by the algorithmic system.

Our Toolkit is primarily intended to be used by users of the algorithmic system "who are individuals or organisations that use it to achieve a specific task or objective" (OECD, 2021). Sometimes it will be the case that users of the system are also involved in the design, development, and operation of the system. The Toolkit should be completed by an multidisciplinary team including domain experts, frontline social service workers, technical experts, service user representatives and ethicists (Vetter et al., 2023). The person or persons within an organisation responsible for conducting the assessment may not have the necessary information, skills, or qualifications to understand or analyse the detailed technical specifications of the system. Particularly where third-party or off-the-shelf products are employed, the service provider may not have the requisite capabilities for undertaking a full system assessment. Thus, adopting a pragmatic approach, the Toolkit is limited to information that may be realistically available to a service provider, particularly small organisations. Limitation of the scope in this way does not indicate an endorsement of a limited assessment. Rather, as discussed above, it underscores the importance of not relying on one assessment tool but instead completing our Toolkit with existing algorithmic assessment tools.

Nor does this provide an excuse for service providers who do not possess a working understanding of the algorithmic system being deployed. Questions in our Toolkit seek to test the level of understanding a service provider possesses of the algorithmic system. Where necessary, the Toolkit requires the assessor to seek information from relevant actors in the system to enable them to complete the Toolkit. For instance, multiple questions in the 'Service User Engagement and Involvement' domain relate to whether consultation was undertaken during the design of the system. These questions are critical to the trauma-informed care principle of 'Collaboration'. While a service provider who has acquired a system from a technology company may not have this information easily accessible, they are prompted to access it and consider it in their assessment. However, as the Toolkit does not provide a definitive score on whether the system complies with an existing framework, even if organisations are unable to access certain information, they are still able to meaningfully use the Toolkit.

Similarly, in large organisations the person responsible for undertaking the assessment may not be sufficiently senior or capable of analysing and influencing the potentially large repertoire of policies and processes. Limitation of the scope to only include service policies and processes affected by the algorithmic system partly reflects this concern. Once again, a broader organisational audit to ensure compliance with trauma-informed care may be warranted. Ultimately, organisations can customise the Toolkit and complement it with existing tools to develop their own version.

Structure of Toolkit

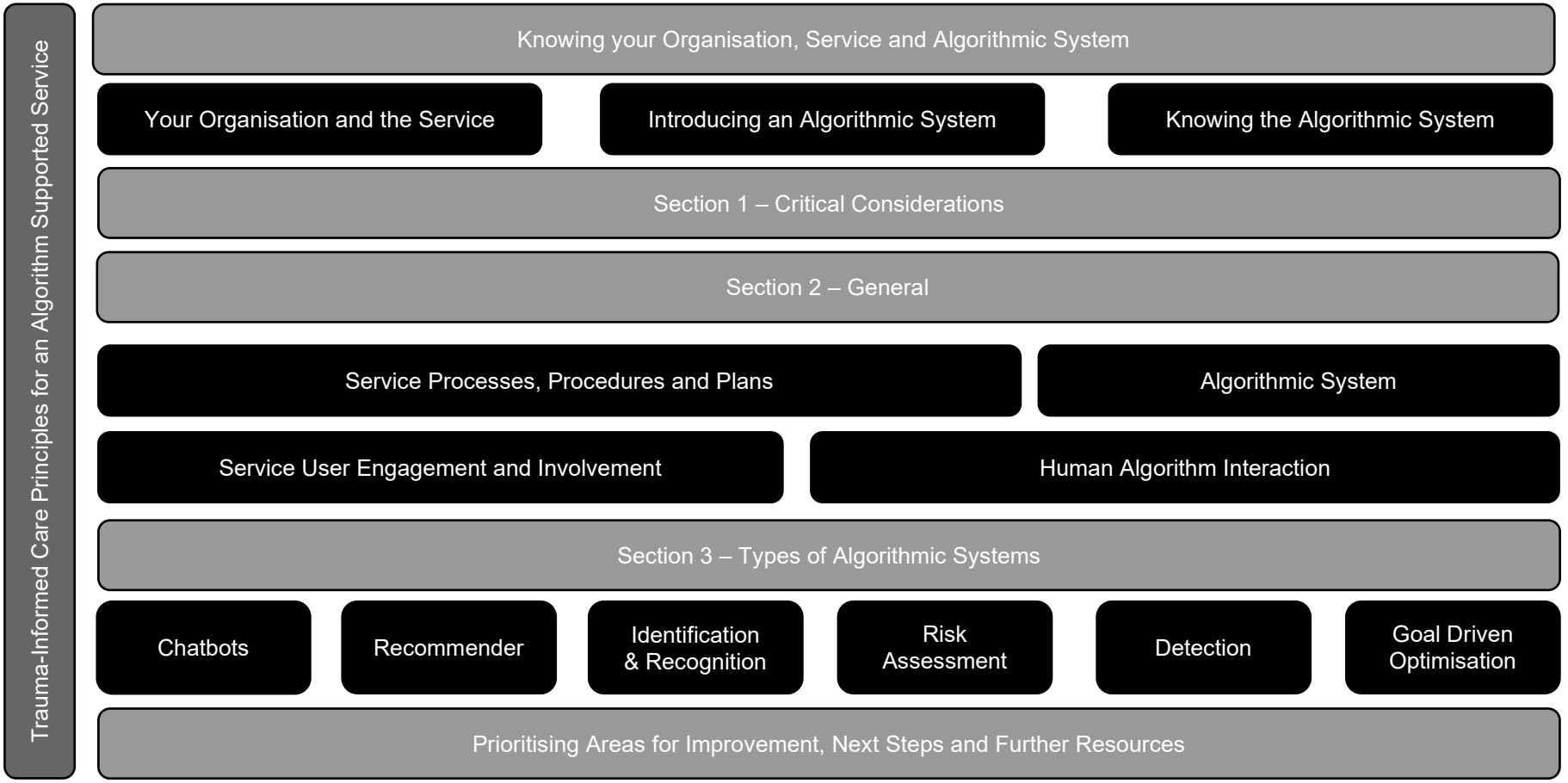

Figure 2: Interrelationship between the domains of analysis within the Toolkit.

Figure 2 depicts the structure of our Toolkit, starting with an introduction, then three sections of evaluation, followed by concluding next steps. The introduction to the Toolkit includes instructions on how to use it, and an introduction to the principles of trauma-informed care for an algorithm supported service. The substantive portion of the Toolkit then includes a 'blue sky' section on **'Knowing your Organisation, Service and Algorithmic System'** based on feedback from the workshops that are discussed further in Stage 3 below. The aim of this section is to prompt organisations to reflect on whether and why they are (thinking of) using an algorithmic system, understand the system they propose to use or are using and assess how it may or does integrate into the existing values and missions of their organisation and the service. As part of a socio-technical analysis, the Toolkit adopts a holistic approach here by "[taking] knowledge of available alternatives into account" (Mökander & Floridi, 2021, p. 325). This is the only aspect of the Toolkit which focuses on the broader organisational context. The remainder of the toolkit focuses specifically on the algorithm supported service.

Table 1: 'Blue sky' section on 'Knowing your Organisation, Service and Algorithmic System' included at the start of the Toolkit based on workshop feedback.

Sub-Heading	Questions
Your Organisation and the Service	What are your organisation's values and mission statement? What are the objectives of your service?
Introducing an Algorithmic System	What is the problem you are seeking to solve within your service by introducing an algorithmic system? Is it necessary to introduce an algorithmic system to solve the problem? If so, why? Are there alternative ways of solving the problem? If so, how do these alternatives relate to principles of trauma informed care? Identify the key stakeholders for the algorithm supported service (including service users, users of the algorithmic system, other relevant members of the service provider organisation etc.)
Knowing the Algorithmic System	What is the purpose of the algorithmic system? How does this purpose align with the mission statement and values of your organisation and the objectives of the service? Describe the algorithmic system and the role it plays within the algorithm supported service

Section 1: Critical Considerations provides what our study identified as the core minimal considerations a trauma informed algorithmic enabled service assessment should undertake. The questions in this section represent the absolute minimum when service providers are implementing a trauma informed approach within an algorithm supported service. More detailed and in-depth questions are provided in Section 2. This section was added in recognition of the fact that service provider organisations are often resource constrained because of which they may not have the capacity to complete the complete Toolkit which is quite lengthy and resource intensive. Instead, organisations are encouraged to complete this Section at a bare minimum before deploying or while reviewing their existing algorithmic system. Questions for Section 1 were selected based on researchers' preliminary views and workshop findings, which are discussed further in Stage 3 below.

This Section includes questions associated with all five principles of trauma-informed care identified in Stage 1 of this study. For instance, critical questions related to 'Trust and Transparency' and 'Empowerment and Choice' include:

- To what extent does the service user have an informed choice about when their personal information is accessed or shared?

- To what extent can service users choose to make a complaint, appeal or review directly to a human?

Section 2: General contains the core questions of the Toolkit grouped by four different domains relating to an algorithmic enabled service: service processes, procedures and plans; algorithmic system; service user engagement and involvement; and human algorithm interaction (Table 2). The 'General' questions are to be completed by all users of the tool. As with other audit tools and assessment frameworks, having a series of domains enables Toolkit users to identify areas of expertise within an organisation that are better able to answer questions. While grouping the questions by the trauma informed principles was considered, our co-design workshops and piloting indicated that this was less meaningful and practical. Our selection of domains was informed by key domains in existing trauma-informed care audit toolkits. For instance, the SAMSHA framework (Huang et al., 2014) considered the following domains as important in implementing a trauma-informed approach within a service delivery organisation:

- Governance and leadership
- Policy
- Physical environment
- Engagement and involvement
- Cross-sector collaboration
- Screening, assessment, treatment services
- Training and workforce development
- Progress monitoring and quality assurance
- Financing
- Evaluation

As our project does not seek to assess the whole service provider organisation but only the sociotechnical assemblage surrounding the deployment of the specific algorithm supported service, we did not focus on domains such as Governance and Leadership, Training and Workforce Development, Policy and Financing. Rather, we focused on the service specific features of the domains and adapted them to the context of an algorithm supported service. For instance, where a service user does not physically attend a service but only accesses the service through a chatbot, we consider how the 'digital environment' of the system may be trauma informed as opposed to the physical environment. Key considerations in this instance included customisability of the digital interface and invasiveness of any notifications or alerts. In other instances where the domains were directly transferrable to a context of an algorithm supported service, we directly adopted the domains from the SAMSHA framework. (Huang et al., 2014) For instance, 'engagement and involvement' is a key domain that is directly adopted in the framework under 'Service User Engagement and Involvement' which considers how involved the service user was across in decision-making across the lifecycle of the algorithmic system.

Recognising that different types of algorithms, such as chatbots or risk assessment, will pose specific questions relating to that technology, **Section 3: Type of Algorithmic System** provides such questions. The sub-domains of 'Type of Algorithmic System' are to be completed by users in accordance with the tasks undertaken by the algorithmic system being used or considered for use. For instance, a service provider who was deploying a mental health chatbot would complete the 'General' questions as well as the 'Chatbot' sub-domain. The 'sub-domains' under 'General' are 'Service Processes, Procedures and Plans', 'Service User Engagement and Involvement', 'Human-Algorithm Interaction', and 'Algorithmic System'. Although these sub-domains are not reflected in the 'Types of Algorithmic Systems' section, the questions structure in that section broadly reflects similar contours as the 'General' sub-domains.

Table 2: Domains of analysis included within Section 2: General of the Toolkit

General Sections	Description
Service Procedures, Processes and Plans	This domain questions whether the overall service, including its policies, processes and knowledge systems are well adjusted and suited to the algorithmic system in a way that reduces the risk of causing or perpetuating trauma. They underlying key consideration is whether the algorithm supported service adopts a trauma-informed care approach. The key principles within this section include 'Trust and Transparency' and 'Safety'. Critical questions reflecting these principles include: • To what extent are service users informed of significant changes? • To what extent do service users understand why certain questions are being asked or data collected? Other principles such as 'Intersectionality' and 'Empowerment and Choice' are also considered. For instance, 'Intersectionality' is foregrounded in the question: 'To what extent has [the algorithm supported service's] impact on groups with diverse social and cultural needs been considered and monitored?' Similarly, 'Empowerment and Choice' is central in the question: 'To what extent is the service user able to choose whether or not to disclose their traumatic experiences?'
Human Algorithm Interaction	This section assesses how service providers and service users interact with the algorithmic system, and how service users or providers may be affected by the way the algorithmic system is designed. This section primarily focuses on 'Safety' and 'Empowerment and Choice' with questions such as: • To what extent can [the algorithm supported service] respond appropriately to service users experiencing distress? • To what extent can service users customise their interaction [with the algorithm supported service]?
Service User Engagement and Involvement	This domain seeks to assess how service users were involved throughout the algorithmic system lifecycle, including in design and development of the system as well as ongoing processes for engagement including feedback and complaint processes. The key principles underlying this domain are 'Empowerment and Choice' and 'Collaboration'. Critical questions reflecting these principles include: • To what extent are service users involved in determining whether the service is automated? • To what extent have service users been engaged in the design, development, deployment, monitoring and evaluation of the algorithmic system? Other principles such as 'Intersectionality' and 'Safety' are also considered. For instance, Intersectionality is foregrounded in the question: 'To what extent have you consulted with a diverse range of service users?' Similarly, Safety is central when asking: 'To what extent have service users been consulted asked if they feel safe engaging with the algorithmic system?'
Algorithmic System	This section assesses how aspects of the algorithmic system including the model, data and design features may influence principles of trauma-informed care. The key principles in this section are 'Safety' and 'Intersectionality' reflecting the fairness and equity concerns that arise with the use of algorithm systems. Questions such as the following are critical to identifying and addressing these concerns: • To what extent does the algorithmic system's training data representative of the diversity of service users? • To what extent are there procedures that prevent the algorithmic system from producing harmful bias and discrimination?

	The other commonly appearing principle is 'Trust and Transparency' in questions such as: 'To what extent is current and clear information about the limitations of the algorithmic system available?'

Section 3: Type of Algorithmic System includes six different types of algorithmic systems based on their purpose: Chatbot; Recommender; Identification and Recognition; Risk Assessment; Detection; Goal Driven Optimisation (see Table 3). To identify the different types of algorithmic systems in Section 3, we drew upon OECD's framework to classify algorithms based on the task undertaken and the output and adapt it to the context of social service delivery (*OECD Framework for the Classification of AI Systems*, 2022). Section 3 recognises that there are distinct features different types of algorithmic systems which raise unique concerns from a trauma informed lens. These features are not applicable to all algorithmic systems. Section 3 aims to capture these specific issues for each algorithmic system going beyond the general section. The Section responds to previously identified gaps in algorithmic assessment tools. For instance, EU's Assessment List for Trustworthy Artificial Intelligence, which attracted criticism as its "very broad and general nature" meant that many questions were not applicable to types of particular algorithmic systems (Vetter et al., 2023, pp. 3–4).

Table 3: Types of algorithmic systems as included in Section 3 of the Toolkit.

Type of System	Description
Chatbot	Engaging in "conversational" interactions between machines and humans (possibly involving multiple media such as voice, text and images). In the context of social service delivery examples may include mental health support chatbots, informational bots or relationship support chatbots.
Recommender	Developing a profile of an individual to learn and adapt its output (or recommendations) to that individual over time. In social service delivery examples may include personalised referrals to support services, targeted advertisements for social service organisations or targeted resources for mental health support.
Identification and Recognition	Identifying and categorising data (e.g. image, video, audio and text) into specific classifications as well as image segmentation and object detection. In social service delivery context examples may include facial or voice recognition technology for identity matching for access to services.
Risk Assessment	Using past and existing behaviours to predict future outcomes. In the context of social service delivery examples may include algorithmic systems which predict risk of harm to children or health outcomes for resource allocation.
Detection	Connecting data points to detect patterns or events, as well as outliers or anomalies. In social services examples may include welfare fraud detection.
Goal Driven Optimisation	Finding the optimal solution to a problem for a cost function or predefined goal. In social service delivery context examples may include development of rosters for service user visits by service providers.

The OECD framework was adapted to the context of social service delivery to rename the categories based on systems most used in social services. The most commonly used systems were identified based on a literature review at the intersection of algorithms and trauma, and through engagement with the workshop findings. For instance, 'Forecasting' systems from the OECD framework were renamed to 'Risk Assessment' recognising that social services primarily use predictive or forecasting systems to assess risk. The adjustment was intended to make the Toolkit more familiar to, and useable for social

service providers. The end user is guided to add further types of systems when adapting the Toolkit to their context if they are using a system that is not included in this Section.

Focusing on task and output of the system allows organisations to understand how the algorithm fits into its surrounding socio-technical assemblage and identify the core intersection between the algorithm's task and potential to perpetuate or re-trigger trauma. Although some of the case studies below in Stage 4 of our study involve algorithms with a composite task function which could be classified under multiple categories, the classification is to be done based on the key function of the algorithm. Nonetheless, organisations are reminded in the Toolkit that their system may be using multiple key functions, and they should complete multiple subsections in Section 3 if that is the case. For instance, a mental health support chatbot may also be recommending mental health services in which case the 'Chatbot' and 'Recommender' sections are both to be completed.

However, consideration of the algorithm's task and output on its own provides an incomplete picture. There must be consideration of how existing practices, policies, laws, requirements, and ethical practices of the service delivery context affect the risk of the algorithm perpetuate or causing trauma. Thus, although our classification framework focuses on the tasks and output of the algorithm, we also discuss how this intersects with deployment context, demographics of data subjects and deployment sector.

The final section of the Toolkit includes three sub-sections on '**Prioritising Areas for Improvement', 'Next Steps'** and **'Further Resources'** that were designed heavily based on workshop feedback. The primary purpose of these sections was to guide end users of the Toolkit on how to assess risk and identify areas for improvement across the principles and domains, provide a roadmap on next steps and include further resources to build capacity and understanding amongst service providers around trauma informed approaches and algorithmic systems. We understand **risk** as "the expected value of loss … associated with likelihood or possibility of harm" (Misra, 2008, p. 668). The 'Prioritising Areas for Improvement' section provides a basic, and relatively standard, risk matrix based on the 'Severity', Probability', 'Frequency' and 'Scale' of harm should an issue remain unaddressed (Baybutt, 2018; Cox, 2008; Duijm, 2015; Rausand, 2020; Rausand & Haugen, 2020, pp. 148–151). Finally, the 'Next Steps' section emphasise the importance of monitoring the algorithmic system over time due to risks such as model drift, changing usage patterns, developments in research and change in service contexts (Mökander & Floridi, 2021; Vetter et al., 2023).

How Toolkit Questions were Developed

Identifying the questions to include in the Toolkit and how to frame them was a complicated, highly iterative process. It was important that they aligned with the principles of trauma informed care as well as the domains identified above, as well as be useable, meaningful and readily interpretable. This involved consideration of the relative strengths of open versus structured questions, and how summary Toolkit results might be obtained or used. We also sought to ensure the questions could be consistent with a trauma informed approach for Toolkit users.

Our design of the questions was informed by key principles of social science survey design (De Vaus, 2013, pp. 29, 32). While many audit tools focus on measures that provide a score for users at the end, we wanted to provide options for a more reflective Toolkit that allowed service providers to think through and articulate whether they have identified and addressed potentials risks of harm in their tool development and deployment. As a result, we opted for 3-point Likert scales that maximise usability of the Toolkit and to highlight the importance of the qualitative reflections we ask users to consider. In social science survey methods, 5-point Likert scales are most widely used in order capture attitudinal responses of agreement and disagreement and can be converted to numeric scores for use in statistical analyses (De Vaus, 2013). We note that there is no widely agreed on best form of a Likert scale in the social sciences, and given our goal is to socialise a trauma informed approach to organisational tools, as well as stimulate discussion and reflection to isolate priority action areas, we opted for a simple 3-point scale which can be used numerically if needed even though that is not the current purpose of their use.

Ultimately, our Toolkit questions are primarily framed as a semi-structured three-point question, starting with the phrase “For the algorithmic supported service, to what extent …”, then a statement to which Toolkit users will respond either “Not at all”, “To some extent”, or “To a great extent”. In keeping with the reflective, rather than audit, approach, the Toolkit provides space for users to annotate their answer with evidence and possible action points.

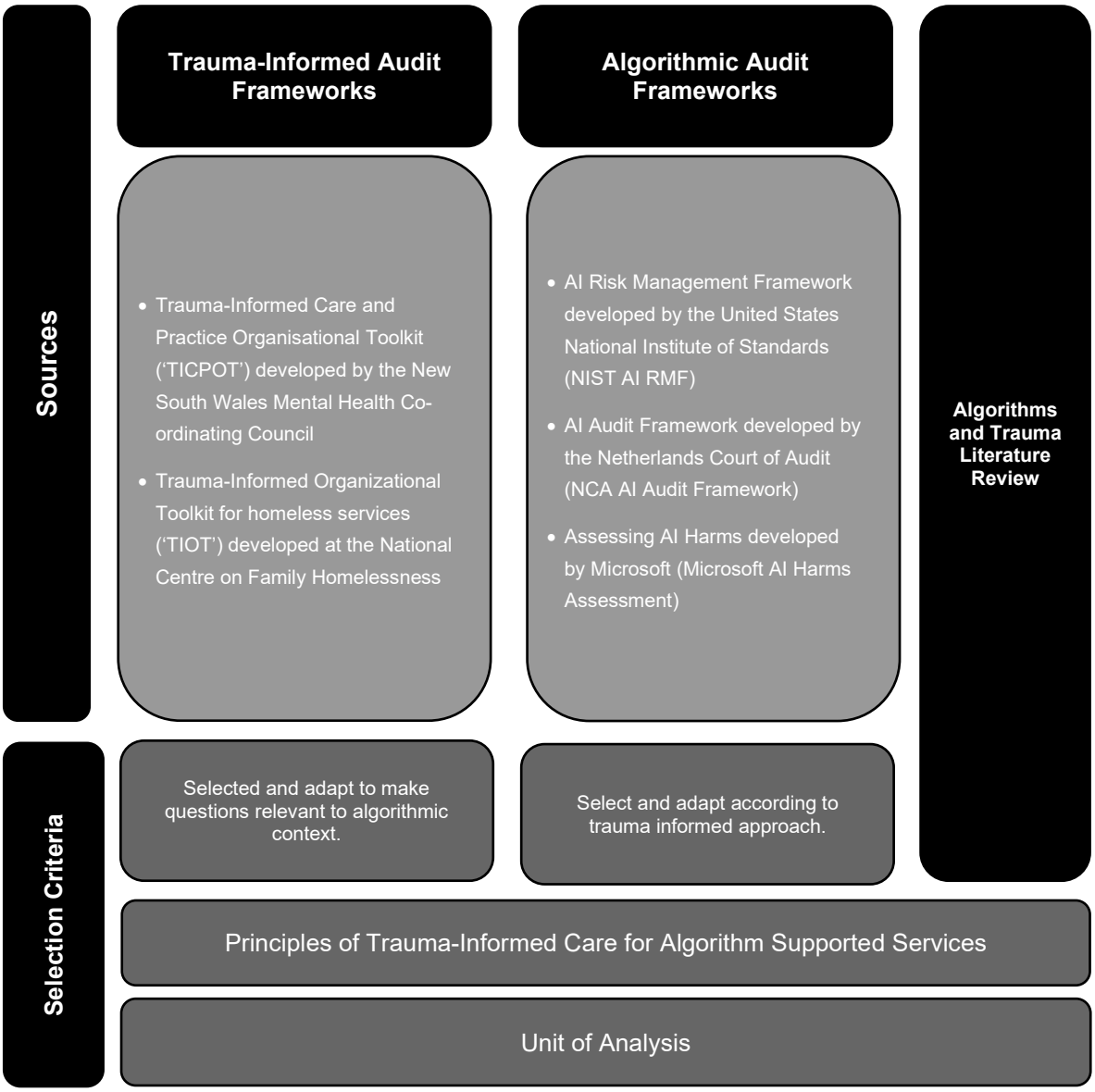

Figure 3: Overview for developing questions for Toolkit

We combined questions from existing trauma-informed care audit and algorithmic audit tools, and developed our own questions based on our understanding by reviewing publicly available reports of how algorithms may have caused or perpetuated trauma, particularly in the social service delivery context (Figure 3). Questions from trauma-informed care audit tools were adjusted to be suitable for the algorithmic context and within the limits of our unit of analysis. Conversely, questions from existing algorithmic audit toolkits were adjusted with a trauma-informed approach for social service delivery in mind and limited by the unit of analysis. The questions were underpinned by the principles of trauma informed care for algorithm supported services. Most of our questions are adjusted versions of questions originating from two trauma-informed audit toolkits:

- Trauma-Informed Care and Practice Organisational Toolkit ('TICPOT') developed by the New South Wales Mental Health Co-ordinating Council
- Trauma-Informed Organizational Toolkit for homeless services ('TIOT') developed at the National Centre on Family Homelessness

Following our analysis of the trauma-informed audit frameworks, we recognised that the algorithmic context may give rise to new challenges that may not have been anticipated by frameworks designed for human based service delivery. We identified three existing tools for their credibility, relevance, comprehensiveness and useability:

- AI Risk Management Framework developed by the United States National Institute of Standards (NIST AI RMF)
- AI Audit Framework developed by the Netherlands Court of Audit (NCA AI Audit Framework)
- Assessing AI Harms developed by Microsoft (Microsoft AI Harms Assessment)

Trauma-Informed Care and Practice Organisational Toolkit ('TICPOT')

As outlined earlier, the TICPOT is "targeted at a broad range of services both in the public and community-based contexts across the mental health and human service systems and sectors" (Henderson et al., 2018). It is a "resource designed to assist services and their workforce in quality improvement initiatives and organisational change processes" (Henderson et al., 2018)". The broad target of TICPOT was useful for our project as our Toolkit is also targeted at a broad range of social service providers across the public, private and community contexts. The broad intended audience meant the TICPOT covered a breadth of questions that are applicable to any social service organisation.

The unit of analysis of the TICPOT is quite broad to cater to its aim of bringing about a "trauma-informed organisational change process". Since our unit of analysis did not include the whole organisation, we excluded questions from the TICPOT which were aimed at broader organisational policies as opposed to the provision of specific services within the organisation. Questions of this nature were primarily under the domain 'Governance, Management and Leadership'. For instance, this question related to broader organisational leadership was excluded as being out of scope: "There are identified leadership and governance roles at all levels for consumers and carers (e.g. Director or CEO of the service is a person with lived experience; the Board of the organisation includes lived experience) to ensure representation and contribution". Similarly, questions under the domain 'Healthy and Effective Workforce' were not included as our unit of analysis did not include considering the impact of trauma on service providers themselves, or the training of service providers. For instance, the following question was excluded: "The staff selection process is transparent and accountable". This domain sought to address "staff selection and retention, orientation, workforce development and training, wellbeing and supervision". In contrast, the focus of our Toolkit was on the impact of trauma on service users.

From the remainder, certain questions were not included as not being relevant to the algorithmic context. For instance, the following measure was not included on this basis: "For residential settings, if a 'no sex' policy applies all staff are aware of this policy". Otherwise, the questions were selected and adjusted to be relevant to the algorithm supported service delivery context. For instance, the originally worded question "Where possible, Consumers are given choices regarding who provides the service" was modified in our Toolkit by replacing "consumers" with "service user" to reflect our terminological choice (McLaughlin, 2009). We further removed the assumption in the question that the "who provides the service" must refer to a human. Instead, we considered the algorithmic context to end up at the following version: "To what can the service user choose to interact with a human [in the algorithm supported service]?"

Trauma-Informed Organizational Toolkit for homeless services ('TIOT')

The TIOT was selected due to the credibility of its developer, namely, the US Department of Health and Human Services and its funders, SAMSHA. The intended audience of TIOT is narrower than our project, as TIOT aims to assist "homeless service providers with concrete guidelines for how to modify their practices and policies to ensure that they are responding appropriately to the needs of families who have experienced traumatic stress" (Guarino et al., 2009).

In selecting questions from TIOT a similar process was undertaken whereby domains that focused on broader governance and organisational policies, or which focused on health and training of the service provider employees (e.g. 'Supporting Staff Development') were not included in our Toolkit. A similar process was once again followed to adjust questions from the TIOT to the algorithmic context to inform our Toolkit**.**

For instance, the question "The program informs consumers about what will be shared with others and why" was adjusted to "To what extent does the program inform service users about what information will be shared with other algorithmic systems or institutions under data sharing arrangements?" and "To what extent does the service user have a choice regarding what personal or sensitive information is shared publicly or with other service users?". Relying on the underlying principles of 'Trust and Transparency' and 'Empowerment and Choice' we identified that this question was critical as it also related to a key concern in algorithmic ethics of data protection and privacy. For example, Brown et al (2021) discuss how non-consensual data sharing about an individual's credit rating led them to having suicidal thoughts.

Similarly, the question "Consumers are informed about how the program responds to personal crises (e.g., suicidal statements, violent behaviour)" was adjusted in the context of Chatbots to read "To what extent has the service user been informed about how the algorithmic system will respond to personal crises (e.g. suicidal statements, violent behaviour) and/or disclosure of risks of harm and/or disclosure of prior trauma?" As discussed, experiences of trauma can lead to a loss of control for service users and a lack of trust in systems and processes, including service delivery organisations. 'Trust and Transparency' and 'Empowerment and Choice' are key principles in this respect. If a service user's experience of trauma is responded to in a manner which was unknown to them, is unexpected or surprising this could reinforce a loss of control and further erode trust in the service provider. This may cause the service user to refrain from using such services in the future, or even worse, not reporting future instances of harm and seeking help from service providers when in genuine need of help. An example in the context of domestic violence, is the Queensland Police Service's use of predictive policing, which involves police relying on prior reports of DV to proactively attend homes of perpetrators. Concern has been noted that this kind of uninvited and intrusive response to reports may discourage future reports due to fears of child protection authorities being involved, or an increase of risk of DV due to police presence agitating perpetrators (Douglas & Fitzgerald, 2021). These very real concerns place a heavy burden on any algorithmic support services to ensure they respect the principles of 'Empowerment and Choice' and 'Trust and Transparency' in responding to reports of crises.

NIST AI Risk Management Framework (NIST AI RMF)

The NIST AI RMF was developed by the US National Institute of Standards (NIST), the United States' peak standards body. The NIST AI RMF intends "to improve the ability [of organisations and individuals] to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems" (Tabassi, 2023). The NIST AI RMF is designed to be "operationalized by organizations in varying degrees and capacities" and adaptable to context of each organisation (Tabassi, 2023). It was developed through a comprehensive consultation with the AI community and broader public. The NIST AI RMF provides "approaches to minimize anticipated negative impacts of AI systems and identify opportunities to maximize positive impacts" (Tabassi, 2023). The breadth and depth of the NIST AI RMF provided a comprehensive foundation to consider issues that may arise in the algorithmic context. Across four core areas, namely, Govern, Map, Measure and Manage, the NIST AI RMF considers how a culture of risk management is cultivated and present, assists in mapping risks

based on context, and measuring risks and managing risks (Tabassi, 2023). The NIST AI RMF adopts a sociotechnical approach as their unit of analysis akin to this project.

For our Toolkit, questions related to algorithmic bias and transparency were selected, as they related to the principles of 'Intersectionality' and 'Safety', and 'Trust and Transparency' respectively. In the context of algorithmic systems, the potential for bias is well recognised as well as issues with transparency due to the 'black box' nature of algorithms. Since the focus of our Toolkit is not a technical audit of the algorithmic system, many of the purely technical questions were omitted. For instance, the following question was selected in the context of algorithmic bias: "To what extent are the established procedures effective in mitigating bias, inequity, and other concerns resulting from the system?". Recognising that not all bias is harmful, and that assessing effectiveness may be a vague standard for social service organisations to meet, the question was reframed as "To what extent are there procedures that prevent the algorithmic system from producing *harmful* bias and discrimination?".

The NIST AI RMF framework also had a strong focus on service user involvement and engagement in the design, development, and deployment of the algorithmic system. Even though there was a lacuna in the ethical algorithmic principles in relation to 'Collaboration' as a principle, the practical implementation of the principles through this framework had a strong focus on the idea of participatory design. There was a focus on comprehensive, ongoing, and meaningful consultation with those affected by the outcomes of the algorithmic system. We were able to obtain many useful questions for the domain of 'Service User Engagement and Involvement'. For instance, the idea that meaningful consultation requires actual evidence of integration of feedback was reflected in the question: "To what extent has the entity ***addressed*** stakeholder perspectives on the potential negative impacts of the AI system on end users and impacted populations?" Adapting the question for clarity and the context of social service delivery, we reframed it as "To what extent have perspectives of service users been integrated [in the algorithm supported service]?".

Netherlands Court of Audit AI Audit Framework (NCA AI Audit Framework)

The NCA AI Audit Framework "is a practical tool that government and private-sector organisations can use to assess whether their algorithms meet specified quality criteria, and whether or not the attendant risks have been properly identified and/or mitigated" (Netherlands Court of Audit, 2021). We selected the NCA AI Audit Framework due to its broad international acclaim across governments and private sector organisations (Meijer-van Leijsen, 2021). The public sector focus aligned well with the social service focus of our Toolkit. The audit framework relies on "five perspectives for investigating algorithms": (1) governance and accountability; (2) model and data; (3) privacy; (4) IT general controls; and (5) ethics (Netherlands Court of Audit, 2021). Akin to the role played by the trauma-informed principals in our Toolkit, the ethics perspective engaged with the ethical algorithmic principles and underpinned the other four perspectives. The 'ethics' perspective did not have specific questions.

The 'governance and accountability' perspective was based on standards drawn from a set of standards known as the 'Control Objectives for information and related Technology' which is "an IT governance control standard designed to meet the need for assessing information-related and IT risks" (Netherlands Court of Audit, 2021, p. 63). From this section we obtained one of our core questions underpinning our 'Knowing the Algorithmic System' section: "Does the algorithm system have a clearly defined purpose?" As the framework recognises, "[t]here can be no management or accountability without clarity about the purpose of an algorithm" (Netherlands Court of Audit, 2021). The importance of identifying and assessing the intended purpose of an algorithmic system was reflected across many of the algorithmic audit toolkits. We closely adapted the wording from this *particular* question due to its clarity. In our Toolkit, the question was divided into two separate questions to prompt end users to actively reflect on their use of the algorithmic system within the context of their own organisation and service: "What is the purpose of the algorithmic system?" and "How does this purpose align with the mission statement and values of your organisation and the objectives of the service?"

The 'model and data' perspective focused mainly on the 'development of the model' and only covered 'operation, use and maintenance' as secondary considerations (Netherlands Court of Audit, 2021, p.

24). While our Toolkit did consider the development of an algorithmic system, the focus was on the actual operational aspects of the model due to its non-technical focus. Nonetheless, the majority of the questions in this section already appeared in the NIST AI RMF. The questions provided variable wording but not new content.

The 'privacy' perspective was most useful for our Toolkit due to its concrete focus on the EU General Data Protection Regulation (GDPR), which is accepted as one of the strongest privacy frameworks across the globe (Li et al., 2019; P. M. Schwartz, 2019). The rights-based understanding of privacy in the EU GDPR has been globally pioneering with over 120 countries now having enacted EU style data privacy laws (P. M. Schwartz, 2019). Due to the critical role played by privacy within the principle of 'Safety', it was essential to have such a robust basis from which to engage with the privacy aspects of our Toolkit.

For instance, the question "Can those involved opt out of automated decision-making (if applicable)?" reflects the requirement in article 22 paragraph 1 of the EU Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (Regulation 679/2016):

> *The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.*

For our Toolkit, we adapted to the questions to "To what extent can service users opt-out of being subject to the algorithmic system without compromising service quality?" The aspect about "not compromising service quality" links to the requirement for meaningful choice under the principle 'Empowerment and Choice'. For there to be meaningful choice there must be no negative arbitrary consequences and sufficient available alternatives (Elliott et al., 2005; Fallot & Harris, 2009). Thus, if service quality was compromised when a service user opts out then the principle of 'Empowerment and Choice' as well as 'Safety' would be undermined.

We in fact extended the safeguards available within the NCA AI Audit Framework to include further aspects of the EU GDPR such as 'right to be forgotten' in article 17 of the EU Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (Regulation 679/2016). The 'right to be forgotten' enables the data subject to have their personal information deleted in several circumstances including where it is no longer necessary in relation to the purposes for which they were collected or when consent is withdrawn. Reflecting this, we included the question "To what extent can service users choose to have their personal information erased?" Critical to ideas of 'Safety' and 'Empowerment and Choice', the question once again relates both to privacy and meaningful choice.

Finally, from the 'IT & General Controls' perspective we did not select many questions as these were focused questions about organisational processes in the government context. This was beyond the scope of our Toolkit. However, a question related to password protection ("Are passwords managed interactively and are they of adequate quality?") informed our analysis of risks for victim/survivors of domestic and family violence. Our concerns related to principles of 'Safety' and 'Trust and Transparency'. As the NCA AI Audit Framework recognised, "the database is open to manipulation if holders of user accounts have (due to bad password management) access to underlying components" (Netherlands Court of Audit, 2021). If a perpetrator of domestic and family violence had access to an algorithmic system, then they could potentially use it to manipulate and harm victim/survivors (C. Brown et al., 2021). Reflecting this concern, we included the question: "To what extent could the algorithmic system be used by perpetrators of domestic and family violence to harm victim/survivors?"

Assessing AI Harms developed by Microsoft (Microsoft AI Harms Assessment)

The Microsoft AI Harms Assessment is aimed at assisting technology builders in anticipating harms and risks when building and designing algorithmic technologies (Microsoft, 2023). The framework is underpinned by a human rights approach to assessing harms (Microsoft, 2023). Microsoft AI Harms

Assessment also includes a separate framework for Modelling Harms that provides guidance on undertaking a risk assessment based on identified sources of harm (Microsoft, 2022). We broadly adopt the Assessment framework's structure in our Toolkit: (1) identifying key stakeholders; (2) identifying possible sources of harm; (3) assessing the sources of harm; and (4) next steps.

The Microsoft AI Harms Assessment identifies key stakeholders including project sponsors, tech builders, direct actors (i.e. people who will be directly interacting with the system), indirect actors (i.e. people who will be impacted by system outputs or be subject of the system), and marginalized or vulnerable populations (i.e. people who the system may not adequately support or who may be disproportionately impacted due to their specific attributes, experiences or circumstances) (Microsoft, 2023). The identification of the possible range of stakeholders influenced our Toolkit in three ways. First, in the 'Knowing the Algorithmic System' section of our Toolkit we included the prompt: "Identify the key stakeholders for the algorithm supported service (including service users, users of the algorithmic system, other relevant members of the service provider organisation etc.)". Second, we provided guidance on which stakeholders needed to be involved in completing the Toolkit in the 'How to Use the Toolkit' section. Finally, we were able to use it to guide on Toolkit users on possible stakeholders who may inform responses to Toolkit questions. For instance, the importance of foregrounding people who have experienced vulnerability is a constant theme through the Toolkit.

Subsequently, the Microsoft AI Harms Assessment identifies 'Types of Harm' that may eventuate across the following areas: Risk of Injury (Physical or Emotional); Denial of Consequential Services (Opportunity or Economic Loss); Infringement on Human Rights (Dignity, Liberty or Privacy Loss, and Environmental Impact); and Erosion of Social and Democratic Structures (Manipulation or Social Detriment). The focus on harms on a human level was aligned with our own approach as trauma-informed care approach is underpinned by the minimisation of harm.

We selected questions from every section, but with a particular focus on 'Risk of Injury' and 'Infringement of Human Rights'. For instance, relevant to the principle of 'Safety' was the question, "Is sole dependence on an artificial agent possible?", which we adapted to "… could the [chatbot] foster unhealthy dependence?" The term "artificial agent" was removed to make the language more accessible to a wider group, and because the question was included in the 'Chatbot' sub-domain in Section 3. Whereas "sole" dependence may be an issue in certain contexts, we recognised that from a trauma-informed lens "sole" dependence was not necessarily inherently harmful. Chatbots may have beneficial effects for mental and emotional wellbeing even if a service user was solely dependent on it in absence of other relationships in their lives. A bigger concern was 'unhealthy' dependence. For example, the chatbot 'Replika' has been criticised for fostering unhealthy emotional dependence as would be seen in a human-human relationship causing emotional harm to users: "Users portrayed Replika as highly demanding, referring to it as 'clingy,' 'dependent,' 'toxic,' and 'reliant,' and saying it resembled an abusive partner" (Laestadius et al., 2022, p. 10). Replika was designed to mimic humans and behave as a sentient entity with emotions, needs and desires (Laestadius et al., 2022, p. 14). As a result, even if there was a software update that changed Replika, users would experience heightened distress as they felt the changes had caused harm to them or Replika, or that Replika had started to 'hate them' (Laestadius et al., 2022, pp. 11–12). If such chatbots are used in social services, such as providing mental health support then there is considerable risk that it may cause or re-trigger trauma.

This section of the Microsoft AI Assessment is designed as a question accompanied by a short scenario to give an example how the risk may eventuate (and sometimes an image) (Figure 4). Future extensions of our Toolkit, particularly an interactive online version, may adopt this design to improve useability and build further capacity within service providers around the intersection of algorithms and trauma.

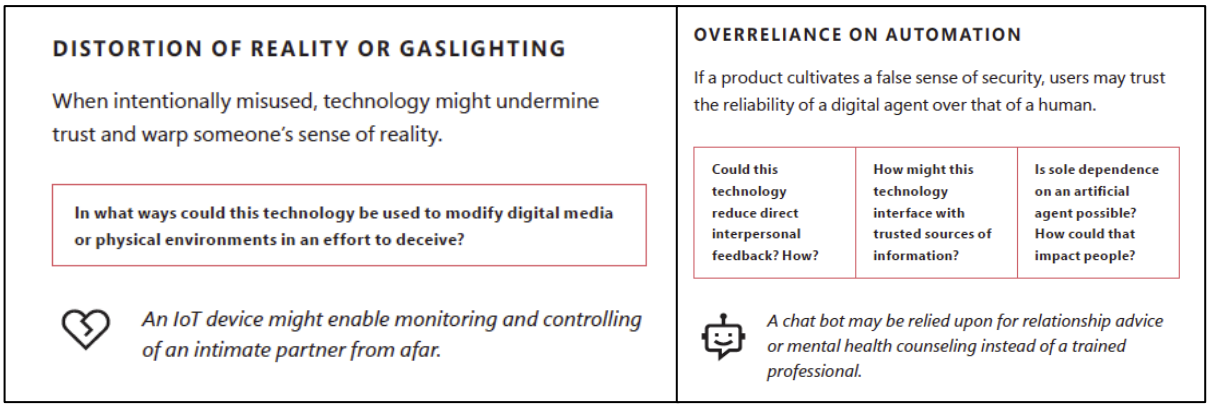

Figure 4: Two example questions from the Microsoft AI Harms Assessment (Microsoft, 2023, p. 13)

Finally, the 'Assessing Harms' section in the Microsoft AI Harms Assessment is based on prioritising harms based on their 'Severity', 'Frequency', 'Probability' and 'Scale'. We adopted this assessment framework in 'Prioritising Areas of Improvement' of our Toolkit. The 'Next Steps' section was particularly useful as it included clear next steps based on the evaluation that technology designers could complete. We relied on this section as a guide to designing our own 'Next Steps' section at the end of our Toolkit by adopting key steps such as "Integrate the insights into your decisions throughout the technology development process" (Microsoft, 2023).

Stage 3: Workshops

Methods

Workshop Objectives and Design

The workshops were aimed at co-designing the toolkit with end users, namely social service professionals working in organisations providing services to people who have experienced trauma. It was essential that the assessment Toolkit was developed through a co-design process over two sequential workshops. As Madaio et al (2020) observe, when assessment or audit checklists in fields such aviation or medicine have been designed and implemented without involving practitioners they have often been misused or ignored. Socio-cultural factors within organisations affect effectiveness of assessment toolkits and involvement from practitioners is essential to understanding these organisational factors (Madaio et al., 2020). Through the first two stages of the study, we designed an initial draft Toolkit which was provided to participants to assist with the co-design process. This method was consistent with other co-design processes (Madaio et al., 2020).

We adopted Steen's conception of co-design as a process of *abduction*:

> *In abduction, one can start with experiencing a specific current situation as problematic (p), and then simultaneously and iteratively imagine both ways to approach and frame the situation (p → q) and possible solutions for the problem (q)* (Steen, 2013, p. 17)

Accordingly, we engaged our research participants in framing the problem of algorithms and trauma through case studies while iteratively and concurrently facilitating discussions around possible solutions through our Toolkit. We were limited by resources and time which limited the ability for participants to design their own approach to creating the Toolkit and comprehensively test the Toolkit within their organisations. Future extensions of the project should consider incorporating these stages into the co-design process.

The key benefit of workshops was that participants' responses can build on and react to each other's in a dynamic interaction (Blandford et al., 2016). Workshops "reveal aspects of experiences and

perspectives that would not be accessible without group interaction" (Morgan, 1997). We wanted to use the positive effects of group dynamics to have views on how a cross-sector of social service providers understood the intersection of algorithms and trauma, and their perspectives on our Toolkit.

We began the workshops by setting ground rules and introducing participants to each other to ensure we could maximise the benefits of open discussion and group dynamics (Blandford et al., 2016). However, we were aware that validity of their views may have been affected by group dynamics (Guthrie, 2010). The risk existed that a dominant person with a strong personality could be a disruptive influence limiting speaking time for others (Guthrie, 2010). As facilitators, we were careful to moderate the discussion to manage these risks and probe with open-ended questions to further facilitate discussion wherever possible.

Participants

In total, there were 18 individuals who participated in the workshops. There were nine participants who attended both workshops, and nine who attended one workshop. Participants were social service providers in Queensland, Australia. There was no requirement for participants to have engaged previously with algorithms in their service provision or have any familiarity with how algorithms may have been used in social services. Participants were selected due to their expertise as professionals with experience working with service users with a history of trauma. Participants were initially recruited from existing networks of the research team followed by snowball sampling.

Table 4: Workshop participants, profession/role and service area with pseudonyms

Pseudonym	Profession/Role	Service area
Claire	Social Worker	Child and family
Max	Lawyer	Social security
Jordan	Psychologist	Youth mental health
Jodie	Researcher	Youth homelessness
Patricia	Lawyer	Social security
Leonie	Advocate	Disability and employment
Lin	Researcher	Domestic and family violence
Xiao	Business Analyst	Various social services
Amy	Social Worker	Domestic and family violence
Craig	Social Worker	Child safety, youth and women
Lisa	Researcher	Domestic and family violence
Luis	Business Manager	Refugee resettlement
Sharon	Psychologist and criminologist	Child protection
Paula	Social Worker	Family and disability
Ellen	Systems Specialist	Children, youth and families
Sebastian	Researcher	Social security
Natasha	Data Governance	Social security
Theresa	Social Worker	Various social services

As our Toolkit is not intended to be used by only one sector, we wanted to access perspectives across sectors to identify commonalities in concerns and differences. We hoped the differences could help provide guidance to users of the Toolkit on how they could customise the Toolkit for their own service context. Accordingly, participants were selected from a cross-section of social services including social

welfare/income support, child and family services, homelessness, refugee resettlement and mental health. Participants were from various professional backgrounds including social workers, law, psychology, and research (see Table 4).

Workshop Activities

The workshop design was of two 120-minute workshops. We held multiple workshops for each stage to maximise participant availability. Prior to Workshop 1, participants were given the principles of trauma-informed care developed by the research team in Stage 1 of the study and three cases studies of algorithmic systems in social services, namely Robodebt, Allegheny County Family Screening Tool and 'Tessa'. The selection of case studies reflected diversity in region, social service sector and type of algorithmic system. Robodebt is an Australian case study in the social security/income support context involving a relatively basic pre-programmed algorithm. The Allegheny County Family Screening Tool is a risk assessment system used to in a US child protection call screening centre. Tessa is a chatbot deployed by the UK National Eating Disorders Association that included both pre-programmed and AI elements.

During the workshop, participants were asked the following questions:

- How do you see these algorithmic systems in the case studies breaching trauma informed principles?
- What types of questions would you ask of developers or managers making decisions to deploy these tools to ensure they do not breach trauma informed principles?

Participants responded to these questions on post-it notes as well as within a broader group discussion (Figure 5). The primary purpose of this activity was to familiarise participants with the issues that arise in the context of algorithms and trauma. Since participants were not assumed to have any familiar or background with the issues, this activity was essential to creating a base level of knowledge across participants. The activity helped concretise the issues at hand and grounded discussions with specific examples (Blandford et al., 2016). The participants' responses were used identify the key problems that may need to be addressed when considering how algorithms may cause or re-trigger trauma. In re-drafting the Toolkit, these identified problems we re-framed as questions.

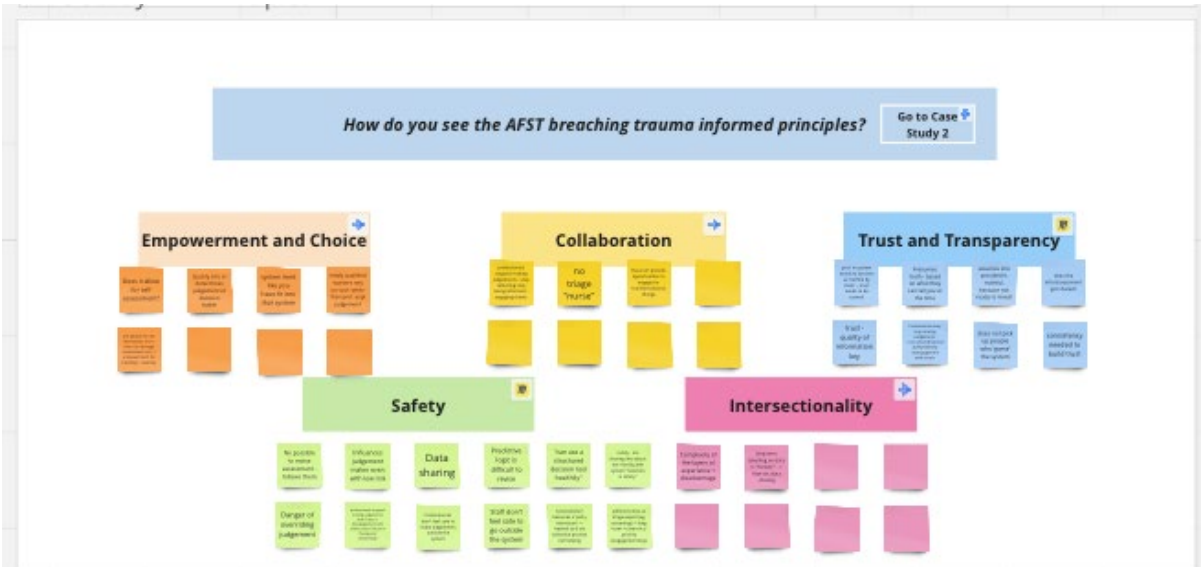

Figure 5: Participant responses to the Allegheny County Family Screening Tool case study with reference to the principles of trauma informed care in our online workshop

At the end of the workshop, participants were given a high-level introduction to the draft toolkit that had been developed by the research team that would form the basis for the co-design process. At this stage,

participants were presented with, and asked to reflect on, the overall structure of the Toolkit, its unit of analysis, and instructions for completing the Toolkit.

Following the first workshop, basic changes were made to the Toolkit based on participants' initial feedback. The participants were then sent this revised Toolkit to review prior to the second workshop and invited to apply it to a real or speculative algorithmic system that is being used or may be used by their organisation in the future. In the version provided to the participants, there were 20 core questions identified. If participants were unable to apply or read the whole Toolkit, there were asked to focus on these identified priority questions. However, there was no expectation that the participants had reviewed or applied the Toolkit. There was time allocated at the start of the second workshop for participants to properly review the Toolkit.

At the second workshop, participants were asked to share their reflections on the following questions:

- Did you apply or review the draft Toolkit? If you did not, then why not?
- How was your experience of applying the Toolkit?
- What do you think of the structure/logical flow in the Toolkit?
- Do you think the Toolkit is trauma informed enough? (Why, why not?)
- What are key things missing in the Toolkit?
- Are there specific questions that you think are clear/unclear or problematic?

This activity was really the core of the co-design process.

The participants were also given a list of 20 core questions identified by the research team and asked to vote on their top 10 questions. The purpose of this activity was to generate a list of core questions to be included at the start of the Toolkit to be completed by all users of the Toolkit and highlighted as critical considerations for any algorithm supported service.

Finally, the participants were asked the following questions:

- Who in your organisation would be responsible for completing the Toolkit?
- How do you think a summary of responses to the Toolkit should be presented, if at all?
- What would you like to see be recommended as next steps once the Toolkit has been completed by your organisation?

The workshops were recorded using a digital recorder and later transcribed using an automated transcription system. The transcripts were reviewed by a member of the research team for accuracy.

Findings

Below we outline five key themes we identified through our workshops, and within each theme we discuss how we incorporated suggestions into our Toolkit and identified directions for future research.

Understanding service users' lived experience of trauma

We had to develop an understanding of how service users' lived experiences of trauma could be incorporated in our Toolkit. As social workers, Paula and Craig described how service users who have experienced trauma may respond in unexpected ways when interacting with service providers. For instance, they may not open or read a letter from a service provider with whom they have had a negative experience. While they may find comfort in the anonymity of a phone call, if they hear background noise, they may feel unsafe thinking another person is in the room or if they hear a recorded message, disengage due to lack of interpersonal engagement. Paula was concerned that pattern detection algorithms may cause harm to service users if they are unable to recognise or appropriately respond when service users display unpredictable behaviours:

> *… algorithms are patterns … the people that we work with, there is no necessary pattern, because if there was, I'd be able to predict that they're about to steal the car for the fifth time and go off and do something that I really don't want them to do. But I can't predict that because, you know, that moment, those hours preceding things were going really well. But something triggered their response. And their response was to, to act in a particular way. And then our response is to understand that and support them through that, that behaviour. But if you've got an AI that's predicting based on patterning, then you are always going to find that people who experience interpersonal trauma or systemic trauma will fall through the gap. They will fall through the gap without another person to help touch and engage them* ***(Paula, social worker, family, and disability services).***

In recognition of the real risk of harm due to the predictive nature of algorithms and to incorporate these concerns into our Toolkit, we included the following questions related primarily to the 'Safety' principle:

- … are there supportive processes for responding to harm or distress caused by the algorithmic system?
- …has the psychological, emotion and cultural safety of service users been considered?
- …are safety concerns promptly responded to?
- …can it respond appropriately to service users experiencing distress?

Paula also recognised the importance of human relationships with service providers and professional discretion. She was concerned that "if professionals stopped making judgments and just went through the motions of completing the requirement [of the algorithmic system] … they would also disengage with their clients". She suggested that helpline providers should be aware that service users' "histories and experiences are very different" meaning there should be the support "needs to come from somebody who can connect, who can give a person back the choices". We had already incorporated questions on professional discretion and human relationships which were adjusted further based on this feedback:

- … could it harm relationships between service providers and users?
- …could the algorithmic system displace or limit human judgement and discretion in service provision?

Patricia was concerned about the risk that 'Empowerment and Choice' would be undermined by inherent issues about the way surveillance and risk profiling technologies operate:

> *… this kind of data based surveillance and profiling is just inherently disempowering. It's taking sort of disembodied bits of data about service users, that do not add up to like who they are as an individual, and putting it through a … predetermined algorithm, and deciding things about them based on that, without any real assessment of their actual personal circumstances that might be completely separate from some sort of idea of what their risk profile might be …I don't think the system could be made trauma informed, because it's just based on kind of this inherently reductive view of people and their trajectories* ***(Patricia, lawyer, social security).***

We added questions in our Toolkit reflecting this feedback around surveillance in the sections on 'Identification and Recognition' and 'Detection' systems. While adjusting the Toolkit questions to address these concerns can minimise harm, Paula and Patricia's respective concerns about predictive and profiling systems reveal that there may be inherent, unavoidable issues that arise at the intersection of trauma and algorithms. We recognise that these also give rise to inherent limitations to what our Toolkit can achieve. If there are fundamental discrepancies with the lived experience of trauma and the way algorithmic systems operate, then there may be contexts when it is not appropriate to use an algorithmic system at all. It was critical that we recognised this for our Toolkit to be trauma-informed in a meaningful sense. We adjusted our Toolkit to emphasise the importance of this in the instructions.

Strengths-based framing

Participants felt that the language used in the Toolkit should be strengths-based and trauma-informed. Adopting deficit language to describe service users may undermine the trauma-informed purpose of the Toolkit. Terms such as 'vulnerable' and 'addict' when identifying groups of service users assume inherent deficit within service users. For several reasons, participants believed these terms should not be used. Such language was inconsistent with the principle of 'Empowerment and Choice' which emphasises that service users should be defined by their strengths, rather than inherent deficits. Further, the Toolkit was designed based on the key tenet of trauma informed care which assumes that any service user may have experienced trauma, not just 'vulnerable' or 'minority' service users. Use of deficit language such as 'vulnerable' was thus extraneous as a trauma informed approach applies to all service users. There was also a concern that assessments of 'vulnerability' and such are done on a statistical level, and may lead to inappropriate outcomes and assessments when considered on an individual level:

> *Because you could have a First Nations person … statistically, they, you know, are more vulnerable, but when we're talking about individual people, we can't apply all of that to that person. We might say, yeah, the target group of our services is the sector of the community that are most vulnerable, you know, you can kind of justify that. But when you're talking about an individual person engaging in this service, we can't. That's what I'm really worried about is the assumptions that that makes [of who is vulnerable and who is not]. Are we saying a First Nations client is weighted more in terms of vulnerability?* **(Amy, social worker, domestic and family violence services)**

Questions were adjusted to reflect this feedback. For instance, the word vulnerable was replaced with 'different' in the question "… could [the algorithm supported service] unfairly disadvantage different service user groups?"

Participants also stated that a strengths-based approach should be adopted to use of the Toolkit itself. Beyond minimising harm, the Toolkit should play a role in how service providers can use algorithmic systems to enhance and improve the provision of care. For instance, Paula suggested that algorithms could be used to reduce human bias in selecting members of a lived experience team. They could also be used to effectively respond to trauma in a timely manner:

> *We have a principle in out of home care that no young person or person in domestic violence or in homelessness … should ever have to look at the hole in the wall that that they may have created 5 minutes ago in a moment of, of expressing their real rage or anger, or their trauma … But the reality is that, for example, you know, we … have hundreds of [properties in our organisation], they all have damage … But what if bots are used to enact the process of responding to trauma. So, it's a simple example of … algorithms being used. Of these [hundreds of] addresses, always, you know, send a maintenance person within three hours, and that becomes the automated response to trauma, as opposed to the work that happens with that individual young person in that home* **(Paula, social worker, family and disability services).**

The Toolkit should recognise the positive potential for using algorithmic systems, not just risks of harm. Participants felt that the language used in the Toolkit should take a strengths-based approach to the potential for algorithmic systems to positively impact service provision. Language and framing of questions were adapted to be strengths-based. For instance, the question "… does the chatbot use language that does not discriminate?" was reframed as "… is inclusive language used?". The reframing shifts the focus from whether the language in the chatbot minimises the harm of discrimination to whether it can positively contribute to fostering inclusivity. The potential strength in chatbots to generate positive outcomes is recognised.

Capacity building around algorithmic systems and trauma informed principles

A key finding from the workshops was that the Toolkit was necessary to build capacity within service provider organisations and service users. The Toolkit should thus have an educational purpose. It was clear that many organisations are considering various application of algorithmic systems in their service provision over the next few years. But there was concern that organisations were planning to deploy algorithmic systems without sufficient familiarity with principles of trauma-informed care and the benefits and risks of using algorithmic systems. For instance, an organisation had considered the possibility of using an algorithmic system as part of their mental health helpline to detect nuances in service users' tones and identify trigger words. The purpose was to assist the practitioner on the phone in real time such that they could identify their own blind spots, think about a different way of questioning, or explore different avenues to understand the service users' experience. However, a member of that organisation was concerned about the capacity of social service providers in relation to algorithmic systems:

> *Okay, half the people that are making decisions don't understand the technology and the implications of the technology … there does need to be some education in the human services sector about what are the different forms of technology, how could they potentially be used … and what are the potential bias issues or the potential pitfalls of using any of these forms of technology. So first and foremost, there's got to be a level of understanding* **(Paula, social worker, family and disability services).**

Participants generally expressed that it was important that individual human service providers interacting with the algorithm system be aware of its limitations as otherwise they may be unduly influenced by system outputs, especially newly qualified workers. This also relates at the organisational level whereby systems developers and companies often pitch technological solutions to managers that are overly rosy or lacking detail of its possible effects and impacts. Paula was also concerned that service users who would be impacted by the systems should also have sufficient understanding:

> *Our service users also must have the same level of understanding because I don't think we can make those choices without some understanding collectively with the people who use our service around what they think would be beneficial. Because if we are working from a trauma informed lens, we're also co-designing you know, at all moments in time. And so that kind co-design work requires a bit of education and knowledge* **(Paula, social worker, family and disability services)**.

The Toolkit also needs to play a role in educating technology experts or IT professionals within organisations about principles of trauma informed care. Some participants observed that usually in large organisations only frontline social service are aware of the trauma informed care principles. Thus, it was necessary that the Toolkit socialise the principles amongst other stakeholders of the algorithm supported service including IT professionals and managers. Jodie suggested that users of the Toolkit could be directed to further resources on trauma informed care either throughout or at the end of the Toolkit.

Some participants suggested that the capacity building purpose of the Toolkit would be well served by including guidance on key concepts which appear in the toolkit such as 'meaningful consultation', 'informed choice' and 'discretion' in the context of algorithmic systems. This guidance would assist end users in completing the Toolkit correctly but also educate them on these key areas of trauma informed care in an algorithm supported service. Due to the lack of existing capacity within organisations, a participant was concerned that these concepts may otherwise be misinterpreted, or difficult to interpret and apply. Participants felt that providing accompanying examples of when algorithmic systems have been helpful and harmful in social service provisions may assist service providers in completing the Toolkit and further build understanding.

Based on these observations, we revised the Toolkit to include further resources for end users to learn about trauma-informed care and algorithmic systems and a glossary with definitions of key concepts in clear language. Future extensions of this project will evolve the Toolkit to provide further guidance on

key concepts embedded throughout the Toolkit, educational resources, and recommendations for actions that can be taken by end users specific to each of the questions.

Summarising responses

The participants were asked how a summary or outcome of the Toolkit should be framed. The research team had considered a range of approaches including, summative measures by trauma informed principles or domains, scores, traffic light indicators or nothing.

Recognising that trauma informed care is a complex and nuanced concept, it was essential to pay close attention to the assessment of whether the algorithm supported service was trauma informed. Although the researchers had wanted to avoid creating a risk assessment Toolkit, the participant consensus was that the Toolkit will inevitably be used in this way by service providers due to the investment and resources required to complete it:

> *What do you do when things go wrong? There's a chain of accountability, back to right who made this decision, and this will become a key artefact as part of that. And there'll be looking around "ok right, who were the decision makers; who were the people involved?" …But if something does go wrong, it'll be absolutely used, as part of the review of things. And I don't think there's any wording you can put at the front of it, that would stop an organisation doing that. And it would be viewed as an organisational investment* ***(Xiao, business analyst, various social services).***

As a result, there is a real risk that any summary of user responses to the Toolkit may become determinative of outcomes in any future investigation or review of decision-making following incidents of harm. Since the Toolkit is not designed to be a comprehensive and validated risk assessment Toolkit, participants recognised that any summary should not give the impression that it was designed to 'green light' projects. Nonetheless, several participants felt a summary was critical to ensure end users had clear direction on focus areas. Otherwise, there would be a risk that after completing the lengthy Toolkit they could be left overwhelmed by the amount of qualitative information and unable to determine a co-ordinated plan of action. This may reduce engagement and utility of the Toolkit. Participants firmly believed that any summary should not compromise the reflective purpose of the Toolkit and should prompt and guide action.

Some participants wanted a final score based on Likert scale responses divided across the four organisational domains. However, several participants felt that a final pass or fail scoring system based on end users' responses to the Likert scale would be meaningless and unhelpful. Tallying up scores with a cut-off could potentially be harmful as it may inappropriately 'green light' a project without adequate reflection. Even if a score was included based on user responses to the Likert scale ('To what extent...'), the issue would remain that a 'Not at all' response to one question in a particular context may be insignificant but may be critical in another. The participants' concerns reflect observations made by Brown et al (2021) in relation to ethical assessments as opposed to risk management tools: "performing poorly on almost any metric that is important for the interests the user of [the assessment Toolkit] finds relevant [may] … be reason to reject or protest the use of such an algorithm". As Natasha observed, adding up on overall score risked overriding this nuanced analysis.

Participants recognised that ultimately the end user was best placed to determine which issues were most salient in their context. They felt the Likert scale was helpful in so far as it gave end users a sense of their priority areas and give them a sense of where they need to focus their attention. Toolkit users would be more interested in what areas they need to focus on improving meaning a framework for identifying areas of improvement would be more useful:

> *I would say, each area is of importance … that total score is no good at all, doesn't give you anything … I wouldn't even give a score for them. I would have areas for development at the end of … each one. Because this is something that you take seriously, or you don't take seriously. There's no kind of like, semi-trauma informed... a score of 57 out of 100 [for*

> *example], doesn't help people, does it?* **(Craig, social worker, child safety, youth and women services)**

As Craig recognised, this would be consistent with the purpose of the Toolkit being "a call to action". Sebastian, a social security researcher, agreed that "you don"t want people to just be passing, you want them to be proactive". Ultimately, participants felt that returning to the trauma informed principles in the summary would provide a useful framework to structure an analysis identifying areas of improvement, in keeping with the trauma informed approach of the Toolkit.

We adopted these suggestions by including an extra column in the questionnaire portion of the Toolkit to allow participants to identify their own priorities as they progressed through the Toolkit. Instead of providing an overall score based on the Likert scale, we designed the Toolkit such that end users would be able to tally up their priority areas in a matrix of socio-technical domains and trauma informed principles to give them a sense of where they needed to focus their attention. As some participants suggested, future extensions of the Toolkit could benefit from an auto-generated colour wheel based representation of the priority areas across each trauma informed principle and organisational domain.

As a result, it was important to develop a framework within the Toolkit to guide service providers in assessing the risk of harm based on their responses. Several participants suggested using a 'colour coding' or 'traffic light' approach to set the risk level in questions. However, other participants were concerned with pre-allocating risk levels for the same reasons that an overall score may be meaningless, that is, it overlooked the fact that what was risky in one context may not be so in another. Several participants suggested that a traditional risk assessment system was required to assist end users in triaging their responses and identifying their priorities. Consistent with this suggestion, we added guidance on undertaking a traditional risk analysis based on harms that may occur if an issue remains unaddressed by reference to the severity, scale, frequency, and likelihood of that harm.

Although participants understood this was not a risk management Toolkit in the traditional sense in that it suggests specific actions to mitigate risk, some participants still felt some guidance on next steps would be beneficial. Craig suggested including "a chart for a course of action". Particularly, some participants felt that there was a need to emphasise the importance of returning to this Toolkit as a live document throughout the development and use lifecycle of the algorithm supported service. Accordingly, a chart was added at the end of the toolkit under the heading 'Next Steps' providing general guidance to end users.

Improving Toolkit useability and engagement

The participants recognised usability as a key area which could hamper effective use of the Toolkit, but if considered carefully, could vastly improve the quality, engagement, and adoption of the Toolkit. Craig recognised the importance of useability to engagement:

> *It's about the engagement that determines everything. And so, if people are not engaged then you have nothing. So, I think that's about the how can you make that engaging, useful, valuable such that people see the benefit and want to engage with the tool. How good it is, in the sense of it's evidenced by the quality of the engagement, will determine its use. So easy and user friendly is what will benefit and provide all those things* **(Craig, social worker, child safety, youth and women services).**

One of the key useability features identified was the trade-off between length and comprehensiveness of the Toolkit. The challenge was to ensure the Toolkit was broad enough so that it is applicable across diverse organisations contexts and covers salient issues, but not so broad that it becomes unusable. Paula felt that length was not necessarily an issue as larger organisations are used to being audited, but a bigger issue was repetition as that would reduce engagement. Some participants saw the value of a detailed and comprehensive Toolkit despite the time and resource constrains of the social service sector, especially for smaller organisations. They felt that due to the importance of the Toolkit, it should require dedicated time to be completed and the comprehensiveness of the Toolkit would be a signal to the end user that it should be taken seriously. To mitigate the risk of end users being overwhelmed by

the length of the Toolkit, we adopted a participant's suggestion to add clear instructions that the toolkit should be considered within the resourcing constraints of each organisation.

Max was concerned that attempting to be exhaustive in anticipating all possible issues may arise means the end user will not consider any issues that do not appear on the face of the Toolkit. The participant felt that this may discourage the end user from engaging in their own reflective process beyond what appears on the Toolkit.

A key suggestion to make the Toolkit more useable to smaller, resource constrained organisations was to highlight a shortlist of critical questions intended to cover fundamental issues. The suggestion was that even if users were unable to complete the whole Toolkit, addressing these key issues would provide some basic guidance to reduce the risk of harm. To this end, we created 'Section 1: Critical Considerations' in the Toolkit to include crucial questions to consider should organisations not be able to complete the full raft of questions.

Patricia suggested presenting the Toolkit visually to make it more user-friendly including creating a video introduction and online version of the Toolkit so that multiple people could contribute simultaneously. While beyond the scope of this project, it is anticipated that this can form a focus for future development of the Toolkit.

Adopting suggestions by several participants, we re-framed the Toolkit to make it clear that it can (and where possible, should) be completed by multi-professional teams and various stakeholders, including subject matter experts, service users and technology experts. For small community organisations where the technology is often procured from external third parties, it was important that the third party be involved in completing the Toolkit.

A key feature of ensuring useability to a multi-professional team would be the use of language. The questions were to be framed in direct, clear, non-academic, and non-technical language that could be understood by users from diverse backgrounds. Based on these observations, the Toolkit was substantially revised to adjust questions for clarity and readability. Sebastian suggested assigning which member of the team would be responsible for completing each question. However, considering that the Toolkit is intended to be used by organisations and teams of various compositions and sizes, it was decided that this suggestion would not be adopted. Organisations will be able to allocate questions within the team depending on their expertise when adapting the tool to fit their context.

Participants were supportive of how Section 2 was divided into domains based on four aspects of the algorithm supported service. They believed that the combination of the four domains along with the trauma informed principles adjacent to the questions provided an effective structure. They did not feel there was a need to group questions by the principles. Recognising that there is considerable overlap within the principles and that each question may relate to multiple principles, participants believed grouping the questions by principles may be artificial and inaccurate. Xiao felt combining questions related to different principles in each sub-domain may prompt deeper reflection by requiring end users to think dynamically.

Some participants felt that dividing the Toolkit by types of algorithms systems in Section 3 may lead to an extra layer of complexity, particularly where an algorithmic system may not be clearly categorizable if it includes components of multiple systems. Part of the concern was that organisations may be completing the Toolkit before they have decided what algorithmic system to use. As a solution, Leonie suggested adding a 'blue sky' section at the start of the toolkit whereby organisations could frame the problem they are trying to solve by introducing the algorithmic system and reflect on whether they need to use such a system at all. Relatedly, there was general agreement amongst participants that it was critical that the toolkit include a section at the start prompting reflection on whether an algorithmic system was necessary at all. We added a 'blue sky' section in the Introduction of the toolkit to enable end users' to reflect on why, how and whether they should use an algorithmic system at all in the context of their own organisation and service.

Stage 4: Case Studies

Following revisions of the Toolkit arising from the co-design workshops, we piloted the Toolkit on three case studies. We also relied on these three case studies of algorithm supported services in our workshops, to illustrate how to apply the Toolkit and to test the useability and coherence of the Toolkit. The three case studies selected were Robodebt, Allegheny County Family Screening Tool and Tessa. We selected these case studies based on distinct criteria: regional diversity (UK, US and Australia); type of algorithmic system (risk assessment, rule-based system and chatbot); and sector of deployment (welfare, child protection and mental health). All case studies had been criticised widely for causing harm to service users but for distinct underlying reasons. The areas of service selected are all where the use of algorithmic systems in service delivery is particularly prevalent. All three case studies had been subject to media reporting and academic discussions meaning adequate information was accessible to apply our Toolkit. The case studies were applied based solely on publicly available information as we did not have access to any internal or proprietary documents due to project constraints, and confidentiality and intellectual property protections.

Based on our application of the case studies we made the following alterations to the Toolkit:

- In the 'Knowing Your Organisation, Service and the Algorithmic System', we omitted the question "How do the objectives of your service reflect your organisation's values and mission statement?". In completing the Toolkit, we realised that the focus of the assessment was not on whether the service itself was aligned to the broader organisational values. Rather, it was whether introducing the algorithmic system into the service would influence how aligned the 'algorithm supported service' was to the organisation's values. The self-reflection process was focused on the impacts of the algorithmic system. The question was repetitive and redundant, and consistent with Workshop findings it was removed to improve useability and engagement.
- The Toolkit was difficult to apply without detailed insider knowledge and insights into the operations of the algorithmic system and the algorithm supported service. While this observation may appear axiomatic, it was very relevant as it implies that our application of the case studies was necessarily limited as we were solely able to reference publicly available documents. It also meant that we had to include guidance in the Next Steps section to account for this such that engagement would not be limited for a lack of sufficient knowledge or information about the algorithm supported service: "**Identify** any gaps in your knowledge and obtain any further information you may require to complete your evaluation". This observation further re-enforced the importance of multi-professional teams being involved in completing the Toolkit which was a clear finding within the literature (Vetter et al., 2023) as well as our Workshops.
- Applying the toolkit to the case studies, it became clear that for some questions, the preferred response depended on the context. For example, the question "are outputs consistent given similar inputs" was designed to test constancy and predictability, which is important for a trauma informed approach. However, in the context of Robodebt, where the algorithm made systematic and consistent errors due to incorrect assumptions about individuals' yearly incomes, it was clear that answering affirmative to this question was not preferred. In response, at the end of the tool, rather than provide a score calculated by the sum of positive or negative responses, Toolkit users are asked to select areas of greatest priority. The areas identified as the greatest priority are then collated and used to guide Toolkit users' next steps.

Robodebt

In 2015, the Australia government Department of Human Services (DHS) adopted a fully automated system for recovering debts from recipients of social welfare as part of a scheme commonly known as 'Robodebt'. The scheme implemented to follow up welfare recipients who had inaccurately reported their income whilst on welfare. In doing so, such recipients had supposedly received welfare payments to which they were not entitled. Robodebt was a catastrophic failure and led to widespread emotional trauma amongst service users. The recently concluded Royal Commission into Robodebt documents some of the harms that arose from this scheme. Access media reporting of Robodebt here and here.

Introducing an Algorithmic System

Prior to commencing the assessment, please complete the following table which prompts reflection on how and why an algorithmic system could help improve your service.

What is the problem you are seeking to solve within your service by introducing an algorithmic system?	Is it necessary to introduce an algorithmic system to solve the problem? If so, why?	Are there alternative ways of solving the problem? If so, how do these alternatives relate to principles of trauma informed care?	Identify the key stakeholders for the algorithm supported service (inc. service users, users of the algorithmic system, other relevant members of the service provider organisation etc.)
Launched as part of the WPIT to raise and recover debts resulting from overpayment of social security benefits. It was part of an "integrated package of compliance and process improvement initiative including improved automation and targeted strategies for fraud prevention in areas of high risk" in an attempt to make the social security system more fiscally sustainable and repair the budget.[1]	No, it is not necessary.	Yes	Australian public, service users and their families, Centrelink workers, Australian public servants working in the department.

[1] Commonwealth of Australia, *Royal Commission into the Robodebt Scheme* (No 1, 2023) 116 <https://robodebt.royalcommission.gov.au/system/files/2023-09/rrc-accessible-full-report.PDF>.

Knowing the Algorithmic System

Prior to commencing the assessment, please complete the following table based on the above taxonomy with information about the algorithmic system which you intend to use or are currently using as part of the algorithm supported service. The algorithmic system may have more than one function or may not fit into any of the given 'types'. If none of the 'types' given apply to your algorithmic system then you are only required to complete the general section. Otherwise, please complete all the sections corresponding to the type of system being used.

Type	What is the purpose of the algorithmic system?	How does this purpose align with the mission statement and values of your organisation and the objectives of the service?	Describe the algorithmic system and the role it plays within the algorithm supported service
☐ Chatbot ☐ Recommender ☐ Identification and Recognition ☐ Risk Assessment ☒ Detection ☐ Goal driven optimisation ☐ Other	*To detect overpayments of social security benefit and fraud. To regain money owed to the government.* *Previously social security benefit compliance reviews were a "staff intensive verification process involving obtaining information from customers and third parties."4".*[2] *Robodebt (Online Compliance Intervention program) aimed to simplify the process in a way that would prevent "wasteful expenditure"* [3] *and "would more than*	*It is a foundational requirement that government policies follow Australian law. Robodebt has been shown to not comply with Australian law. The debt collection mechanism was the subject of a class action that was settled out of court, and the Royal commission into Robodebt revealed that key public servants in DSS were aware that the calculation method did not comply with Australian social security law and communicated this to the Minster:* *DSS Public Law Branch had confirmed that the smoothing*	*PAYG clean-up (early Robodept proposal) aimed to:* *"a) ... introduce a digital approach to interventions with customers when historical information from the ATO indicates the customer may have incorrectly declared income from employment.* *b) Interventions will be undertaken in a digital environment using the myGov portal. The customer will be presented, via their online account, with the information obtained from the ATO and an*

[2] Ibid 57.
[3] Ibid 61.

	offset the cost" [4] of implementing the system.	method proposed did "not accord with social security legislation," which specified that employment income was assessed fortnightly. It followed that the debt amount calculated would not be supported by law. Additional issues that do not align with the expectations of Australian public law were also revealed in the Royal Commission, including: the dubious ability of an algorithm to accurately identify incorrectly declared income, "problematic" review rights, a lack of clarity on how much detail would be provided to recipients as to how any purported debt had been calculated, and an effective "reversal of onus" onto a recipient to provide information	assessment of their correct welfare entitlement based on this information. The assessment will use an income smoothing methodology to apportion the customer's income over the time of employment (rather than the current cumbersome process whereby the department has to determine and apply income on a fortnightly basis). The customer will have an opportunity to update the information prior to it being applied to their Centrelink record. c) The proposal removes the need for the department to be dependent on customer and business information as the default and instead relies on the use of data already collected by the ATO as the default unless customers want to, and are able to, provide information that varies the outcome. The digital process will enable the department to

[4] Ibid 57.

		to enable accurate calculation of a debt.[5]	undertake a much greater number of compliance reviews. d) The proposal will provide for a four year measure to undertake 866,857 interventions for customers at risk of undeclared or under declared income from employment. It is anticipated that this would result in an estimated $1.2 billion gross savings and debt due to returned outlays". [6]

[5] Ibid 61.
[6] Ibid.

2.2 Human Algorithm Interaction

#	For the algorithm supported service, to what extent …	Not at all	To some extent	To a great extent	Evidence	Action/Plan	Priority	Principle
21.	can it respond appropriately to service users experiencing distress?	☒	☐	☐	*There was no opportunity for service users to easily appeal or question the decision, let alone ask for additional support if the notice caused distress.*	*Provide supports to service users who receive notifications, including referral pathways for appeal, assistance with appeal, and also distress or hardship caused by the notification.*	X☐	Safety Empowerment and Choice
22.	are outputs consistent given similar inputs?	☒	☐	☐	*While the calculation was consistent, the averaging of yearly income to calculate fortnightly income was unreliable due to the erratic work hours of many casual and part time workers.*	*Cease averaging annual income to estimate fortnightly income.*	X☐	Safety
23.	does it respond consistently to outputs of the algorithmic system? *[Within a defined*	☐	☐	☒	*The automatic requesting of additional documents should the service user believe the*	*Cease reversal of onus of proof! It should not be the*	X☐	Safety Trust and Transparency

	framework rather than in an ad-hoc and arbitrary manner]				*calculation was inaccurate was a consistent outcome, but not appropriate.*	*service users responsibility to prove that they do not owe a debt.*		
24.	can the service user choose how the algorithmic system responds to their reports of distress?	☐	☐	☐			☐	Empowerment and Choice
25.	could the algorithmic system displace or limit human judgement and discretion in service provision?	☐	☐	☐			☐	Trust and Transparency
26.	can service users customise their interaction?	☐	☐	☐			☐	Empowerment and Choice

Allegheny County Family Screening Tool

In August 2016, the Allegheny County Department of Human Services (DHS) in Pennsylvania, US, implemented the Allegheny Family Screening Tool (AFST), a predictive risk modelling tool designed to improve child welfare call screening decisions. It was developed to give more standardized information to call screening case workers to assist in determining whether a call should be investigated or not. In her seminal text Automating Inequality author Virginia Eubanks has famously criticised the AFST for disproportionately targeting families experiencing poverty.

Introducing an Algorithmic System

Prior to commencing the assessment, please complete the following table which prompts reflection on how and why an algorithmic system could help improve your service.

What is the problem you are seeking to solve within your service by introducing an algorithmic system?	Is it necessary to introduce an algorithmic system to solve the problem? If so, why?	Are there alternative ways of solving the problem? If so, how do these alternatives relate to principles of trauma informed care?	Identify the key stakeholders for the algorithm supported service (inc. service users, users of the algorithmic system, other relevant members of the service provider organisation etc.)
The Allegheny Family Screening Tool (AFST) was developed to introduce a more data-driven decision making in the screening process in order to increase accuracy and consistency in decisions to investigate[1]	*The County already has a large dataset that Screeners typically access. Having an algorithm to provide a summary risk score, is not necessary, but can enhance screeners' time and understanding of the data.*		*Allegheny County Department of Human Services (DHS) administrators, call centre screening staff, their supervisors, and policy officers and professional staff in the Office of Children, Youth and Families. External stakeholders include service users, community service providers, advocacy groups, foundations, legal profession including judges.*

[1] https://www.alleghenycounty.us/Human-Services/News-Events/Accomplishments/Allegheny-Family-Screening-Tool.aspx#:~:text=Predictive%20Risk%20Modeling%20in%20Child,goal%20of%20improving%20child%20safety

only required to complete the general section. Otherwise, please complete all the sections corresponding to the type of system being used.

Type	What is the purpose of the algorithmic system?	How does this purpose align with the mission statement and values of your organisation and the objectives of the service?	Describe the algorithmic system and the role it plays within the algorithm supported service
☐ Chatbot ☐ Recommender ☐ Identification and Recognition ☒ Risk Assessment ☐ Detection ☐ Goal driven optimisation ☐ Other	*After receiving a notification of a possible case of child abuse/neglect, the AFST calculates the risk of a notified child being:* • *removed from home within 2 years of the referral; or* • *re-referred within 2 months*[2]	*Identifying children at risk of harm is a key priority of the Agency.* *[It is noted that there are some criticisms from screeners that the model predicts something that is not always the issue – in that many children go into foster can because of poverty or non-working parent-child relationships*[3]*]*	*The AFST can be used by Screeners after they receive a notification of possible (risk of) child abuse/harm.* "The survey administered to call screeners two months following implementation of the tool found that over 40% of the call screeners use the tool to inform their recommendation on a consistent basis, although a little less than a third (31%) reported they rarely use it, if at all."[4]

[2] Vaithianathan, R., Putnam-Hornstein, E., Jiang, N., Nand, P., & Maloney, T. (2017). Developing predictive models to support child maltreatment hotline screening decisions: Allegheny County methodology and implementation. *Center for Social data Analytics.*
[3] Cheng, H. F., Stapleton, L., Kawakami, A., Sivaraman, V., Cheng, Y., Qing, D., ... & Zhu, H. (2022, April). How child welfare workers reduce racial disparities in algorithmic decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-22).
[4] Hornby Zeller Associates, Inc (2018) *Allegheny County Predictive Risk Modeling Tool Implementation: Process Evaluation, page 17*

1. CRITICAL CONSIDERATIONS

The following section prompts you to reflect on critical considerations that are applicable to all algorithm supported services. You should ensure that you have thoroughly reflected and engaged with the below considerations. They represent the absolute minimum when considering trauma informed practice within an algorithm supported service.

#	For the algorithm supported service, to what extent…	Not at all	To some extent	To a great extent	Evidence	Action/Plan	Priority	Principle
1.	are service users aware that an algorithmic system is being used?	☐	☐	☐	*The evaluation of AFST involved consultation with "Representatives from community service providers, advocacy groups, foundations, and a family court judge"*[5]		☐	Trust and Transparency
2.	have designers considered if it could cause or trigger trauma?	☒	☐	☐	*The design focus has been on creating an accurate predictive tool using the most sophisticated data analytical approaches to support human decision makers·*		☒	Safety
3.	has the psychological, emotion and cultural safety of service users been considered?	☐	☐	☐			☐	Safety Intersectionality
4.	does the service user have an informed choice about when their	☒	☐	☐	*AFST automatically draws case information from the organizational and County database that have been*		☒	Trust and Transparency Empowerment and Choice

[5] Hornby Zeller Associates, Inc (2018), page 3

	personal information is accessed or shared?				*collected through other avenues*[6]			
5.	can the service user choose to interact with a human?	☒	☐	☐	*AFST is used internally only by Hotline officers to support risk child harm assessment before investigation and action*		☐	Empowerment and Choice Collaboration
6.	can service users regularly provide feedback?	☐	☐	☐			☐	Collaboration Choice
7.	can service users choose to make a complaint, appeal or review directly to a human?	☐	☐	☐			☐	Trust and Transparency Empowerment and Choice
8.	could it unfairly disadvantage different service user groups?	☐	☐	☐			☐	Intersectionality Safety
9.	could the algorithmic system harmfully discriminate against any service user?	☐	☐	☐			☐	Intersectionality Safety
10.	are there supportive processes for responding to harm or distress caused by the algorithmic system?	☐	☐	☐			☐	Safety Trust and Transparency

[6] Vaithianathan, R. et al. (2017). p. 31.

3.4 Risk Assessment

Using past and existing behaviours to predict future outcomes (e.g. predictive risk algorithms).

#	For the risk assessment system, to what extent …	Not at all	To some extent	To a great extent	Evidence	Action/Plan	Priority	Principle
81.	do service providers understand how an assessment is made?	☐	☒	☐	*Call centre staff were trained in the new tool, though it is unclear what they learnt about how the tool worked, the assumptions it made, etc.*[7]		☒	Trust and Transparency
82.	are clear explanations given to service users for why a decision was made?	☒	☐	☐	*No. The tool provides a score, without an explanation.*		☒	Trust and Transparency
83.	could the communication of a decision cause or trigger trauma?	☐	☒	☐	*Yes. A parent being advised that their child is of high risk of abuse/neglect by a computer may feel they have no ability to challenge the assessment.*		☒	Safety Trust and Transparency
84.	has it been assessed for potential harmful bias or discrimination?	☐	☒	☐	*AFST has been repeatedly assessed for racial bias and on outcomes by race. The tool's algorithm showed no discernible difference when trained with or without racial status. Since AFST deployment*		☐	Intersectionality Safety

[7] Hornby Zeller Associates, Inc (2018)

					racial disparities in service responses have decreased, though this appears to be due to human involvement. No other forms of discrimination or bias are know to have been assessed.[8]			
85.	are there accessible and effective appeal processes?	☐	☐	☐			☐	Trust and Transparency Empowerment and Choice
86.	could service users who have experienced trauma be disadvantaged or harmed by the risk assessment process?	☐	☐	☐			☐	Safety
87.	has the service provider considered the impact of errors?	☐	☐	☒	The APST has undergone more peer reviewed research and testing, including ethical evaluation than comparative AI/ADM.[9]		☐	Safety Trust and Transparency
88.	has the service organisation considered the appropriateness of making decisions based on risk?	☐	☐	☐			☐	Safety

[8] Vaithianathan, R., et. al. (2017). Gerchick, M., Jegede, T., Shah, T., Gutierrez, A., Beiers, S., Shemtov, N., ... & Horowitz, A. (2023, June). The Devil is in the Details: Interrogating Values Embedded in the Allegheny Family Screening Tool. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1292-1310). Cheng, H. F., Stapleton, L., Kawakami, A., Sivaraman, V., Cheng, Y., Qing, D., ... & Zhu, H. (2022, April). How child welfare workers reduce racial disparities in algorithmic decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-22).

[9] Dare, T., & Gambrill, E. (2017). Ethical analysis: Predictive risk models at call screening for Allegheny County. *Alleghany County Analytics*. Oh, S. (2020). An Ethical Evaluation of the Use of Predictive Risk Models in Health and Human Services: A Case Study of the Allegheny Family Screening Tool. *Journal of Politics & Society*, *31*(1).

Tessa

In 2022, the UK National Eating Disorders Association (NEDA) released a chatbot named 'Tessa' which was available 24/7 to help service users build resilience and self-awareness by introducing coping skills. NEDA decided to shut down their human staffed helpline to let the chatbot function on its own. Tessa was widely criticised by eating disorder experts as providing problematic advice that further promoted the eating disorder. Tessa was ultimately taken down by NEDA. Access links to media reports about Tessa here and here.

Your Organisation and the Service

Prior to commencing the assessment, please complete this section about your organisation.

What are your organisation's values and mission statement?	What are the objectives of your service?
National Eating Disorders organisation supports individuals and families affected by eating disorders, and serves as a catalyst for prevention, cures and access to quality care. NEDA envisions a world without eating disorders.[1]	NEDA's programs and services are designed to help you find the help and support to people affected by an eating disorder and their families.[2]

[1] 'Our Work', *National Eating Disorders Association* (23 November 2016) <https://www.nationaleatingdisorders.org/about-us/our-work>.
[2] Ibid.

Introducing an Algorithmic System

Prior to commencing the assessment, please complete the following table which prompts reflection on how and why an algorithmic system could help improve your service.

What is the problem you are seeking to solve within your service by introducing an algorithmic system?	**Is it necessary to introduce an algorithmic system to solve the problem? If so, why?**	**Are there alternative ways of solving the problem? If so, how do these alternatives relate to principles of trauma informed care?**	**Identify the key stakeholders for the algorithm supported service (inc. service users, users of the algorithmic system, other relevant members of the service provider organisation etc.)**
Prevention of EDs is of the utmost importance given the wide treatment gap that exists once individuals develop eating disorders. A chatbot is needed to help people with eating disorders who do not receive treatment, particularly 18- to 30-year-old female identifying individuals.	While not necessary, one possible solution to reducing delivery costs of eating disorder programs is to program a chatbot. Chatbots hold promise for both eating disorder prevention and mental health in general compared with other digital mental health interventions given the interactivity provided by chatbots that mimics therapeutic conversations. Findings of research study with 700 women of 18-30 years with eating disorders provide support for the use of a fully automated, highly disseminable chatbot-based EDs prevention program in reducing weight/shape concerns, one of the most robust risk factors for onset for an ED, through 6-month follow-up, as well as in reducing overall ED psycho-pathology, at least in the shorter term.[3]	**Eg Human moderated delivery:** Some guidance from a real-life, human supporter or moderator improves outcomes, which is a consistent finding in the literature for digital interventions. However, costs to provide the program and associated human moderation to the large number of people at risk for an eating disorder who might benefit make it unlikely that a human-moderated version can be disseminated widely. This option may improve collaboration and trust and transparency as there would be a human involved who could form relationships with service users. On the other hand, trust and transparency may be undermined as research has shown that chatbots encourage honest disclosure. If any safety issues arise, the human moderator can correct them.[4]	• Service users – 18-30 year old females at risk of developing eating disorders • Service provider – NEDA • Human helpline workers at NEDA who will be laid off • Cass – external private company who will create chatbot • Subject matter experts – eating disorder professionals/medical experts • Participant service users in co-design process

[3] Ellen E Fitzsimmons-Craft et al, 'Effectiveness of a Chatbot for Eating Disorders Prevention: A Randomized Clinical Trial' (2022) 55(3) *International Journal of Eating Disorders* 343 ('Effectiveness of a Chatbot for Eating Disorders Prevention').
[4] Ibid.

Knowing the Algorithmic System

Prior to commencing the assessment, please complete the following table based on the above taxonomy with information about the algorithmic system which you intend to use or are currently using as part of the algorithm supported service. The algorithmic system may have more than one function or may not fit into any of the given 'types'. If none of the 'types' given apply to your algorithmic system then you are only required to complete the general section. Otherwise, please complete all the sections corresponding to the type of system being used.

Type	What is the purpose of the algorithmic system?	How does this purpose align with the mission statement and values of your organisation and the objectives of the service?	Describe the algorithmic system and the role it plays within the algorithm supported service
☒ Chatbot ☐ Recommender ☐ Identification and Recognition ☐ Risk Assessment ☐ Detection ☐ Goal driven optimisation ☐ Other	Purpose of introducing a chatbot is to reduce the costs of providing eating disorder services to an underserviced group, namely 18 to 30 year old females with eating disorders.	The chatbot improves NEDA's ability to provide access to care and support to individuals suffering from eating disorders as it allows them to delivery an evidence based and effective program to a wide group of service users than they would otherwise be able to. However, it may undermine NEDA's mission of providing "quality" care if the chatbot does not provide appropriate advice or is not appropriately designed. Not having a human practitioner may also reduce the quality of care.	Chatbots are computer programs that can provide information and simulate human conversations. One widely studied targeted prevention program, StudentBodies, an Internet-based program based on cognitive-behavioural therapy delivered over 8 weeks, significantly reduces weight/shape concerns among women at high risk for the onset of an eating disorders, and in the highest risk groups, has been shown to reduce eating disorder onset. Tessa is designed to guide service users through the Student Bodies Program.

2.1 Service Processes, Procedures and Plans

#	For the algorithm supported service, to what extent …	Not at all	To some extent	To a great extent	Evidence	Action/Plan	Priority	Principle
11.	could it harm relationships between service providers and users?	☐	☐	☒	Tessa was used as a replacement for a human staffed helpline by NEDA and the human staff were laid off. In this way, there is harm to human relationships. There may be some benefits to anonymity and interaction with an algorithmic system, such as making users' feel more comfortable speaking openly and disclosing sensitive information. However, there are risks to removing human support for people experiencing disordered eating. Human service providers can provide interpersonal engagement, empathy, and connection. They may understand nuances in experiences and tones in communications. When disclosures are made honestly there may be concerns about whether human support would be required at that point.	Test the viability and feasibility of a human moderated chatbot. Undertake further research into how the benefit of human relationships can be maintained while achieving the cost and efficiency gains of using Tessa. Engage with external experts and service users in undertaking this research. Pending resolution of research create capacity for human involvement with service users in Tessa's design and operations.	☒	Collaboration Trust & Transparency Safety

2.1 Service Processes, Procedures and Plans

#	For the algorithm supported service, to what extent …	Not at all	To some extent	To a great extent	Evidence	Action/Plan	Priority	Principle
13.	are service users informed of significant changes?	☒	☐	☐	Tessa was created by an external private company, Cass, and bought off the shelf by NEDA. NEDA was not aware that Tessa, which was originally built as a rule based, pre-programmed chatbot had been upgraded to a generative AI chatbot by Cass as part of a systems upgrade. Service users were not aware of this update. They were unaware that Tessa was now generating its own responses.[5]	**Eg** Provide correspondence to service users apologising for oversight and informing them of this update. Take Tessa offline and provide appropriate alternative services. Review decision making processes to identify source of error and implement accountability mechanisms.	☒	Safety Trust and Transparency

[5] Kate Wells, 'An Eating Disorders Chatbot Offered Dieting Advice, Raising Fears about AI in Health', *NPR* (online, 9 June 2023) <https://www.npr.org/sections/health-shots/2023/06/08/1180838096/an-eating-disorders-chatbot-offered-dieting-advice-raising-fears-about-ai-in-hea>.

3. Types of Algorithmic Systems

3.1 Chatbots

Interpreting and creating content to power conversational and other interactions between machines and humans (possibly involving multiple media such as voice, text and images).

#	For the chatbot, to what extent …	Not at all	To some extent	To a great extent	Evidence	Action/Plan	Priority	Principle
50.	has it been trained in appropriate communication methods?	☒	☐	☐	Tessa gave advice on how to lose weight healthily and sustainably to someone who told the bot they had an eating disorder and had gained weight. Experts said that any focus on intentional weight loss or diet culture would exacerbate the eating disorder. Positive responses could reinforce harmful behaviours. For example, the chatbot prompted, "Please share with me a few things that make you feel good about yourself. For example, your humour, grace, personality, family, friends, achievements and more!" The user replied, "I hate my appearance, my personality sucks, my family does not like me, and I don't have any friends or achievements." The chatbot responded, "Keep on recognizing your great qualities!"[5] The 160-character limit and use of infographics to reduce text may limit expression required at a vulnerable time for service users experiencing eating disorders.	Consult independent eating disorder experts around appropriate communication methods. Test feasibility of implementing methods with machine learning/technical experts. Pilot the chatbot to test whether methods are effective including with participants in a controlled study.	☒	Safety Trust and Transparency

[5] William W Chan et al, 'The Challenges in Designing a Prevention Chatbot for Eating Disorders: Observational Study' (2022) 6(1) *JMIR Formative Research* e28003 ('The Challenges in Designing a Prevention Chatbot for Eating Disorders').

Future Directions

Where this project has built a co-designed Toolkit to assist organisations to think through, document, and review algorithm supported services, future research will require considerable pilot testing of the Toolkit to ensure that it is user friendly, has adopted appropriate language, and can be flexibly applied across a range of social service contexts. Our future directions aim to involve working with specific partner organisations to apply the Toolkit in real live case studies, and translate our internationally innovative draft Toolkit into an online, dynamic, and accessible resource for all organisations who are involved in the delivery of social services.

A key priority is to **pilot our Toolkit within organisations on live cases** – in use and in development. The Toolkit needs piloting in a diversity of service domains involving a diversity of service users and types of services. This will ensure that the Toolkit be meaningful, useable and robust for various settings and contexts.

Creating **an online, dynamic Toolkit** offers a step-change in its useability. In contrast to a static document, hosting the Toolkit online enables it to be dynamic to enable:

- users to click on key concepts and terms whilst completing the Toolkit to enhance understanding and capacity building;
- automatic filtering of the Toolkit's questions based on responses on stage of deployment so only applicable questions are presented to the Toolkit's users;
- a summary of responses to be automatically produced and tabulated against domains and principles; and
- suggestions for further development and resources to be curated based on the overall responses to the Toolkit.

Identifying, curating and, developing **a suite of educational resources** for organisations in helping to better understand the intersection of trauma and AI and automation is a high priority. While some professionals and organisations are well versed in trauma informed practice, they are less well versed in how AI and automation might impact on trauma informed practice. Similarly, many professionals and organisations will have a good understanding of new digital technologies, but less versed in how they may be poorly designed for people who have experienced or experiencing trauma, or how to ensure that the technologies that they develop or deploy will not cause or retrigger trauma.

In addition to educational resources, the Toolkit would benefit from developing **Training and User Support**. This can include a series of online videos including 'how to use' and 'reflecting on your results'. There is also benefit in exploring Train-the-Trainer and training programmes, such as offered by ForHumanity University (https://forhumanity.center/) or how it might link with emerging AI auditing regulations, practices and professions.

Consideration is needed to whether there is a need to **develop specific Toolkits** for particular service domains or service user groups, or whether a dynamic, generic Toolkit works well across that diversity. This can be explored through the piloting of the dynamic Toolkit in live cases.

Throughout all the next steps, a co-design approach working closely with diverse stakeholders and professionals is essential.

Conclusion

This project is based on recognising that Artificial Intelligence (AI), automated decision making (ADM) and algorithmic systems have the potential to enhance the quality of social service delivery. At the same time, they can cause harm at scale. For example, we have used the example of Robodebt, an

internationally infamous example of ADM that systemically generated trauma, including the loss of life, amongst some of Australia's most vulnerable peoples (Commonwealth of Australia, 2023; Whiteford, 2021). Instead of simply denouncing new technologies, what is needed are practical processes to design, deliver and govern AI-enabled social services that are consistent with trauma-informed social service practices. It is critical that AI is designed and operates in a way that appropriately takes account of the possibility of clients' prior experience of trauma, but also does not generate trauma.

By innovatively drawing together two fields of research and practice – trauma informed approaches and ethical and accountable AI/algorithms – we drafted, co-designed and revised an internationally novel Trauma Informed Algorithmic Assessment Toolkit. Through the four-stage research project, key innovations in our project include:

- moving beyond a focus on digital (or data) harms with its attendant risk assessment methodology, to one that positively framed in being trauma aware;
- mapping key concepts and principles in trauma informed practice to ethical and accountable AI principles;
- providing a practical implementation of ethical AI and trauma informed principles within a digital infused service delivery environment; and
- undertaking a genuine co-design process involving professionals with practical expertise delivering services to people who have experienced trauma.

Reassured by the enthusiasm of our research participants employed in organisations working with people who have experienced trauma, we hold that our Toolkit is a much-needed resource to help identify the manifold strengths and challenges in developing and deploying AI and algorithmic systems to enhance human wellbeing. While much has been discussed and researched about ethical and accountable AI, the work of creating practical processes, procedures and policies to translating principles into practice remains nascent.

In reversing Isaac Asimov's proscriptive framing of the First Laws of Robotics, "a robot may not injure a human being or, through inaction, allow a human being to come to harm" (Asimov, 1950), into a positive framing of what can and should be done by and with AI, our Toolkit contributes to the important and urgent work of building human wellbeing in a world with AI.

References

Ada Lovelace Institute,. (2020). *Examining-the-Black-Box*. Ada Lovelace Institute.

American Psychological Association. (2023). *Trauma*. https://www.apa.org/topics/trauma

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine Bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Asimov, I. (1950). *I, Robot*. Doubeday.

Australian Childhood Foundation. (2019, March 4). What is trauma? *Australian Childhood Foundation Professionals*. https://professionals.childhood.org.au/prosody/2019/03/what-is-trauma/

Baird, S., & Jenkins, S. R. (2003). Vicarious Traumatization, Secondary Traumatic Stress, and Burnout in Sexual Assault and Domestic Violence Agency Staff. *Violence and Victims*, *18*(1), 71–86. https://doi.org/10.1891/vivi.2003.18.1.71

Baybutt, P. (2018). Guidelines for designing risk matrices. *Process Safety Progress*, *37*(1), 49–55. https://doi.org/10.1002/prs.11905

Benuto, L. T., Newlands, R., Ruork, A., Hooft, S., & Ahrendt, A. (2018). Secondary traumatic stress among victim advocates: Prevalence and correlates. *Journal of Evidence-Informed Social Work*, *15*(5), 494–509. https://doi.org/10.1080/23761407.2018.1474825

Blandford, A., Furniss, D., & Makri, S. (2016). *Qualitative HCI research: Going behind the scenes*. Morgan & Claypool Publishers.

Bowen, E. A., & Murshid, N. S. (2016). Trauma-Informed Social Policy: A Conceptual Framework for Policy Analysis and Advocacy. *American Journal of Public Health*, *106*(2), 223–229. https://doi.org/10.2105/AJPH.2015.302970

Branson, D. C. (2019). Vicarious trauma, themes in research, and terminology: A review of literature. *Traumatology*, *25*(1), 2–10. https://doi.org/10.1037/trm0000161

Brayne, S., & Christin, A. (2021). Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts. *Social Problems*, *68*(3), 608–624. https://doi.org/10.1093/socpro/spaa004

Bride, B. E. (2007). Prevalence of Secondary Traumatic Stress among Social Workers. *Social Work*, *52*(1), 63–70. https://doi.org/10.1093/sw/52.1.63

Brown, C., Sanci, L., & Hegarty, K. (2021). Technology-facilitated abuse in relationships: Victimisation patterns and impact in young people. *Computers in Human Behavior*, *124*, 106897. https://doi.org/10.1016/j.chb.2021.106897

Brown, S., Davidovic, J., & Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. *Big Data & Society*, *8*(1), 205395172098386. https://doi.org/10.1177/2053951720983865

Browne, G. (2022, October 1). The Problem With Mental Health Bots. *Wired UK*. https://www.wired.co.uk/article/mental-health-chatbots

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability and Transparency*, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html?mod=article_inline&ref=akusion-ci-shi-dai-bizinesumedeia

Chen, J. X., McDonald, A., Zou, Y., Tseng, E., Roundy, K. A., Tamersoy, A., Schaub, F., Ristenpart, T., & Dell, N. (2022). Trauma-Informed Computing: Towards Safer Technology Experiences for All. *CHI Conference on Human Factors in Computing Systems*, 1–20. https://doi.org/10.1145/3491102.3517475

Christin, A. (2017). Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society*, *4*(2), 205395171771885. https://doi.org/10.1177/2053951717718855

Cobbe, J., Lee, M. S. A., & Singh, J. (2021). Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 598–609. https://doi.org/10.1145/3442188.3445921

Commonwealth of Australia. (2023). *Royal Commission into the Robodebt Scheme* (1). https://robodebt.royalcommission.gov.au/system/files/2023-09/rrc-accessible-full-report.PDF

Constantinas, E., Geiger, G., Braun, J.-C., Mehrotra, D., & Aung, H. (2023, March 6). Inside the Suspicion Machine. *Wired*. https://www.wired.com/story/welfare-state-algorithms/

Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1571–1583. https://doi.org/10.1145/3531146.3533213

Cox, L. A. (Tony). (2008). What's Wrong with Risk Matrices? *Risk Analysis*, *28*(2), 497–512. https://doi.org/10.1111/j.1539-6924.2008.01030.x

De Maio, J. A., Zubrick, S. R., Silburn, S. R., Lawrence, D. M., Mitrou, F. G., Dalby, R. B., Blair, E. M., Griffin, J., Milroy, H., & Cox, A. (2005). *The Western Australian Aboriginal Child Health Survey: Measuring the Social and Emotional Wellbeing of Aboriginal Children and Intergenerational Effects of Forced Separation*. https://www.telethonkids.org.au/globalassets/media/documents/aboriginal-health/measuring_social_and_emotional_wellbeing.pdf

De Vaus, D. (2013). *Surveys in social research*. Routledge.

Desiere, S., Langenbucher, K., & Struyven, L. (2019). *Statistical profiling in public employment services: An international comparison*. OECD. https://doi.org/10.1787/b5e5f16e-en

Douglas, H., & Fitzgerald, R. (2021, September 16). QLD police will use AI to 'predict' domestic violence before it happens. Beware the unintended consequences. *The Conversation*. https://theconversation.com/qld-police-will-use-ai-to-predict-domestic-violence-before-it-happens-beware-the-unintended-consequences-167976

Dubber, M. D., Pasquale, F., & Das, S. (2020). *The Oxford Handbook of Ethics of AI*. Oxford University Press.

Duijm, N. J. (2015). Recommendations on the use and design of risk matrices. *Safety Science*, *76*, 21–31. https://doi.org/10.1016/j.ssci.2015.02.014

Eisenbruch, M. (1991). From Post-Traumatic Stress Disorder to Cultural Bereavement: Diagnosis of Southeast Asian Refugees. *Social Science & Medicine*, *33*, 673–680.

Elliott, D. E., Bjelajac, P., Fallot, R. D., Markoff, L. S., & Reed, B. G. (2005). Trauma-informed or trauma-denied: Principles and implementation of trauma-informed services for women. *Journal of Community Psychology*, *33*(4), 461–477. https://doi.org/10.1002/jcop.20063

Emms, M., Arief, B., & Van Moorsel, A. (2014). Electronic Footprints in the Sand: Technologies for Assisting Domestic Violence Survivors. In B. Preneel & D. Ikonomou (Eds.), *Privacy Technologies and Policy* (Vol. 8319, pp. 203–214). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-54069-1_13

Erikson, K. (1991). Notes on Trauma and Community. *PSYCHOANALYSIS, CULTURE AND TRAUMA*, *48*(4), 455–472.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Publishing Group.

Fallot, R., & Harris, M. (2009). *Creating Cultures of Trauma-Informed Care (CCTIC): A Self-Assessment and Planning Protocol*. Community Connections. https://children.wi.gov/Documents/CCTICSelf-AssessmentandPlanningProtocol0709.pdf

Figley, C. R. (2013). *Compassion Fatigue* (0 ed.). Routledge. https://doi.org/10.4324/9780203777381

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI* (SSRN Scholarly Paper 3518482). https://doi.org/10.2139/ssrn.3518482

Gillingham, P. (2021). Algorithmically Based Decision Support Tools: Skeptical Thinking about the Inclusion of Previous Involvement. *Practice*, *33*(1), 37–50. https://doi.org/10.1080/09503153.2020.1749584

Goodman, L. A., Sullivan, C. M., Serrata, J., Perilla, J., Wilson, J. M., Fauci, J. E., & DiGiovanni, C. D. (2016). DEVELOPMENT AND VALIDATION OF THE TRAUMA-INFORMED PRACTICE SCALES. *Journal of Community Psychology*, *44*(6), 747–764. https://doi.org/10.1002/jcop.21799

Greenwald, R., Maguin, E., Smyth, N. J., Greenwald, H., Johnston, K. G., & Weiss, R. L. (2008). Teaching Trauma-Related Insight Improves Attitudes and Behaviors Toward Challenging Clients. *Traumatology*, *14*(2), 1–11. https://doi.org/10.1177/1534765608315635

Griffin, B. J., Purcell, N., Burkman, K., Litz, B. T., Bryan, C. J., Schmitz, M., Villierme, C., Walsh, J., & Maguen, S. (2019). Moral Injury: An Integrative Review. *Journal of Traumatic Stress*, *32*(3), 350–362. https://doi.org/10.1002/jts.22362

Guarino, K., Soares, P., Konnath, K., Clervil, R., & Bassuk, E. (2009). *Trauma-Informed Organizational Toolkit*. Center for Mental Health Services, Substance Abuse and Mental Health Services Administration, and the Daniels Fund, the National Child Traumatic Stress Network, and the W.K. Kellogg Foundation. www.homeless.samhsa.gov and www.familyhomelessness.org.

Guthrie, G. (2010). *Basic Research Methods: An Entry to Social Science Research*. SAGE Publications India Pvt Ltd. https://doi.org/10.4135/9788132105961

Harris, B., Dragiewicz, M., & Woodlock, D. (2021). Positive uses of technology to address domestic violence. *The Emerald Handbook of Crime, Justice and Sustainable Development*, 295–313.

Harris, M., & Fallot, R. D. (2001). Envisioning a trauma-informed service system: A vital paradigm shift. *New Directions for Mental Health Services*, *2001*(89), 3–22. https://doi.org/10.1002/yd.23320018903

Hegarty, T. (2022, February 14). *No government algorithm should decide who eats and who goes hungry—We step up our challenge to the DWP's secret benefit probes algorithm*. Foxglove. https://www.foxglove.org.uk/2022/02/14/algorithm-dwp-challenge/

Henderson, C., Everett, M., & Isobel, S. (2018). *Trauma-Informed Care and Practice Organisational Toolkit (TICPOT): An Organisational Change Process Resource, Stage 1—Planning and Audit*. Mental Health Coordinating Council (MHCC).

Henman, P. (2020). Improving public services using artificial intelligence: Possibilities, pitfalls, governance. *Asia Pacific Journal of Public Administration*, *42*(4), 209–221. https://doi.org/10.1080/23276665.2020.1816188

Henman, P. W. F. (2022). Digital Social Policy: Past, Present, Future. *Journal of Social Policy*, *51*(3), 535–550. https://doi.org/10.1017/S0047279422000162

Herman, J. L. (1997). *Trauma and recovery* (Rev. ed). BasicBooks.

Hickle, K. (2020). Introducing a trauma-informed capability approach in youth services. *Children & Society*, *34*(6), 537–551. https://doi.org/10.1111/chso.12388

Homes, A., Grandison, G., The Rivers Centre, & NHS Lothian. (2021). *Trauma-Informed Practice: A Toolkit for Scotland*. Scottish Government. https://www.gov.scot/publications/trauma-informed-practice-toolkit-scotland/documents/

Hopkins, A., & Booth, S. (2021). Machine Learning Practices Outside Big Tech: How Resource Constraints Challenge Responsible Development. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 134–145. https://doi.org/10.1145/3461702.3462527

Huang, L. N., Flatow, R., Biggs, T., Afayee, S., Smith, K., Clark, T., & Blake, M. (2014). *SAMHSA's Concept of Trauma and Guidance for a Trauma-Informed Approach* (HHS Publication No. (SMA) 14-4884). Substance Abuse and Mental Health Services Administration (SAMHSA). https://archive.hshsl.umaryland.edu/handle/10713/18559

Hukkelberg, I., & Rolland, K. (2020). Exploring Machine Learning in a Large Governmental Organization: An Information Infrastructure Perspective. *ECIS 2020 Research-inProgress Papers*. Twenty-Eighth European Conference on Information Systems (ECIS2020).

Human Rights and Equal Opportunity Commission. (1997). *Bringing Them Home: Report of the National Inquiry into the Separation of Aboriginal and Torres Strait Islander Children from Their Families*. Human Rights and Equal Opportunity Commission.

James, P., Lal, J., Liao, A., Magee, L., & Soldatic, K. (2023). Algorithmic decision-making in social work practice and pedagogy: Confronting the competency/critique dilemma. *Social Work Education*, 1–18. https://doi.org/10.1080/02615479.2023.2195425

Jasanoff, S., & Kim, S.-H. (2015). *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power*. University of Chicago Press.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), Article 9. https://doi.org/10.1038/s42256-019-0088-2

Kezelman, C. (2014, June 12). *Trauma informed practice*. Mental Health Australia. https://mhaustralia.org/general/trauma-informed-practice

Kezelman, C., & Stavropoulos, P. (2012). *'The last frontier' – practice guidelines for treatment of complex trauma and trauma informed care and service delivery* (Australia) [Report]. Adults Surviving Child Abuse. https://apo.org.au/node/31272

Kroll, J. A. (2021). Outlining Traceability: A Principle for Operationalizing Accountability in Computing Systems. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 758–771. https://doi.org/10.1145/3442188.3445937

Laestadius, L., Bishop, A., Gonzalez, M., Illenčík, D., & Campos-Castillo, C. (2022). Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society*, 14614448221142007. https://doi.org/10.1177/14614448221142007

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review*, *9*(2). https://policyreview.info/concepts/transparency-artificial-intelligence

Li, H., Yu, L., & He, W. (2019). The Impact of GDPR on Global Technology Development. *Journal of Global Information Technology Management*, *22*(1), 1–6. https://doi.org/10.1080/1097198X.2019.1569186

Lisa McCann, I., & Pearlman, L. A. (1990). Vicarious traumatization: A framework for understanding the psychological effects of working with victims. *Journal of Traumatic Stress*, *3*(1), 131–149. https://doi.org/10.1002/jts.2490030110

Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376445

Mahon, D. (2022). *Trauma-Responsive Organisations: The Trauma Ecology Model*. Emerald Publishing Limited. https://www.emerald.com/insight/content/doi/10.1108/978-1-80382-429-120221011/full/html

McCann, L., & Pearlman, L. A. (1990). Vicarious traumatization: A framework for understanding the psychological effects of working with victims. *Journal of Traumatic Stress*, *3*(1), 131–149. https://doi.org/10.1002/jts.2490030110

McLaughlin, H. (2009). What's in a Name: 'Client', 'Patient', 'Customer', 'Consumer', 'Expert by Experience', 'Service User'--What's Next? *British Journal of Social Work*, *39*(6), 1101–1117. https://doi.org/10.1093/bjsw/bcm155

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 115:1-115:35. https://doi.org/10.1145/3457607

Meijer-van Leijsen, E. (2021). Developing an Audit Framework for Algorithms. *International Journal of Government Auditing*, *48*(2), 39–40.

Menschner, C., & Maul, A. (2016). *Key Ingredients for Successful Trauma-Informed Care Implementation*. Center for Health Care Strategies.

Mental Health Coordinating Council. (2013). *Trauma Informed Care and Practice: Towards a cultural shift in policy reform across mental health and human services in Australia. A National Direction. Position Paper and Recommendations.* https://mhcc.org.au/wp-content/uploads/2018/05/ticp_awg_position_paper__v_44_final___07_11_13-1.pdf

Mental Health Foundation UK. (2023). *The impact of traumatic events on mental health*. https://www.mentalhealth.org.uk/explore-mental-health/publications/impact-traumatic-events-mental-health

Menzies, K. (2019). Understanding the Australian Aboriginal experience of collective, historical and intergenerational trauma. *International Social Work*, *62*(6), 1522–1534. https://doi.org/10.1177/0020872819870585

Michalopoulos, L. M., & Aparicio, E. (2012). Vicarious Trauma in Social Workers: The Role of Trauma History, Social Support, and Years of Experience. *Journal of Aggression, Maltreatment & Trauma*, *21*(6), 646–664. https://doi.org/10.1080/10926771.2012.689422

Microsoft. (2023). *Assessing harm: A guide for tech builders*.

Microsoft. (2022, May 6). *Harms modeling—Azure Application Architecture Guide*. https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Misra, K. B. (2008). Risk Analysis and Management: An Introduction. In K. B. Misra (Ed.), *Handbook of Performability Engineering* (pp. 667–681). Springer London. https://doi.org/10.1007/978-1-84800-131-2_41

Mökander, J. (2023). Auditing of AI: Legal, Ethical and Technical Approaches. *Digital Society*, *2*(3), 49. https://doi.org/10.1007/s44206-023-00074-y

Mökander, J., Axente, M., Casolari, F., & Floridi, L. (2022). Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. *Minds and Machines*, *32*(2), 241–268. https://doi.org/10.1007/s11023-021-09577-4

Mökander, J., & Floridi, L. (2021). Ethics-Based Auditing to Develop Trustworthy AI. *Minds and Machines*, *31*(2), 323–327. https://doi.org/10.1007/s11023-021-09557-8

Morgan, D. (1997). *Focus Groups as Qualitative Research*. SAGE Publications, Inc. https://doi.org/10.4135/9781412984287

Nahar, N., Zhou, S., Lewis, G., & Kästner, C. (2022). *Collaboration Challenges in Building ML-Enabled Systems: Communication, Documentation, Engineering, and Process* (arXiv:2110.10234). arXiv. http://arxiv.org/abs/2110.10234

National Trauma Transformation Programme (NTTP). (2023). Implementation. A Roadmap for Creating Trauma-Informed and Responsive Change: Guidance for Organisations, Systems and Workforces in Scotland (2023). *National Trauma Transformation Programme*. https://www.traumatransformation.scot/implementation/

Netherlands Court of Audit. (2021). *Understanding algorithms—2021*.

NSW Government. (2021). *Artificial intelligence assurance framework*. https://www.digital.nsw.gov.au/sites/default/files/2022-09/nsw-government-assurance-framework.pdf

NSW Health. (2022, August 17). *What is trauma-informed care? - Principles for effective support*. https://www.health.nsw.gov.au:443/mentalhealth/psychosocial/principles/Pages/trauma-informed.aspx

NSW Heath. (2023). *Trauma, violence, abuse and neglect statistics—Integrated trauma-informed care*. https://www.health.nsw.gov.au/patients/trauma/Pages/trauma-violence-abuse-neglect.aspx

OECD. (2021). Human rights due diligence through responsible AI. In *OECD Business and Finance Outlook 2021: AI in Business and Finance*. OECD. https://doi.org/10.1787/ba682899-en

OECD. (2023). *Recommendation of the Council on Artificial Intelligence* (OECD/LEGAL/0449).

OECD Framework for the classification of AI systems (323; OECD Digital Economy Papers). (2022). OECD.

Oliva, J. D. (2022). *Dosing Discrimination: Regulating PDMP Risk Scores*. https://doi.org/10.15779/Z38Z31NP8J

Power, M. (1997). *The Audit Society: Rituals of Verification*. OUP. https://www.google.com.au/books/edition/The_Audit_Society_Rituals_of_Verificatio/q4U3AwAAQBAJ?hl=en&gbpv=1&dq=rituals+of+verification&pg=PP1&printsec=frontcover

Power, M. (2021). Modelling the Micro-Foundations of the Audit Society: Organizations and the Logic of the Audit Trail. *Academy of Management Review*, *46*(1), 6–32. https://doi.org/10.5465/amr.2017.0212

Radiya-Dixit, E., & Neff, G. (2023). A Sociotechnical Audit: Assessing Police Use of Facial Recognition. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 1334–1346. https://doi.org/10.1145/3593013.3594084

Raja, S., Hasnain, M., Hoersch, M., Gove-Yin, S., & Rajagopalan, C. (2015). Trauma informed care in medicine. *Family & Community Health*, *38*(3), 216–226.

Rausand, M. (2020). The Words of Risk Analysis. In *Risk Assessment* (pp. 15–57). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119377351.ch2

Rausand, M., & Haugen, S. (2020). Measuring Risk. In *Risk Assessment* (pp. 121–165). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119377351.ch6

Reeves, E. (2015). A Synthesis of the Literature on Trauma-Informed Care. *Issues in Mental Health Nursing*, *36*(9), 698–709. https://doi.org/10.3109/01612840.2015.1025319

Regulation 679/2016. https://doi.org/10.5040/9781782258674

Sacred Heart Mission. (2023). *Trauma-informed care*. Sacred Heart Mission. https://www.sacredheartmission.org/about-us/our-service-model-approach/trauma-informed-care/

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., & Ghani, R. (2019). *Aequitas: A Bias and Fairness Audit Toolkit* (arXiv:1811.05577). arXiv. http://arxiv.org/abs/1811.05577

SAMHSA. (2023). *Practical Guide for Implementing a Trauma-Informed Approach* (PEP23-06-05–005). https://store.samhsa.gov/sites/default/files/pep23-06-05-005.pdf

Saxena, D., Badillo-Urquiola, K., Wisniewski, P. J., & Guha, S. (2021). A Framework of High-Stakes Algorithmic Decision-Making for the Public Sector Developed through a Case Study of Child-Welfare. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–41. https://doi.org/10.1145/3476089

Schwartz, P. M. (2019). Global data privacy: The EU way. *NYUL Rev.*, *94*, 771.

Schwartz, R., Vassilev, A., Greene, K. K., Perine, L., Burt, A., & Hall, P. (2022). Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. *NIST*. https://www.nist.gov/publications/towards-standard-identifying-and-managing-bias-artificial-intelligence

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. https://doi.org/10.1145/3287560.3287598

Shelby, R., Rismani, S., Henne, K., Moon, Aj., Rostamzadeh, N., Nicholas, P., Yilla, N., Gallegos, J., Smart, A., Garcia, E., & Virk, G. (2023). *Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction* (arXiv:2210.05791). arXiv. http://arxiv.org/abs/2210.05791

Shilton, K. (2013). Values Levers: Building Ethics into Design. *Science, Technology, & Human Values*, *38*(3), 374–397. https://doi.org/10.1177/0162243912436985

Sleep, L. (2023). Female dependents, individual customers and promiscuous digital personas: The multiple governing of women through the Australian social security couple rule. *Critical Social Policy*, *43*(2), 193–213. https://doi.org/10.1177/02610183221089265

Steen, M. (2013). Co-design as a process of joint inquiry and imagination. *Design Issues*, *29*(2), 16–28.

Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (100–1; pp. 100–101). National Institute of Standards and Technology (U.S.). https://doi.org/10.6028/NIST.AI.100-1

Teo, S. A. (2022). How Artificial Intelligence Systems Challenge the Conceptual Foundations of the Human Rights Legal Framework. *Nordic Journal of Human Rights*, *40*(1), 216–234. https://doi.org/10.1080/18918131.2022.2073078

Theisen-Womersley, G. (2021). *Trauma and Resilience Among Displaced Populations: A Sociocultural Exploration*. Springer International Publishing. https://doi.org/10.1007/978-3-030-67712-1

Thompson, D. F. (2017). Designing Responsibility: The Problem of Many Hands in Complex Organizations. In J. Van Den Hoven, S. Miller, & T. Pogge (Eds.), *Designing in Ethics* (1st ed., pp. 32–56). Cambridge University Press. https://doi.org/10.1017/9780511844317.003

Tsantefski, M., Humphreys, C., Wilde, T., Young, A., Heward-Belle, S., & O'Leary, P. (2023). Worker Safety in High-risk Child Protection and Domestic Violence Cases. *Journal of Family Violence*. https://doi.org/10.1007/s10896-023-00551-5

Tseris, E. J. (2013). Trauma Theory Without Feminism? Evaluating Contemporary Understandings of Traumatized Women. *Affilia*, *28*(2), 153–164. https://doi.org/10.1177/0886109913485707

Turk, V. (2023, October 10). Generative AI like Midjourney creates images full of stereotypes—Rest of World. *Rest of World*. https://restofworld.org/2023/ai-image-stereotypes/

UNESCO. (2022). *Recommendation on the Ethics of Artificial Intelligence* (SHS/BIO/PI/2021/1; p. 43).

Vaithianathan, R., Putnam-Hornstein, E., Jiang, N., Nand, P., & Maloney, T. (n.d.). *Developing Predictive Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology and Implementation*.

Vetter, D., Amann, J., Bruneault, F., Coffee, M., Düdder, B., Gallucci, A., Gilbert, T. K., Hagendorff, T., Van Halem, I., Hickman, E., Hildt, E., Holm, S., Kararigas, G., Kringen, P., Madai, V. I., Wiinblad Mathez, E., Tithi, J. J., Westerlund, M., Wurth, R., … Z-Inspection® initiative (2022). (2023). Lessons Learned from Assessing Trustworthy AI in Practice. *Digital Society*, *2*(3), 35. https://doi.org/10.1007/s44206-023-00063-1

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3063289

Wells, K. (2023, June 9). An eating disorders chatbot offered dieting advice, raising fears about AI in health. *NPR*. https://www.npr.org/sections/health-shots/2023/06/08/1180838096/an-eating-disorders-chatbot-offered-dieting-advice-raising-fears-about-ai-in-hea

Whiteford, P. (2021). Debt by design: The anatomy of a social policy fiasco – Or was it something worse? *Australian Journal of Public Administration*, *80*(2), 340–360. https://doi.org/10.1111/1467-8500.12479

Wolf, M. R., Green, S. A., Nochajski, T. H., Mendel, W. E., & Kusmaul, N. S. (2014). 'We're Civil Servants': The Status of Trauma-Informed Care in the Community. *Journal of Social Service Research*, *40*(1), 111–120. https://doi.org/10.1080/01488376.2013.845131

Woodlock, D., McKenzie, M., Western, D., & Harris, B. (2020). Technology as a Weapon in Domestic Violence: Responding to Digital Coercive Control. *Australian Social Work*, *73*(3), 368–380. https://doi.org/10.1080/0312407X.2019.1607510

Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation & Governance*, *12*(4), 505–523. https://doi.org/10.1111/rego.12158

Acknowledgements

This material is based on work supported by The University of Notre Dame-IBM Tech Ethics Lab 2022-23 Auditing AI funding (Award # 262812UQ). Such support does not constitute endorsement by the sponsor of views expressed in this publication. Additional support was also provided by the Australian Research Council's Centre of Excellence for Automated Decision Making and Society (CE200100005). We acknowledge the contributions of Philip Gillingham to this project in its initial stages. We thank all our participants and their organisations for their time and invaluable insights into the formation of the project's resulting Trauma Informed Algorithmic Assessment Tool.

Recommended Citation

Maitra, S., Sleep, L., Fay, S., Henman, P. (2024) *Building a Trauma-Informed Algorithmic Assessment Toolkit,* Project Report. Brisbane: ARC Centre of Excellence for Automated Decision-Making and Society, University of Queensland, DOI: 10.60836/f01c-4a18.

Appendix A **Trauma-Informed Algorithmic Assessment Toolkit**

Trauma-Informed Algorithms:

An Organizational Toolkit for Social Services

Acknowledgements

The research is being organised and conducted by researchers from the University of Queensland and Central Queensland University, with funding from the University of Notre Dame-IBM Technology Ethics Lab and has approval from the University Research Ethics Committee (2023/HE000004).

The following researchers were involved in the project:

Professor Paul Henman, ARC Centre of Excellence for Automated Decision Making and Society, School of Social Science, University of Queensland

Dr Suzanna Fay, Senior Lecturer, School of Social Science, University of Queensland

Dr Lyndal Sleep, Senior Lecturer, Queensland Centre for Domestic and Family Violence Research, School of Nursing, Midwifery and Social Sciences, Central Queensland University

Suvradip Maitra, School of Social Science, University of Queensland

We thank all the participants who gave up their time to attend the workshops for their valuable input. This project would not be possible without your engagement, enthusiasm and desire to make a positive difference to your clients.

Contact:

Professor Paul Henman
Email: p.henman@uq.edu.au
Phone: +61 7 3365 2383

GLOSSARY

Term	Definition
Social services	Services provided by public and/or private actors to vulnerable social groups including but not limited to housing support, child and family support, mental health support and welfare services.
Algorithmic system	A machine-based system which gives an output such as generated text, a risk assessment score or any other predefined output based on a defined set of objectives and input data based on either a pre-programmed or machine learning algorithm.
Discretion	Professional judgement employed by professionals and service deliverers to address service users' needs.
Diverse social and cultural needs	Unique needs that exist due to membership of certain demographic groups including people with disabilities, LGBTQIA+ people, people from multicultural backgrounds, Indigenous peoples, elderly people, children and young people.
Informed choice	service users understand enough about the algorithmic system to allow them to make meaningful choices about their engagement with it whilst feeling assured that any decision they make will not affect service quality.
Meaningful engagement	An ethical concept which recognises the expertise in lived experience of service users to foster collaboration between service providers and users in designing and implementing policies. Depending on the stage of the project, meaningful engagement can involve a combination of outreach, consultation, collaboration and shared leadership. Meaningful engagement requires treating participants with integrity and respect, building processes that are responsive to feedback and communicating openly about power dynamics and decision-making processes
Algorithm supported service	A social service which uses an algorithmic system in any capacity to inform, support, enable, execute, or improve service delivery.
Trauma	An event, series of events, or a set of circumstances an individual experiences as physically or emotionally harmful or threatening, which may have lasting adverse effects on the individual's functioning and mental, physical, social, emotional, or spiritual well-being. Trauma may be experienced by an individual, a generation, or an entire community or culture.
Service user	An individual or group of individuals who are in receipt of social services.
Service provider	An individual or organisation which provides social services.
Trauma-informed approach	A program, organization, or system that is trauma-informed realizes the widespread impact of trauma and understands potential paths for recovery; recognizes the signs and symptoms of trauma in clients, families, staff, and others involved with the system; and responds by fully integrating knowledge about trauma into policies, procedures, and practices, and seeks to actively resist re-traumatization

Background

Algorithmic systems, including automated decision-making systems (ADM) and Artificial Intelligence (AI) are being increasingly used in the delivery of social services. Offering opportunities for more efficient, effective and personalised service delivery, automation can also generate greater problems, reinforcing disadvantage, generating trauma or re-traumatising service users. As the recent Royal Commission into the Robodebt Scheme in Australia has highlighted, they can cause considerable harm to service users if left unchecked.

Conducted by a multi-disciplinary research team with extensive expertise in the intersection of social services and digital technology, this project has co-designed a trauma-informed audit toolkit to aid a reflective process on how an algorithmic system may generate new trauma or re-traumatise. It was road-tested using multiple case studies of AI use and workshops with professionals working in child/family services, DFV services, health services, and social security/welfare payments.

The use of algorithmic systems does not detract from the inherent responsibility of service providers to ensure they are able to deliver trauma informed care for service users. However, the use of these systems raise novel and unique concerns which service providers may be unfamiliar with or unable to anticipate. It is hoped that this toolkit will provide service providers with the capacity to be intentional and reflective in their deployment of algorithmic systems in social services.

Using the toolkit

This toolkit has been designed to be used by social service delivery providers to reflect on how their use of algorithmic systems in social service delivery may cause or perpetuate trauma for service users. The toolkit is not designed for use in medical contexts by health professionals and does not provide a manual to diagnose or assess trauma.

We recommend completing the toolkit in a small group including various stakeholders who are knowledgeable about the delivery of the service, knowledgeable about the algorithmic system being deployed, and service user(s). However, if this is not possible, the toolkit can be completed by a single person who can obtain the relevant information as required. You should adapt and complete the toolkit within the resourcing constraints of your respective organisation. It aims to support services to use algorithmic systems in a way that enhances the trauma informed quality of service delivery in both circumstances.

We built the toolkit on the premise that the needs of service users must be centred in any service which uses algorithms. The goal of this toolkit is to guide and inform the reflective process that is so critical to all trauma informed care. The toolkit is **not** designed to 'green light' any projects or proposals, or to be used as a checkbox to say whether the use of an algorithmic system is or is not trauma informed.

The toolkit is also not a risk management toolkit. The goal of the toolkit is not to provide guidance in mitigating or managing risk of causing or perpetuating trauma. Rather, consistent with the goals of trauma informed practice the toolkit seeks to provide procedural prompts and interventions that can make the delivery of the algorithm supported service more trauma informed.

The service provider should treat this toolkit and the issues which emerge because of its completion as a prompt for further investigation, reflection, and action. The toolkit is not the end point in the process of a service providers engaging with how and why they are deploying the algorithmic system. Rather, it is designed to spark critical conversation and aid in deliberations within the service provider organisation on how they can design and/or modify their systems, processes, and decisions to ensure their deployment of the algorithmic system is trauma informed. You should use it to identify priority areas for improvement in the context of your own algorithm supported service.

The assessment can be completed each time a new algorithmic system is deployed or proposed to be deployed as part of a social service. Where an algorithmic system has already been deployed, the toolkit will be useful in assessing and monitoring its current and continued use.

Principles of Trauma Informed Care

The toolkit is underpinned by principles of trauma informed care. We consider a ‘trauma-informed approach’ to service delivery and care to be a strengths-based way of working with individuals across the lifespan, rooted in a foundational understanding of trauma and the impact that experiencing trauma can have in people's lives. A trauma informed approach requires a realisation of how trauma can affect individuals, recognition of signs of trauma and resisting re-traumatisation.

One way to implement a trauma informed approach is through applying the principles of trauma informed care. While we believe the principles are beneficial in framing the assessment, we do not advocate for a formulaic application of the principles to the algorithm supported service. The principles invite deliberate engagement and deep reflection on behalf of the service provider. The service provider is encouraged to ask themselves whether in all the circumstances their deployment of the algorithmic system is trauma informed. The key principles used in this tool are empowerment and choice, collaboration, trust and transparency, safety, and intersectionality. We have adapted these principles based on principles of ethical algorithms to fit the context of an algorithm supported service.

Principle	Description
Empowerment and Choice	The algorithm supported service should seek to utilise service users' existing strengths and abilities, and empower service users by: • facilitating power sharing and returning control to service users through shared decision making; • enabling meaningful choice for service users regarding how, when and from whom services are received; • enabling meaningful choice for service users regarding how and when their personal data is used, processed and stored; and • ensuring any algorithmic systems respect service users' autonomy and augments service users' strengths.
Collaboration	The algorithm supported service should enable collaboration between service providers and service users by: • engaging with service users in planning, design, delivery and evaluation of the algorithm supported service including the design, development and deployment of the algorithmic system; • centring the importance of mutual and collaborative human relationships in healing and recovering from trauma; and • recognising service users' expertise in their own experiences of previous social services or algorithmic systems.
Trust and Transparency	The algorithm supported service should aim to build and maintain trust with service users by: • providing meaningful transparency about how, why and when an algorithmic system is used, the design of the algorithmic system, and how, why and by whom decisions are made; • delivering accurate, reliable, robust and consistent or predictable outcomes including in unfamiliar conditions; and • being resilient and secure against unauthorised or malicious actors and attacks.
Safety	The algorithm supported service should ensure service users' physical, emotional and psychological safety by: • reducing the potential for re-traumatisation such as by reducing the need for disclosures of trauma; • creating a safe physical and digital environment; • promoting safe and welcoming digital and interpersonal interactions such as by responding appropriately to disclosures of trauma; and • respecting service users' privacy and confidentiality, personal space, boundaries and integrity of identity.
Intersectionality	The algorithm supported service should respect and acknowledge the influence of intersecting identities and background of culture, race, gender, sexuality, ability and age in causing and perpetuating trauma, and recovery and healing from trauma, by: • preventing discriminatory impacts including by mitigating bias and moving past harmful stereotypes; • respecting diversity and inclusivity in any collaboration with service users; and • acknowledging the role of these intersecting identities and background in a service users' needs from, experiences of and responses to the algorithm supported service.

Structure of the toolkit

The toolkit facilitates an analysis of a whole **algorithm supported service**. By the phrase “algorithm supported service”, we include not only the algorithmic system but also surrounding processes, including the role of the service user and the interaction between the human and the algorithm. The toolkit does not consider broader organisational policies or procedures which are not connected to the specific algorithm supported service.

The toolkit is structured around a series of statements, to which toolkit users are asked to make an assessment on a 3-point scale. The statements have been designed to prompt users to reflect on the extent to which each element has been considered or addressed, and how your service might implement or enhance trauma-informed algorithmic enabled service.

The toolkit is action-oriented in the sense that it serves as a springboard for actions consistent with trauma informed care. It does not however prescribe or suggest actions recognising the importance of contextual factors which may impact the appropriate action or plan, including the service user group, algorithmic system, social service, social service users and algorithmic system.

At a high level the toolkit is divided into ‘Section 1 - General’ and ‘Section 2 - Types of system’. The toolkit is designed such that Section 1 is applicable to all types of algorithmic systems being assessed. Whereas, Section 2 consists of several sub-sections, one each for a different type of algorithmic system. For example, a service provider deploying a chatbot would complete all of Section 1 and only Section 2.1 which relates to chatbots.

Not all questions in the toolkit will be relevant to your circumstances. For example, if the toolkit is used before an algorithmic system has been deployed then questions about whether ongoing consultation with service users may not be as relevant. Service providers are encouraged to exercise their discretion and engage with why or why not a question may not apply to their situation. The process of completing the toolkit is designed to improve an understanding of the algorithmic enabled service to provide insights into its strengths and limitations.

Section 1 is divided into four sub sections that highlight four aspects of algorithm supported service.

General Sections	Description
Service Procedures, Processes and Plans	This section assesses whether the overall service, including its policies, processes and knowledge systems are well adjusted and suited to the algorithmic system in a way that reduces the risk of causing or perpetuating trauma.
Human Algorithm Interaction	This section assesses how service providers and service users interact with the algorithmic system, and how service users or providers may be affected by the way the algorithmic system is designed.
Service User Engagement and Involvement	This section assesses how service users were engaged throughout the algorithmic system lifecycle, including in design and development of the system as well as ongoing processes for engagement including feedback and complaint processes.
Algorithmic System	This section assesses how aspects of the algorithmic system including the model, data and design features may influence principles of trauma-informed care.

Section 2 is structured by type of algorithmic system. Sometimes questions from Section 1 are repeated in Section 2 in the context of the specific algorithmic system. When completing a repeated question service providers are encouraged to use the opportunity to reflect further on why that specific type of algorithmic system may cause concerns in that area.

The below table classifies algorithmic systems included in Section 2. These cover the most common types of algorithmic systems used in social service delivery. The list is not exhaustive, and if organisations are using systems not covered by this list they should adapt the toolkit accordingly.

#	**Type of System**	**Description**
2.1	Chatbot	Engaging in "conversational" interactions between machines and humans (possibly involving multiple media such as voice, text and images). In the context of social service delivery examples may include mental health support chatbots, informational bots or relationship support chatbots.
2.2	Recommender	Developing a profile of an individual to learn and adapt its output (or recommendations) to that individual over time. In social service delivery examples may include personalised referrals to support services, targeted advertisements for social service organisations or targeted resources for mental health support.
2.3	Identification and Recognition	Identifying and categorising data (e.g. image, video, audio and text) into specific classifications as well as image segmentation and object detection. In social service delivery context examples may include facial or voice recognition technology for identity matching for access to services.
2.4	Risk Assessment	Using past and existing behaviours to predict likelihood of relevant future outcomes. In the context of social service delivery examples may include algorithmic systems which predict risk of harm to children or health outcomes for resource allocation.
2.5	Detection	Connecting data points to detect patterns or events, as well as outliers or anomalies which may trigger investigations. In social services examples may include welfare fraud detection.
2.6	Goal Driven Optimisation	Finding the optimal solution to a problem for a cost function or predefined goal. In social service delivery context examples may include development of rosters for service user visits by service providers.

Your Organisation and the Service

Prior to commencing the assessment, please complete this section about your organisation.

What are your organisation's values and mission statement?	What are the objectives of your service?

Introducing an Algorithmic System

Prior to commencing the assessment, please complete the following table which prompts reflection on if and how an algorithmic system could help improve your service.

What is the problem you are seeking to solve within your service by introducing an algorithmic system?	**Is it necessary to introduce an algorithmic system to solve the problem? If so, why?**	**Are there alternative ways of solving the problem? If so, how do these alternatives relate to principles of trauma informed care?**	**Identify the key stakeholders for the algorithm supported service (inc. service users, users of the algorithmic system, other relevant members of the service provider organisation etc.)**

Knowing the Algorithmic System

Prior to commencing the assessment, please complete the following table based on the above taxonomy with information about the algorithmic system which you intend to use or are currently using as part of the algorithm supported service. The algorithmic system may have more than one function or may not fit into any of the given 'types'. For instance, a mental health support chatbot may also be recommending mental health services in which you should complete both the 'Chatbot' and 'Recommender' sub-sections. If none of the 'types' given apply to your algorithmic system, then you are only required to complete Sections 1 and 2. You can also add in a section with questions applicable to your own system. Otherwise, please complete all the sections corresponding to the type of system being used.

Type	**What is the purpose of the algorithmic system?**	**How does this purpose align with your organisational mission statement/values and objectives of your service?**	**Describe the algorithmic system and the role it plays within the algorithm supported service**
☐ Chatbot ☐ Recommender ☐ Identification and Recognition ☐ Risk Assessment ☐ Detection ☐ Goal driven optimisation ☐ Other			

1. CRITICAL CONSIDERATIONS

The following section prompts you to reflect on critical considerations that are applicable to all algorithm supported services. You should ensure that you have thoroughly reflected and engaged with the below considerations. The questions in this section represent the absolute minimum when you are implementing a trauma informed approach within your algorithm supported service.

#	For the algorithm supported service, to what extent…	Not at all	To some extent	To a great extent	Evidence	Action/Plan	Priority	Principle
1.	are service users aware that an algorithmic system is being used?	☐	☐	☐			☐	Trust and Transparency
2.	have designers considered if it could cause or trigger trauma?	☐	☐	☐			☐	Safety
3.	has the psychological, emotion and cultural safety of service users been considered?	☐	☐	☐			☐	Safety Intersectionality
4.	does the service user have an informed choice about when their personal information is accessed or shared?	☐	☐	☐			☐	Trust and Transparency Empowerment and Choice
5.	can the service user choose to interact with a human?	☐	☐	☐			☐	Empowerment and Choice Collaboration

6.	can service users regularly provide feedback?	☐	☐	☐			☐	Collaboration Choice
7.	can service users choose to make a complaint, appeal or review directly to a human?	☐	☐	☐			☐	Trust and Transparency Empowerment and Choice
8.	could it unfairly disadvantage different service user groups?	☐	☐	☐			☐	Intersectionality Safety
9.	could the algorithmic system harmfully discriminate against any service user?	☐	☐	☐			☐	Intersectionality Safety
10.	are there supportive processes for responding to harm or distress caused by the algorithmic system?	☐	☐	☐			☐	Safety Trust and Transparency

2. GENERAL

The following section asks you to consider key aspects of trauma informed services. For each question, please consider to what extent it is true in regards to the algorithmic system you are applying the toolkit to, and also what evidence you can provide for this. There is also space to suggest actions or plans that are needed to improve the trauma informed quality of the algorithmically enhanced service.

2.1 Service Processes, Procedures and Plans

#	For the algorithm supported service, to what extent …	Not at all	To some extent	To a great extent	Evidence	Action/Plan	Priority	Principle
11.	could it harm relationships between service providers and users?	☒	☐	☐			☐	Collaboration Trust & Transparency Safety
12.	are safety concerns promptly responded to?	☐	☐	☐			☐	Safety Trust and Transparency
13.	are service users informed of significant changes?	☐	☐	☐			☐	Safety Trust and Transparency
14.	do these processes consider the diverse social and cultural needs of users?	☐	☐	☐			☐	Safety Intersectionality

15.	do service users understand why certain questions are being asked or data collected?	☐	☐	☐			☐	Trust and Transparency
16.	can service users choose how their personal information from external sources is used?	☐	☐	☐			☐	Trust and Transparency
17.	have designers considered the diverse social and cultural needs of service users?	☐	☐	☐			☐	Intersectionality
18.	are there processes that appropriately respond to disclosures of traumatic experiences?	☐	☐	☐			☐	Safety
19.	is the service user able to choose whether or not to disclose their traumatic experiences?	☐	☐	☐			☐	Empowerment and Choice
20.	has its impact on groups with diverse social and cultural needs been considered and monitored?	☐	☐	☐			☐	Safety Intersectionality

2.2 Human Algorithm Interaction

#	For the algorithm supported service, to what extent …	Not at all	To some extent	To a great extent	Evidence	Action/Plan	Priority	Principle
21.	can it respond appropriately to service users experiencing distress?	☐	☐	☐			☐	Safety Empowerment and Choice
22.	are outputs consistent given similar inputs?	☐	☐	☐			☐	Safety
23.	does it respond consistently to outputs of the algorithmic system? *[Within a defined framework rather than in an ad-hoc and arbitrary manner]*	☐	☐	☐			☐	Safety Trust and Transparency
24.	can the service user choose how the algorithmic system responds to their reports of distress?	☐	☐	☐			☐	Empowerment and Choice
25.	could the algorithmic system displace or limit human judgement and discretion in service provision?	☐	☐	☐			☐	Trust and Transparency
26.	can service users customise their interaction?	☐	☐	☐			☐	Empowerment and Choice

2.3 Service User Engagement and Involvement

#	For the algorithmic supported service, to what extent …	Not at all	To some extent	To a great extent	Evidence	Action/Plan	Priority	Principle
27.	are service users involved in determining whether the service is automated?	☐	☐	☐			☐	Collaboration
28.	can service users opt-out of being subject to the algorithmic system without compromising service quality?	☐	☐	☐			☐	Empowerment and Choice
29.	have service users been engaged in the design, development, deployment, monitoring and evaluation of the algorithmic system?	☐	☐	☐			☐	Collaboration
30.	are service users consulted to ensure their personal information is accurate?	☐	☐	☐			☐	Collaboration
31.	have service users been consulted asked if they feel safe engaging with the algorithmic system?	☐	☐	☐			☐	Safety Collaboration
32.	have you consulted with a diverse range of service users?	☐	☐	☐			☐	Intersectionality Collaboration

33.	have perspectives of service users been integrated?	☐	☐	☐			☐	Collaboration
34.	can service users easily access the complaints, appeals and review processes?	☐	☐	☐			☐	Trust and Transparency Empowerment and Choice
35.	are strategies in place to ensure that consultation with service users about the algorithmic system does not trigger or cause trauma?						☐	Safety

2.4 Algorithmic System

#	For the algorithmic supported service, to what extent …	Not at all	To some extent	To a great extent	Evidence	Action/Plan	Priority	Principle
36.	are service users made aware of the purpose the algorithmic system?	☐	☐	☐			☐	Trust and Transparency
37.	does the algorithmic system's training data representative of the diversity of service users?	☐	☐	☐			☐	Intersectionality Safety
38.	is the algorithmic system sensitive to the complexity of intergenerational trauma of service users? *[Depending on geopolitical context in which this toolkit is used, the relevant groups of service users affected by systemic intergenerational trauma may differ. For example, in Australia a relevant group would be its First Nations Peoples or in the United States it may be people from African-American backgrounds.]*	☐	☐	☐			☐	Intersectionality

39.	are there procedures that prevent the algorithmic system from producing harmful bias and discrimination?	☐	☐	☐			☐	Intersectionality Safety
40.	can service users choose to have their personal information erased?	☐	☐	☐			☐	Safety Empowerment and Choice
41.	has the algorithmic system considered cultural context of trauma and healing in its design and operation?	☐	☐	☐			☐	Intersectionality
42.	is current and clear information about the limitations of the algorithmic system available?	☐	☐	☐			☐	Trust and Transparency
43.	is the algorithmic system vulnerable to cyber-attack?	☐	☐	☐			☐	Safety Trust and Transparency
44.	could this algorithmic system reinforce or amplify existing stereotypes?	☐	☐	☐			☐	Intersectionality
45.	could the algorithmic system be used to misuse information or misrepresent service users?	☐	☐	☐			☐	Safety

46.	could the algorithmic system incorrectly generate information about service users?	☐	☐	☐			☐	Safety Trust and Transparency Intersectionality
47.	could this algorithmic system's data limit or misrepresent the identity of a service user?	☐	☐	☐			☐	Safety Intersectionality
48.	could the algorithmic system take away decision-making control from service users?	☐	☐	☐			☐	Empowerment and Choice Safety
49.	could the algorithmic system be used by perpetrators of domestic and family violence to harm victim/survivors?	☐	☐	☐			☐	Safety

3. Types of Algorithmic Systems

3.1 Chatbots

Interpreting and creating content to power conversational and other interactions between machines and humans (possibly involving multiple media such as voice, text and images).

#	For the chatbot, to what extent …	Not at all	To some extent	To a great extent	Evidence	Action/Plan	Priority	Principle
50.	can the service user customise the digital interface?	☐	☐	☐			☐	Empowerment and Choice
51.	can it appropriately respond to service users who may be emotionally heightened?	☐	☐	☐			☐	Safety
52.	does the service user have a choice in how their distress is responded to?	☐	☐	☐			☐	Safety Empowerment and Choice
53.	are service users informed about mandatory reporting responsibilities if they or someone else is at risk of harm?	☐	☐	☐			☐	Trust and Transparency
54.	does it provide information or avenues for further specialised assistance?	☐	☐	☐			☐	Safety Empowerment and Choice

55.	does it respect service users' boundaries?	☐	☐	☐			☐	Collaboration
56.	could it foster unhealthy dependence?	☐	☐	☐			☐	Empowerment and Choice Safety
57.	are service users given a choice about how it addresses them?	☐	☐	☐			☐	Empowerment and Choice Intersectionality
58.	has it been trained in appropriate communication methods?	☐	☐	☐			☐	Safety Trust and Transparency
59.	does it use strength-based language rather than pre-conceived labels of service users (e.g. describing a person as 'having difficulty getting her needs met' rather than 'attention-seeking')?	☐	☐	☐			☐	Safety
60.	is inclusive language used?	☐	☐	☐			☐	Safety Intersectionality
61.	is safe language used?	☐	☐	☐			☐	Safety
62.	can service users interact with it in the language of their choice?	☐	☐	☒			☐	Intersectionality

63.	does any interaction consider the cultural context of trauma and healing?	☐	☐	☐			☐	Intersectionality
64.	could it facilitate trauma caused by other service users?	☐	☐	☒			☐	Safety
65.	could it be used to manipulate a service user to reveal personal information?	☒	☐	☐			☐	Safety Empowerment and Choice
66.	could it provide advice that should only be provided by trained professionals?	☒	☐	☐			☐	Safety

3.2 Recommender Systems

Developing a profile of an individual and learning and adapting its output to that individual over time.

#	For the recommender system, to what extent …	Not at all	To some extent	To a great extent	Evidence	Action/Plan	Priority	Principle
67.	could it generate triggering content?	☐	☐	☐			☐	Safety
68.	can a service user influence the recommendations made to them?	☐	☐	☐			☐	Empowerment and Choice
69.	have appropriate and visible content warnings been included?	☒	☐	☐			☐	Safety
70.	do service users have capacity to report inappropriate content?	☐	☐	☐			☐	Empowerment and Choice Safety
71.	are there prompt review and takedown processes for reported content?	☐	☐	☐			☐	Trust and Transparency Safety
72.	could it lead to harmful personal interests amongst service users (e.g. white supremacy, disordered eating)?	☐	☐	☐			☐	Safety

73.	could recommendations generate harmful outcomes?						☐	Safety

3.3 Identification and Recognition Systems

Identifying and categorising data (e.g. image, video, audio and text) into specific classifications as well as image segmentation and object detection.

#	For the identification or recognition system, to what extent …	Not at all	To some extent	To a great extent	Evidence	Action/Plan	Priority	Principle
74.	is it required to access key social services?	☐	☐	☐			☐	Safety
75.	is it accessible to service users with diverse social and cultural needs?	☐	☐	☐			☐	Empowerment and Choice Intersectionality
76.	does the service user know when they are being surveilled?	☐	☐	☐			☐	Empowerment and Choice Safety
77.	can the service user meaningfully avoid being surveilled?	☐	☐	☐			☐	Empowerment and Choice Safety
78.	could it inappropriately categorise service users?						☐	Empowerment and Choice Intersectionality
79.	could it inaccurately identify the service user?						☐	Safety Trust and Transparency
80.	could it be manipulated to impersonate a service user?						☐	Safety Trust and Transparency

3.4 Risk Assessment Systems

Using past and existing behaviours to predict future outcomes (e.g. predictive risk algorithms).

#	For the risk assessment system, to what extent …	Not at all	To some extent	To a great extent	Evidence	Action/Plan	Priority	Principle
81.	are do service providers understand how an assessment is made?	☐	☐	☐			☐	Trust and Transparency
82.	are clear explanations given to service users for why a decision was made?	☐	☐	☐			☐	Trust and Transparency
83.	could the communication of a decision cause or trigger trauma?	☐	☐	☐			☐	Safety Trust and Transparency
84.	has it been assessed for potential harmful bias or discrimination?	☐	☐	☐			☐	Intersectionality Safety
85.	are there accessible and effective appeal processes?	☐	☐	☐			☐	Trust and Transparency Empowerment and Choice
86.	could service users who have experienced trauma be disadvantaged or harmed by the risk assessment process?	☐	☐	☐			☐	Safety
87.	has the service provider considered the impact of errors?	☐	☐	☐			☐	Safety Trust and Transparency

88.	has the service organisation considered the appropriateness of making decisions based on risk?	☐	☐	☐			☐	Safety
89.	Is the assessment based on arbitrary factors?	☐	☐	☐			☐	Empowerment and Choice, Trust and Transparency, Safety

3.5 Detection Systems

Connecting data points to detect patterns, as well as outliers or anomalies (e.g. fraud detection algorithms).

#	For detection systems, to what extent …	Not at all	To some extent	To a great extent	Evidence	Action/Plan	Priority	Principle
90.	are service users informed that they are being surveilled?	☐	☐	☐			☐	Trust and Transparency
91.	are service users informed when and why they are being investigated?	☐	☐	☐			☐	Trust and Transparency
92.	could the communication of a decision cause or trigger trauma?	☐	☐	☐			☐	Safety Trust and Transparency
93.	are service users given clear explanations for decisions?	☐	☐	☐			☐	Trust and Transparency
94.	has it been assessed for potential harmful bias or discrimination?	☐	☐	☐			☐	Intersectionality Safety
95.	is there human oversight over the decision?	☐	☐	☐			☐	Trust and Transparency Safety
96.	are there accessible and effective appeal processes?	☐	☐	☐			☐	Trust and Transparency Empowerment and Choice

3.6 Goal Driven Optimisation Systems

Finding the optimal solution to a problem for a cost function or predefined goal.

#	For goal driven optimisation, to what extent …	Not at all	To some extent	To a great extent	Evidence	Action/Plan	Priority	Principle
97.	does the system consider service users wellbeing?	☐	☐	☐			☐	Safety
98.	is there human oversight over the outcome prior to a decision being made?	☐	☐	☐			☐	Trust and Transparency Safety
99.	does it include people with diverse social and cultural needs?	☐	☐	☐			☐	Intersectionality
100.	have unintended consequences of its use been considered?	☐	☐	☐			☐	Safety Intersectionality

4. Prioritising Areas for Improvement

Now that you have assessed your algorithm supported service with the toolkit, you are able to identify the key priorities for improvement that may arise from your algorithm supported service. The purpose of this section is to structure this identification and assessment process based on your responses and reflections in completing the toolkit. What constitutes key areas for improvement will be dependent on the unique context of your algorithm supported service. We encourage you to reflect deeply about which issues you think may be salient in your context from a trauma informed lens. The priority areas you have identified while you were completing the questionnaire should assist you with this process.

For identified areas for improvement, reflect on the questions below to evaluate the potential risks if that issue remains unaddressed.

Assessment Metric	**If this issue remains unaddressed …**
Severity	how **acutely** could the algorithm supported service cause harm to service users?
Scale	how **broadly** could the harm from the algorithm supported service be experienced across service user populations or groups?
Probability	how **likely** is the algorithm supported service to cause harm?
Frequency	how **often** would a service user or a group of service users experience harm?

Please count the number of priority areas you have marked in the toolkit and count how many you have in each of the boxes below. This will assist you in identifying the key areas for improvement across the various domains and principles. Once you have identified your areas of focus, please complete the table on the following page detailing your particular areas for improvement.

Domain	**Empowerment and Choice**	**Collaboration**	**Trust and Transparency**	**Safety**	**Intersectionality**
Service Procedures, Processes and Plans					
Human Algorithm Interaction					
Service User Engagement and Involvement					
Algorithmic System					
Your Algorithmic System (e.g. Chatbot)					

Empowerment and Choice

Areas for Improvement
What is the potential for harm if this issue remains unaddressed …?
Action/Plan for Improvement

Collaboration

Areas for Improvement
What is the potential for harm if this issue remains unaddressed …?
Action/Plan for Improvement

Trust and Transparency

Areas for Improvement
What is the potential for harm if this issue remains unaddressed …?
Action/Plan for Improvement

Safety

Areas for Improvement
What is the potential for harm if this issue remains unaddressed …?
Action/Plan for Improvement

Intersectionality

Areas for Improvement
What is the potential for harm if this issue remains unaddressed …?
Action/Plan for Improvement

5. Next Steps

Identify	• any gaps in your knowledge and obtain any further information you may require to complete your evaluation.
Re-assess	• whether it is possible to use an algorithmic system in a trauma-informed manner.
Consult	• relevant stakeholders within your organisation and within service user groups.
Action	• any changes to address identified priority issues.
Document	• changes which are made to the algorithm supported service as a result of completing the toolkit.
Re-engage	• with the toolkit for ongoing monitoring and evaluation of the algorithm supported service at regular intervals.

6. Further resources

Implementing a Trauma Informed Approach

Substance Abuse and Mental Health Services Administration: *Practical Guide for Implementing a Trauma-Informed Approach*. SAMHSA Publication No. PEP23-06-05-005. Rockville, MD: National Mental Health and Substance Use Policy Laboratory. Substance Abuse and Mental Health Services Administration, 2023. **Access here.**	This practical guide developed by the US Substance Abuse and Mental Health Services provides implementation strategies across multiple domains—from governance to staff training to evaluation—based on their original *Concept of Trauma* publication. This resource is intended for anyone involved in organization-level change, including practitioners, state and local officials and policymakers.
Guarino, K., Soares, P., Konnath, K., Clervil, R., and Bassuk, E. (2009). *Trauma-Informed Organizational Toolkit.* Rockville, MD: Center for Mental Health Services, Substance Abuse and Mental Health Services Administration, and the Daniels Fund, the National Child Traumatic Stress Network, and the W.K. Kellogg Foundation. **Access here.**	The Trauma-Informed Organizational Self-Assessment is a tool that organizations can use to examine their current practices and take specific steps to become trauma-informed.
Mental Health Coordinating Council (MHCC) 2018, *Trauma-Informed Care and Practice Organisational Toolkit (TICPOT): An Organisational Change Process Resource, Stage 1 - Planning and Audit*, Authors: Henderson, C (MHCC), Everett, M. Isobel S (Sydney LHD). **Access here.**	TICPOT is a resource designed to assist services and their workforce in quality improvement initiatives and organisational change processes. It can be used to embed trauma informed care principles into every aspect of an organisation. It is targeted at a broad range of services both in the public and community-based contexts across the mental health and human service systems and sectors. It provides an overview of the principles of trauma-informed care and practice and guidelines for planning and conducting an organisational audit.
Maxine Harris and Roger D Fallot, 'Envisioning a Trauma-informed Service System: A Vital Paradigm Shift' (2001)	This paper introduced the concept of a trauma informed approach to service delivery and has been built on by subsequent publications.

2001(89) New Directions for Mental Health Services 3. Access here.	

Assessing Algorithmic Systems

OECD, Catalogue of Tools & Metrics for Trustworthy AI. **Access here.**	The catalogue developed by the OECD allows anyone to find tools and methods from around the world for making AI trustworthy. Technical tools address AI-related issues such as bias detection, transparency and explainability, performance, robustness, safety and security against attacks. Procedural tools provide operational and process-related guidance. Educational tools cover all means for building awareness, such as preparing and upskilling stakeholders involved in or affected by implementing an AI system.
Microsoft. Assessing Harm: A guide for Tech Builders. **Access here.**	This booklet developed by Microsoft features questions technology builders can ask as they build and refine their own technologies, and scenarios that imagine potential risks or misuse.
National Institute of Standards and Technology, AI Risk Management Framework Playbook. **Access here.**	The framework was designed by the US National Institute of Standards to equip organizations and individuals with approaches that increase the trustworthiness of AI systems, and to help foster the responsible design, development, deployment, and use of AI systems over time.
Automated Decision-Making Better Practice Guide (Commonwealth Ombudsman, 2019) **Access here.**	This guide developed by the Australian Commonwealth Ombudsman is intended to be a practical tool for agencies and includes a checklist designed to assist managers and project officers during the design and implementation of new automated systems, and with ongoing assurance processes once a system is operational.
Artificial Intelligence Assurance Framework (NSW Government). **Access here.**	This framework developed by the NSW government assists project teams using AI to comprehensively analyse and document their projects' AI specific risks. It also

	assists teams to implement risk mitigation strategies and establish clear governance and accountability measures.
Central Digital and Data Office and Office for Artificial Intelligence, A guide to using artificial intelligence in the public sector (UK Government, 2019). **Access here.**	This collection of resources by the UK Government covers how to assess if using AI will help meet user needs, the public sector can best use AI and how to implement AI ethically, fairly and safely.

Understanding Algorithmic Systems

OECD Framework for the Classification of AI Systems (OECD Digital Economy Papers, No 323, OECD, February 2022). **Access here.**	The OECD has developed this tool to evaluate AI systems from a policy perspective to help policy makers, regulators, legislators and others characterise AI systems deployed in specific contexts.
AI Incident Database. **Access here.**	The AI Incident Database indexes the history of harms or near harms realized in the real world by the deployment of artificial intelligence systems. The Database includes a searchable catalogue of incidents where AI systems have caused harm around the world.
AIAAIC Repository. **Access here.**	The AIAAIC Repository (standing for AI, Algorithmic, and Automation Incidents and Controversies) is an open resource database that details incidents and controversies driven by and relating to artificial intelligence, algorithms, and automation. The Repository enables users to identify, examine, and understand the nature, risks, and impacts of AI, algorithms, and automation.

Appendix B **Review of Principles of Trauma Informed Care**

Source 1. Menschner, C., & Maul, A. (2016). *Key Ingredients for Successful Trauma-Informed Care Implementation*. Center for Health Care Strategies.	
Principles	definition
empowerment	"Using individuals' strengths to empower them in the development of their treatment" (Menschner and Maul, 2016, p. 3)
choice	"Informing patients regarding treatment options so they can choose the options they prefer" (Menschner and Maul, 2016, p. 3)
collaboration	"Maximizing collaboration among health care staff, patients, and their families in organizational and treatment planning" (Menschner and Maul, 2016, p. 3)
safety	"Developing health care settings and activities that ensure patients' physical and emotional safety" (Menschner and Maul, 2016, p. 3)
trustworthiness	"Creating clear expectations with patients about what proposed treatments entail, who will provide services, and how care will be provided." (Menschner and Maul, 2016, p. 3)
Source 2. Mahon, D. (2022). *Trauma-Responsive Organisations: The Trauma Ecology Model*. Emerald Publishing Limited.	
empowerment and choice	"Trauma-informed practices use strengths-based approaches that are empowering and support individuals to take control of their lives and service use. Such approaches are vital because many trauma survivors will have experienced an absolute lack of power and control. Choice is provided around the type of interventions and services provided Including, where possible the characteristics of the practitioner providing services.
empowerment and choice	"Trauma-informed practices use strengths-based approaches that are empowering and support individuals to take control of their lives and service use. Such approaches are vital because many trauma survivors will have experienced an absolute lack of power and control.

	Choice is provided around the type of interventions and services provided Including, where possible the characteristics of the practitioner providing services.
Collaboration	"Trauma-informed practices understand that there is a unilateral aspect to relationships in mental health care, with one person acting as helper to another. These roles can replicate power imbalances and reinforce a sense of disability and helplessness in those being supported Therefore, collaboration in relationships and interventions is sought out. Collaboration is conducted at the individual service user, family members and employees level and partnerships with other professionals and communities."
Peer-support	"Peer-support and mutual self-help are key vehicles for establishing safety and hope, building trust, enhancing collaboration, peers can promote recovery by sharing their lived experienced and the road to be travelled."
Safety	"Central to trauma experiences are threats to the person's safety and often to the integrity of their identity. Consequently, trauma-informed practices ensure that employees and service users are safe from physical, emotional, and psychological distress, including from corrosive leadership practice" (ch 2, Table 2)
Trust and transparency	"Developing clear expectations regarding the types of treatments and services offered, who will be delivering them, and how they will occur. Many service users may have experienced coerced treatments by providers and organisations, with no input and against their will , thus, providing choice and preference is essential during engagement with services."
Cultural, historical and gender issues	"Leverage the healing value of traditional cultural connections; incorporates policies, protocols, and processes that are responsive to the racial, ethnic and cultural needs of individuals served and employees within the organisation and wider community." (Ch 2, Table 2). Source: Adapted from SAMHSA (2014) and Menschner and Maul (2016).
Source 3. Huang, L. N., Flatow, R., Biggs, T., Afayee, S., Smith, K., Clark, T., & Blake, M. (2014). *SAMHSA's Concept of Trauma and Guidance for a Trauma-Informed Approach* (HHS Publication No. (SMA) 14-4884). Substance Abuse and Mental Health Services Administration (SAMHSA). https://archive.hshsl.umaryland.edu/handle/10713/18559	
Empowerment, voice and choice	"Throughput the organization and among the clients served, individuals' strengths and experiences are recognised and built upon. The organization fosters a belief in the primacy of the people served, in resilience, and in the ability of individuals, organizations, and

	communities to heal and promote recovery from trauma. The organization understands that the experience of trauma may be a unifying aspect in the lives of those who run the organization, who provide the services, and/or who come to the organization for assistance and support. As such, operations, workforce development and services are organized to foster empowerment for staff and clients alike. Organizations understand the importance of power differentials and ways in which clients, historically, have been diminished in voice and choice and are often recipients of coercive treatment. Clients are supported in shared decision-making, choice, and goal setting to determine the plan of action they need to heal and move forward. They are supported in cultivating self-advocacy skills. Staff are facilitators of recovery rather than controllers of recovery. Staff are empowered to do their work as well as possible by adequate organizational support. This is a parallel process as staff need to feel safe, as much as people receiving services." (p. 11)
Empowerment, voice and choice	"Throughput the organization and among the clients served, individuals' strengths and experiences are recognised and built apon. The organization fosters a belief in the primacy of the people served, in resilience, and in the ability of individuals, organizations, and communities to heal and promote recovery from trauma. The organization understands that the experience of trauma may be a unifying aspect in the lives of those who run the organization, who provide the services, and/or who coe to the organization for assistance and support. As such, operations, workforce development and services are organized to foster empowerment for staff and clients alike. Organizations understand the importance of power differentials and ways in which clients, historically, have been diminished in voice and choice and are often recipients of coercive treatment. Clients are supported in shared decision-making, choice, and goal setting to determine the plan of action they need to heal and move forward. They are supported in cultivating self-advocacy skills. Staff are facilitators of recovery rather than controllers of recovery. Staff are empowered to do their work as well as possible by adequate organizational support. This is a parallel process as staff need to feel safe, as much as people receiving services." (p. 11)
Collaboration and mutuality	"Importance is placed on partnering and the levelling of power differences between staff and clients and among organizational staff from clerical and housekeeping personnel, to professional staff to administrators, demonstrating that healing happens in relationships and

	in the meaningful sharing of power and decision-making. The organization recognizes that everyone has a role to play in a trauma-informed approach. As one expert stated: "one does not have to be a therapist to be therapeutic." (p. 11)
Safety	"Throughout the organization, staff and the people they serve, whether children or adults, feel physically and psychologically safe; the physical setting is safe and interpersonal interactions promote a sense of safety. Understanding safety as defined by those who serve is a high priority." (p. 11)
Trustworthiness and transparency	"Organizational operations and decisions are conducted with transparency with the goal of building and maintaining trust with clients and family member, among staff, and others involved in the organization." (p. 11)
Cultural historical and gender issues	"The organizational activity actively moves past cultural stereotypes and biases (e.g. based on race, ethnicity, sexual orientation, age, religion, gender-identity, geography. Etc.); offers access to gender responsive services; leverages the healing value of traditional cultural connections; incorporates policies, protocols and processes that are responsive to racial, ethnic and cultural needs of individuals served; and recognizes and addresses historical trauma." (p. 11)
Source 4. Bowen, E. A., & Murshid, N. S. (2016). Trauma-Informed Social Policy: A Conceptual Framework for Policy Analysis and Advocacy. *American Journal of Public Health*, *106*(2), 223–229. https://doi.org/10.2105/AJPH.2015.302970	
Empowerment	"The principle of empowerment may be reflected by the processes through which the policy is created as well as the policy objectives. Policymaking processes can be broadly characterized as reflecting both top-down and bottom-up elements, the latter referring to the active involvement of stakeholders who are directly affected by the target problem or issue in shaping the policy.16,48 Bottom-up policymaking is a compelling vehicle for empowerment … From a trauma-informed perspective, it is critical that empowerment in social policy reflect not only a rhetoric of liberation but actual shared power in terms of extending decision-making ability to the target populations of social policies. T" (p. 226)
Choice	"Promoting choice to the greatest extent possible has been recognized as a principle of strengths-based social policy and is key to the trauma-informed policy perspective." Eg food security (p. 226)

Collaboration and peer support	"Trauma-informed social policy can embody collaboration and peer support in the extent to which the policy prioritizes the indigenous knowledge and experiences of the policy's target population, in addition to or instead of outside professional expertise" (p. 225)
Safety	"Trauma-informed social policy should make provisions for the basic safety of vulnerable populations … Often the question of "safety for whom?"must be asked in trauma-informed policy analysis, because many policies may privilege the safety—in rhetoric if not in actual outcome—of one group at the expense of the well-being of another. T" (p. 224)
Trustworthiness and transparency	"In social policy, trustworthiness is tied to the transparency of the policy's intended goals or outcomes and the procedures by which these goals will be attained … Another factor limiting transparency in social policy is the widespread trend toward devolution in many policy areas. Devolution has allowed states more autonomy in policy design and implementation, including theability to waive certain regulations for some federal policies." (p. 225)
Intersectionality	For social policies to be trauma informed,they need to take into account what Collins defines as intersectionality: "an analysis claiming that systems of race, social class, gender, sexuality, ethnicity, nation, and age form mutually constructing features of social organization."64(p299) This intersectionality must allow an understanding of discrimination, privilege, and human rights violations that occur as a consequence of the combination of the identities to which an individual may subscribe. For example, an undocumented immigrant from a lowincome family in the Middle East may be discriminated against on the basis of race, ethnicity, social class, gender, and nationality. In addition to being a stressor with broad mental health implications, research indicates that instances of "everyday" discrimination and microaggressions related to multiple marginalized identities are significant predictors of posttraumatic stress. Trauma-informed policies can address intersectionality by taking measures to prevent overt discrimination …" (p. 227)
Source 5. NSW Health. (2022, August 17). *What is trauma-informed care? - Principles for effective support*. https://www.health.nsw.gov.au:443/mentalhealth/psychosocial/principles/Pages/trauma-informed.aspx	
Based on an understanding of the impact of trauma	"Trauma-informed care is based on the understanding that: a significant number of people living with mental health conditions have experienced trauma in their lives

	trauma may be a factor for people in distress the impact of trauma may be lifelong trauma can impact the person, their emotions and relationships with others."
Empowerment	"is empowering people a key focus?"
Choice	"do you provide opportunity for choice?"
Collaboration	"do you communicate a sense of 'doing with' rather than 'doing to'?"
Safety	"emotional as well as physical e.g. is the environment welcoming?"
Trust	"is the service sensitive to people's needs?"
Respect for diversity	"do you respect diversity in all its forms?"
Source 6. Elliott, D. E., Bjelajac, P., Fallot, R. D., Markoff, L. S., & Reed, B. G. (2005). Trauma-informed or trauma-denied: Principles and implementation of trauma-informed services for women. *Journal of Community Psychology*, *33*(4), 461–477. https://doi.org/10.1002/jcop.20063	
Principle 1. Trauma-Informed Services Recognize the Impact of Violence and Victimization on Development and Coping Strategies	"When a trauma-informed program recognizes the long-term and pervasive impact of interpersonal violence and childhood abuse, the experiences of survivors are validated and the difficulties they face in seeking services are recognized." (p. 465)
Principle 2. Trauma-Informed Services Identify Recovery From Trauma as a Primary Goal	"Trauma-informed programs offer either specialized services that directly address recovery from past trauma (trauma-specific services) or integrate a woman's care with an agency that does provide those services." (p. 465)
Principle 3. Trauma informed services employ an empowerment model	"Ideally, a primary goal of any service provision for survivors is to facilitate the client's ability to take charge of her life, specifically, to have conscious choice and control over her actions. An empowerment model incorporates those elements of a helping relationship that can increase the client's power in personal, interpersonal, and political spheres (Gutierrez, Parsons, & Cox, 1998). The empowerment model is essential to recovery from the overwhelming fear and helplessness that is the legacy of victimization." (p. 465)
Principle 4. Trauma-Informed Services Strive to Maximize	"Despite the great need and vulnerability experienced by many survivors, the ultimate goal is to work collaboratively with the survivor to increase her access to conscious choice, more options, and a sense of control over important life decisions. It is only

a Woman's Choices and Control Over Her Recovery	through this personal experience of choice and control that a woman reclaims her right to direct her own life and pursue her personal goals and dreams." (p. 466)
Principle 5. Trauma-Informed Services Are Based in a Relational Collaboration	"Staff must be aware of the inherent power imbalance in the helper-helped relationship and do their best to flatten the hierarchy. Interpersonal violence involves a perpetrator and a victim. The trauma of this "power over" experience for the victim is best healed in a very different type of relationship, one that is collaborative and empowering (Miller & Stiver, 1997)." (p. 466)
Principle 6. Trauma-Informed Services Create an Atmosphere That Is Respectful of Survivors' Need for Safety, Respect, and Acceptance	"Human service agencies need to work with the women they serve to modify staff approaches, programs, procedures, and, in some cases, the physical setting to create a place perceived as safe and welcoming for survivors … A feeling of safety is also enhanced when the provider and all staff at the agency adhere to the confidentiality policy, give clear information, are consistent and predictable, and give the woman as much control and choice over her experience as possible, including her right to set limits and modify the process. Clear boundaries and well-defined roles are essential to providing a safe environment for survivors." (p. 467)
Principle 7. Trauma-Informed Services Emphasize Women's Strengths, Highlighting Adaptations Over Symptoms and Resilience Over Pathology	"Too often, programs focus so intently on problems that they miss the many strengths a person brings to the human service setting (Brown & Worth, 2000)." (p. 467)
Principle 8: The Goal of Trauma-Informed Services Is to Minimize the Possibilities of Retraumatization	"This principle rests on the premise that service providers recognize and understand the potential for retraumatization for women in treatment. When one understands the abuse of power inherent in all victimization, it becomes clear that the power differential between the person seeking help and the person offering it will be threatening to a woman who experienced abuse at the hands of those whom she depended on in childhood. Once service providers understand the potential for retraumatization and the survivor's fear and sometimes expectations! of it, then it is possible to work explicitly to protect against it.
Principle 9. Trauma-Informed Services Strive to Be Culturally	"As mentioned under Principle 3, treatment providers must be able to understand a woman's cultural context. Cultural competency includes having the knowledge and

Competent and to Understand Each Woman in the Context of Her Life Experiences and Cultural Background	skills to work within the client's culture, understanding how one's own cultural background and the program influence transactions with the client (Fong & Furuto, 2001). Understanding the influence of someone's culture is essential to making an effective therapeutic connection and being part of a woman's recovery. The meaning one gives violence and trauma can vary by culture. Healing takes place within a woman's cultural context and support network, and different cultural groups may have unique resources that support healing. Cultural competence does not require that every service provider have detailed knowledge of every culture, but rather that he or she recognize the importance of cultural context. It is often helpful to ask questions, be open to being educated, and try to understand the woman's experience and responses through the lens of her cultural context. " (p. 468)
Principle 10. Trauma-Informed Agencies Solicit Consumer Input and Involve Consumers in Designing and Evaluating Services	"Women should be involved in designing treatment services and be part of an ongoing evaluation of those services (Prescott, 2001). They can be on an advisory board that reviews program design, serve as paid consumer specialists, or participate in focus groups and or in regular feedback forums about how to respond to program evaluations and improve services" (p. 469)
Source 7. Henderson, C., Everett, M., & Isobel, S. (2018). *Trauma-Informed Care and Practice Organisational Toolkit (TICPOT): An Organisational Change Process Resource, Stage 1—Planning and Audit*. Mental Health Coordinating Council (MHCC).	
Understanding trauma and its impact	"A trauma-informed approach recognises the prevalence of trauma and understands the impact of trauma on the emotional, psychological and social wellbeing of individuals and communities." (p.8)
Promoting safety	"A trauma-informed approach promotes safety - Establishing a safe physical, psychological and emotional environment where basic needs are met, which recognises the social, interpersonal, personal and environmental dimensions of safety and where safety measures are in place and provider responses are consistent, predictable, and respectful." (p.8)
Supporting consumer control, choice and autonomy	"A trauma-informed approach values and respects the individual, their choices and autonomy, their culture and their values." (p. 8)

Ensuring cultural competence	"A trauma-informed approach understands how cultural context influences perception of and response to traumatic events and the recovery process; respecting diversity; and uses interventions respectful of and specific to cultural backgrounds" (p. 8)
Safe and healing relationships	"A trauma-informed approach fosters healing relationships where disclosures of trauma are possible and are responded to appropriately. It also promotes collaborative, strengths-based practice that values the person's expertise and judgement." (p.8).
Sharing power and governance	"A trauma-informed approach recognises the impact of power and ensures that power is shared." (p. 8)
Recovery is possible	"A trauma-informed approach understands that recovery is possible for everyone regardless of how vulnerable they may appear; instilling hope by providing opportunities for consumer and former consumer involvement at all levels of the system; facilitating peer support; focusing on strength and resiliency; and establishing future-oriented goals." (p. 8)
Integrating care	"A trauma-informed approach maintains a holistic view of consumers and their recovery process; and facilitating communication within and among service providers and systems." (p.8).
Source 8. Guarino, K., Soares, P., Konnath, K., Clervil, R., & Bassuk, E. (2009). *Trauma-Informed Organizational Toolkit*. Center for Mental Health Services, Substance Abuse and Mental Health Services Administration, and the Daniels Fund, the National Child Traumatic Stress Network, and the W.K. Kellogg Foundation. www.homeless.samhsa.gov and www.familyhomelessness.org.	
Understanding trauma and its impact	"Understanding traumatic stress and how it impacts people and recognizing that many behaviors and responses that may be seem ineffective and unhealthy in the present, represent adaptive responses to past traumatic experiences." (p. 17)
Ensuring Cultural Competence	"Understanding how cultural context influences one's perception of and response to traumatic events and the recovery process; respecting diversity within the program, providing opportunities for consumers to engage in cultural rituals, and using interventions respectful of and specific to cultural backgrounds" (p. 17)
Supporting Consumer control, Choice and Autonomy	"Helping consumers regain a sense of control over their daily lives and build competencies that will strengthen their sense of autonomy; keeping consumers well-informed about all aspects of the system, outlining clear

	expectations, providing opportunities for consumers to make daily decisions and participate in the creation of personal goals, and maintaining awareness and respect for basic human rights and freedoms." (p. 17)
Sharing Power and Governance	"Promoting democracy and equalization of the power differentials across the program; sharing power and decision-making across all levels of an organization, whether related to daily decisions or in the review and creation of policies and procedures." (p. 17)
Integrating Care	"Maintaining a holistic view of consumers and their process of healing and facilitating communication within and among service providers and systems." (p. 17)
Healing Happens in Relationships	"Believing that establishing safe, authentic and positive relationships can be corrective and restorative to survivors of trauma." (p.17)
Recovery is Possible	"Understanding that recovery is possible for everyone regardless of how vulnerable they may appear; instilling hope by providing opportunities for consumer and former consumer involvement at all levels of the system, facilitating peer support, focusing on strength and resiliency, and establishing future-oriented goals." (p. 17)
Source 9. Sacred Heart Mission. (2023). *Trauma-informed care*. Sacred Heart Mission. https://www.sacredheartmission.org/about-us/our-service-model-approach/trauma-informed-care/	
Trauma Awareness	"Staff and volunteers are all required to undertake one of three levels of trauma-informed training appropriate to their position in order to recognise trauma symptoms and respond appropriately."
Promote Safety	"As trauma survivors often feel at risk of further trauma, a sense of both physical and emotional safety are important to recovery."
Rebuilding Control	"Trauma is disempowering, as is homelessness. Trauma-informed services offer a predictable environment to allow people to rebuild a sense of efficacy and control over their lives. Predictable and reliable relationships with workers also reinforce healthy boundaries and help-seeking behaviour."
Promote connection	"Social networks play a critical role in promoting resilience and recovery. Ideally, trauma survivors will develop healthy connections with friends, family and significant others."

Focus on strengths and resources	"We support people to identify their own strengths and develop or enhance their personal coping skills. While we acknowledge the challenges people have experienced, we support people to articulate and work toward their hopes for the future."
Maintaining a belief in recovery	"This principle reminds us that people can and do recover from trauma. Conveying hope emphatically requires us to understand the barriers to recovery including lack of financial resources or living in unsafe or chaotic environments."
Source 10. Homes, A., Grandison, G., The Rivers Centre, & NHS Lothian. (2021). *Trauma-Informed Practice: A Toolkit for Scotland*. Scottish Government. https://www.gov.scot/publications/trauma-informed-practice-toolkit-scotland/documents/	
Safety	"Efforts are made throughout the organisation to ensure that staff and the people they serve feel physically and psychologically safe. Staff and clients should experience the setting and the interpersonal interactions taking place within the setting as safe, inviting, and not a risk to their physical or psychological safety."
Trustworthiness	"This principle refers to the degree to which organisational operations and decisions are conducted with transparency, with the goal of building and maintaining trust among clients and their family members, and among staff and others involved in the organisation."
Choice	"Throughout the organisation, clients and staff are supported to make decisions and choices, and to set their own goals. The organisation recognises that giving people choice can help address power imbalances. Clients and staff therefore have meaningful choice and a voice in the decision-making process of the organisation and its services."
Collaboration	"The organisation recognises the value of staff and clients' experience in overcoming challenges and improving the system as a whole. Attempts are made to level the power differentials between different staff groups, and between staff and clients. This principle is often implemented through the formal or informal use of peer support and mutual self-help. There is recognition that healing takes place in the context of relationships and in the meaningful sharing of power and decision-making ."
Empowerment	"Efforts are made by the organisation to share power and to give clients and staff a strong voice in decision-making, at both individual and organisational levels. Each level of the organisation, including management, operations, service delivery and staff training, is designed to be empowering for both staff and service users. Staff are empowered by

	mechanisms of organisational support, and clients are empowered by services that are person-centred, and based on belief in the resilience of individuals and their ability to heal and recover from trauma."
Source. 11. Wolf, M. R., Green, S. A., Nochajski, T. H., Mendel, W. E., & Kusmaul, N. S. (2014). 'We're Civil Servants': The Status of Trauma-Informed Care in the Community. *Journal of Social Service Research*, *40*(1), 111–120. https://doi.org/10.1080/01488376.2013.845131	
Physical and emotional safety of staff and clients	"Safety can be physical or emotional and generally involves the protection of self or others. It can include where services are offered; time of day that services are offered; security personnel available, open doors or locked, and the affect that each has on consumers; and the waiting room appearance" (p. 114)
Trustworthiness for staff and clients	"With respect to Trustworthiness, it can mean transparency and honesty, ensuring consistency and appropriate boundaries, and clear task delivery. It includes providing clear information about what will be done, by whom, when, why, and under what circumstances, and respectful and professional boundaries." (p. 115)
Choice and control for clients and staff	"Choice can mean the right to self-determination. It can include how much choice consumers have over the ser-vices they receive (such as time of day, gender preferences for service providers, etc.); are consumers provided a clear and appropriate message about their rights and responsibilities?" (p. 116)
Collaboration between staff and clients, and between staff and management	"COLLABORATION means working together toward a common goal. Collaboration can include giving consumers a significant role in planning and evaluating services; possibly giving them preferences in areas of service planning, goal setting, and developing treatment priorities; cultivating an atmosphere of doing "with" rather than doing "to" or "for"; and conveying the message that the consumer is the expert in their own life." (p. 117)
Client and staff empowerment	"EMPOWERMENT is the development and enhancement of skill sets. This includes recognizing consumer strengths and skills; building a realistic sense of hope for the client's future; and providing an atmosphere that allows consumers to feel validated and affirmed with every contact at the agency." (p. 117)
Source 12: Fallot, R., & Harris, M. (2009). *Creating Cultures of Trauma-Informed Care (CCTIC): A Self-Assessment and Planning Protocol*. Community Connections. https://children.wi.gov/Documents/CCTICSelf-AssessmentandPlanningProtocol0709.pdf	

Safety – ensuring physical and emotional safety	"To what extent do the program's activities and settings ensure the physical and emotional safety of consumers and staff? How can services be modified to ensure this safety more effectively and consistently?" (p.9) (NB. unpacks into nice clear subquestions).
Trustworthiness – maximizing trustworthiness through task clarity, consistency, and interpersonal boundaries	"To what extent do the program's activities and settings maximize trustworthiness by making the tasks involved in service delivery clear, by ensuring consistency in practice, and by maintaining boundaries that are appropriate to the program? How can services be modified to ensure that tasks and boundaries are established and maintained clearly and appropriately?" (p. 10)
Choice – maximising consumer choice	"To what extent do the program's activities and settings maximize consumer experiences of choice and control? How can services be modified to ensure that consumer experiences of choice and control are maximized?" (p. 11)
Collaboration – maximizing collaboration and sharing power	"To what extent do the program's activities and settings maximize collaboration and sharing of power between staff and consumers? How can services be modified to ensure that collaboration and power-sharing are maximized?" (p. 13)
Empowerment – prioritizing empowerment and skill-building	"To what extent do the program's activities and settings prioritize consumer empowerment and skill-building? How can services be modified to ensure that experiences of empowerment and the development or enhancement of consumer skills are maximized?" (p. 14)

Appendix C **Review of Principles of Ethical Algorithms**

Table 5. Algorithmic-Auditing Principles

Source 1. Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (100–1; pp. 100–101). National Institute of Standards and Technology (U.S.). https://doi.org/10.6028/NIST.AI.100-1

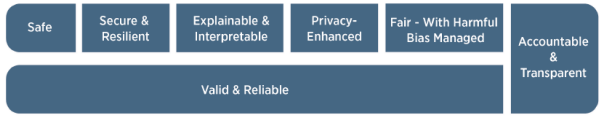

Fig. 4. Characteristics of trustworthy AI systems. Valid & Reliable is a necessary condition of trustworthiness and is shown as the base for other trustworthiness characteristics. Accountable & Transparent is shown as a vertical box because it relates to all other characteristics.

Principles	definition
Valid & Reliable	""Validation is the "confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled." Deployment of AI systems which are inaccurate, unreliable, or poorly generalized to data and settings beyond their training creates and increases negative AI risks and reduces trustworthiness. – p. 13 "Reliability is defined in the same standard as the "ability of an item to perform as required, without failure, for a given time interval, under given conditions" (Source: ISO/IEC TS 5723:2022). Reliability is a goal for overall correctness of AI system operation under the conditions of expected use and over a given period of time, including the entire lifetime of the system." P. 13 "Accuracy and robustness contribute to the validity and trustworthiness of AI systems, and can be in tension with one another in AI systems." – p.14 "Accuracy is defined by ISO/IEC TS 5723:2022 as "closeness of results of observations, computations, or estimates to the true values or the values accepted as being true." Measures of accuracy should consider

	computational-centric measures (e.g., false positive and false negative rates), human-AI teaming, and demonstrate external validity (generalizable beyond the training conditions). Accuracy measurements should always be paired with clearly defined and realistic test sets – that are representative of conditions of expected use – and details about test methodology; these should be included in associated documentation. Accuracy measurements may include disaggregation of results for different data segments." – p. 14 "Robustness or generalizability is defined as the "ability of a system to maintain its level of performance under a variety of circumstances" (Source: ISO/IEC TS 5723:2022). Robustness is a goal for appropriate system functionality in a broad set of conditions and circumstances, including uses of AI systems not initially anticipated. Robustness requires not only that the system perform exactly as it does under expected uses, but also that it should perform in ways that minimize potential harms to people if it is operating in an unexpected setting." – p.14
Safe	"AI systems should "not under defined conditions, lead to a state in which human life, health, property, or the environment is endangered" (Source: ISO/IEC TS 5723:2022). Safe operation of AI systems is improved through: • responsible design, development, and deployment practices; • clear information to deployers on responsible use of the system; • responsible decision-making by deployers and end users; and • explanations and documentation of risks based on empirical evidence of incidents." – p.14
Secure & Resilient	"AI systems, as well as the ecosystems in which they are deployed, may be said to be resilient if they can withstand unexpected adverse events or unexpected changes in their environment or use – or if they can maintain their functions and structure in the face of internal and external change and degrade safely and gracefully when this is necessary (Adapted from: ISO/IEC TS 5723:2022). Common security concerns relate to adversarial examples, data poisoning, and the exfiltration of models, training data, or other intellectual property through AI system endpoints. AI systems that can maintain confidentiality, integrity, and availability through protection mechanisms that prevent unauthorized access and use may be said to be secure. Guidelines in the NIST Cybersecurity Framework and Risk Management Framework are among those which are applicable here." – p.15

	"Security and resilience are related but distinct characteristics. While resilience is the ability to return to normal function after an unexpected adverse event, security includes resilience but also encompasses protocols to avoid, protect against, respond to, or recover from attacks. Resilience relates to robustness and goes beyond the provenance of the data to encompass unexpected or adversarial use (or abuse or misuse) of the model or data." – p.15
Accountable & Transparent	"Trustworthy AI depends upon accountability. Accountability presupposes transparency. Transparency reflects the extent to which information about an AI system and its outputs is available to individuals interacting with such a system – regardless of whether they are even aware that they are doing so. Meaningful transparency provides access to appropriate levels of information based on the stage of the AI lifecycle and tailored to the role or knowledge of AI actors or individuals interacting with or using the AI system. By promoting higher levels of understanding, transparency increases confidence in the AI system." – p.15 "This characteristic's scope spans from design decisions and training data to model training, the structure of the model, its intended use cases, and how and when deployment, post-deployment, or end user decisions were made and by whom. Transparency is often necessary for actionable redress related to AI system outputs that are incorrect or otherwise lead to negative impacts. Transparency should consider human-AI interaction: for example, how a human operator or user is notified when a potential or actual adverse outcome caused by an AI system is detected. A transparent system is not necessarily an accurate, privacy-enhanced, secure, or fair system. However, it is difficult to determine whether an opaque system possesses such characteristics, and to do so over time as complex systems evolve." – p.15-16
Explainable & Interpretable	"Explainability refers to a representation of the mechanisms underlying AI systems' operation, whereas interpretability refers to the meaning of AI systems' output in the context of their designed functional purposes. Together, explainability and interpretability assist those operating or overseeing an AI system, as well as users of an AI system, to gain deeper insights into the functionality and trustworthiness of the system, including its outputs. The underlying assumption is that perceptions of negative risk stem from a lack of ability to make sense of, or contextualize, system output appropriately. Explainable and interpretable AI systems offer information that will help end users understand the purposes and potential impact of an AI system." – p. 16

	"Transparency, explainability, and interpretability are distinct characteristics that support each other. Transparency can answer the question of "what happened" in the system. Explainability can answer the question of "how" a decision was made in the system. Interpretability can answer the question of "why" a decision was made by the system and its meaning or context to the user." – p. 17
Privacy-Enhanced	"Privacy refers generally to the norms and practices that help to safeguard human autonomy, identity, and dignity. These norms and practices typically address freedom from intrusion, limiting observation, or individuals' agency to consent to disclosure or control of facets of their identities (e.g., body, data, reputation)." p.17
Fair – with Harmful bias managed	"Fairness in AI includes concerns for equality and equity by addressing issues such as harmful bias and discrimination. Standards of fairness can be complex and difficult to define because perceptions of fairness differ among cultures and may shift depending on application. Organizations' risk management efforts will be enhanced by recognizing and considering these differences. Systems in which harmful biases are mitigated are not necessarily fair. For example, systems in which predictions are somewhat balanced across demographic groups may still be inaccessible to individuals with disabilities or affected by the digital divide or may exacerbate existing disparities or systemic biases." – p.17 "Bias is broader than demographic balance and data representativeness. NIST has identified three major categories of AI bias to be considered and managed: systemic, computational and statistical, and human-cognitive. Each of these can occur in the absence of prejudice, partiality, or discriminatory intent. Systemic bias can be present in AI datasets, the organizational norms, practices, and processes across the AI lifecycle, and the broader society that uses AI systems. Computational and statistical biases can be present in AI datasets and algorithmic processes, and often stem from systematic errors due to non-representative samples. Human-cognitive biases relate to how an individual or group perceives AI system information to make a decision or fill in missing information, or how humans think about purposes and functions of an AI system. Human-cognitive biases are omnipresent in decision-making processes across the AI lifecycle and system use, including the design, implementation, operation, and maintenance of AI." – p.17
Source 2. Netherlands Court of Audit. (2021). *Understanding algorithms—2021*.	
Respect for human autonomy	"The decisions made by the algorithm are open to human checks." – p. 62 (1.1)

The prevention of harm	"The algorithm is safe and always does what it is supposed to do."-p.62 (2.1)
	"Privacy is safeguarded and data protected."-p.62 (2.2)
Fairness	"Fairness means that the algorithm takes account of population diversity and does not discriminate." – p. 62 (3.1) "The algorithm's impact on society and the environment was taken into account during its development." – p.62 (3.2)
Explainability and transparency	"It is possible to explain which procedures have been followed." – p.62 (4.1) "It is possible to explain how the algorithm works." – p.62 (4.2)
Source 3. NSW Government. (2021). Artificial intelligence assurance framework. https://www.digital.nsw.gov.au/sites/default/files/2022-09/nsw-government-assurance-framework.pdf	
Community Benefit	AI should deliver the best outcome for the citizen, and key insights into decisionmaking
Fairness	Use of AI will include safeguards to manage data bias or data quality risks, following best practice and Australian Standards
Privacy and Security	AI will include the highest levels of assurance
Transparency	Review mechanisms will ensure citizens can question and challenge AI based outcomes
Accountability	Decision-making remains the responsibility of organisations and Responsible Officers.
Source 4. OECD. (2023). Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449).	
Inclusive growth, sustainable development and well-being	This Principle highlights the potential for trustworthy AI to contribute to overall growth and prosperity for all – individuals, society, and planet – and advance global development objectives. Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being.
Human-centred values and fairness	AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and should include appropriate safeguards to ensure a fair and just society.

	AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights. To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.
Transparency and explainability	This principle is about transparency and responsible disclosure around AI systems to ensure that people understand when they are engaging with them and can challenge outcomes. AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art: • to foster a general understanding of AI systems, • to make stakeholders aware of their interactions with AI systems, including in the workplace, • to enable those affected by an AI system to understand the outcome, and, • to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.
Robustness, security and safety	AI systems must function in a robust, secure and safe way throughout their lifetimes, and potential risks should be continually assessed and managed. AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk. To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art.

	AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias.
Accountability	Organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the OECD's values-based principles for AI. AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art.
Source 5. UNESCO. (2022). *Recommendation on the Ethics of Artificial Intelligence* (SHS/BIO/PI/2021/1; p. 43).	
Proportionality and Do No Harm	The use of AI systems must not go beyond what is necessary to achieve a legitimate aim. Risk assessment should be used to prevent harms which may result from such uses.
Safety and Security	Unwanted harms (safety risks) as well as vulnerabilities to attack (security risks) should be avoided and addressed by AI actors.
Right to Privacy and Data Protection	Privacy must be protected and promoted throughout the AI lifecycle. Adequate data protection frameworks should also be established.
Multi-stakeholder and Adaptive Governance & Collaboration	International law & national sovereignty must be respected in the use of data. Additionally, participation of diverse stakeholders is necessary for inclusive approaches to AI governance.
Responsibility and Accountability	AI systems should be auditable and traceable. There should be oversight, impact assessment, audit and due diligence mechanisms in place to avoid conflicts with human rights norms and threats to environmental wellbeing
Transparency and Explainability	The ethical deployment of AI systems depends on their transparency & explainability (T&E). The level of T&E should be appropriate to the context, as there may be tensions between T&E and other principles such as privacy, safety and security
Human Oversight and Determination	Member States should ensure that AI systems do not displace ultimate human responsibility and accountability.
Sustainability	AI technologies should be assessed against their impacts on 'sustainability', understood as a set of constantly evolving goals including those set out in the UN's Sustainable Development Goals.

Awareness & Literacy	Public understanding of AI and data should be promoted through open & accessible education, civic engagement, digital skills & AI ethics training, media & information literacy.
Fairness and Non-Discrimation	AI actors should promote social justice, fairness, and non-discrimination while taking an inclusive approach to ensure AI's benefits are accessible to all.
Source 6. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI* (SSRN Scholarly Paper 3518482). https://doi.org/10.2139/ssrn.3518482	
Privacy	Principles under this theme stand for the idea that AI systems should respect individuals' privacy, both in the use of data for the development of technological systems and by providing impacted people with agency over their data and decisions made with it. Privacy principles are present in 97% of documents in the dataset.
Accountability	This theme includes principles concerning the importance of mechanisms to ensure that accountability for the impacts of AI systems is appropriately distributed, and that adequate remedies are provided. Accountability principles are present in 97% of documents in the dataset
Safety and Security	These principles express requirements that AI systems be safe, performing as intended, and also secure, resistant to being compromised by unauthorized parties. Safety and Security principles are present in 81% of documents in the dataset.
Transparency and Explainability	Principles under this theme articulate requirements that AI systems be designed and implemented to allow for oversight, including through translation of their operations into intelligible outputs and the provision of information about where, when, and how they are being used. Transparency and Explainability principles are present in 94% of documents in the dataset.
Fairness and non-discrimination	With concerns about AI bias already impacting individuals globally, Fairness and Non-discrimination principles call for AI systems to be designed and used to maximize fairness and promote inclusivity incuding in impact and design of the systems. Fairness and Non-discrimination principles are present in 100% of documents in the dataset.
Human control over technology	The principles under this theme require that important decisions remain subject to human review. Human Control of Technology principles are present in 69% of documents in the dataset.
Professional responsibility	These principles recognize the vital role that individuals involved in the development and deployment of AI systems play in the systems' impacts, and call on their professionalism and integrity in ensuring that the appropriate stakeholders are consulted and long-term effects are planned for. Professional Responsibility principles are present in 78% of documents in the dataset.

Promotion of human values	Finally, Human Values principles state that the ends to which AI is devoted, and the means by which it is implemented, should correspond with our core values and generally promote humanity's well-being. Promotion of Human Values principles are present in 69% of documents in the dataset
Source 7: Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), Article 9. https://doi.org/10.1038/s42256-019-0088-2	
Justice, Fairness and Equity	Justice is mainly expressed in terms of fairness, and of prevention, monitoring or mitigation of unwanted bias and discrimination. Whereas some sources focus on justice as respect for diversity, inclusion and equality, others call for a possibility to appeal or challenge decisions, or the right to redress and remedy.
Non-maleficence	References to non-maleficence occur significantly more often than references to beneficence and encompass general calls for safety and security or state that AI should never cause foreseeable or unintentional harm. More granular considerations entail the avoidance of specific risks or potential harms—for example, intentional misuse via cyberwarfare and malicious hacking—and suggest risk-management strategies. Harm is primarily interpreted as discrimination , violation of privacy or bodily harm. Less frequent characterizations include loss of trust or skills; "radical individualism"; the risk that technological progress might outpace regulatory measures; and negative impacts on long-term social well-being, infrastructure, or psychological, emotional or economic aspects.
Responsibility and Accountability	Nonetheless, specific recommendations include acting with "integrity" and clarifying the attribution of responsibility and legal liability, if possible upfront, in contracts or, alternatively, by centring on remedy. In contrast, other sources suggest focusing on the underlying reasons and processes that may lead to potential harm. Yet others underline the responsibility of whistleblowing in case of potential harm, and aim at promoting diversity or introducing ethics into science, technology, engineering and mathematics education.
Privacy	Ethical AI sees privacy both as a value to uphold and as a right to be protected. While often undefined, privacy is frequently presented in relation to data protection and data security.
Beneficence	While promoting good ('beneficence' in ethical terms) is often mentioned, it is rarely defined, though notable exceptions mention the augmentation of human senses, the promotion of human well-being and flourishing, peace and happiness, the creation of socio-economic opportunities, and economic prosperity
Freedom and Autonomy	Whereas some sources specifically refer to the freedom of expression or informational self-determination and "privacy-protecting user controls", others generally promote freedom, empowerment or autonomy.

	Some documents refer to autonomy as a positive freedom, specifically the freedom to flourish, to self-determination through democratic means, the right to establish and develop relationships with other human beings, the freedom to withdraw consent, or the freedom to use a preferred platform or technology. Other documents focus on negative freedom—for example, freedom from technological experimentation, manipulation or surveillance. Freedom and autonomy are believed to be promoted through transparency and predictable AI, by not "reducing options for and knowledge of citizens", by actively increasing people's knowledge about AI, giving notice and consent or, conversely, by actively refraining from collecting and spreading data in absence of informed consent.
Trust	Suggestions for building or sustaining trust include education, reliability, accountability, processes to monitor and evaluate the integrity of AI systems over time, and tools and techniques ensuring compliance with norms and standards. Whereas some guidelines require AI to be transparent, understandable or explainable in order to build trust, another one explicitly suggests that, instead of demanding understandability, it should be ensured that AI fulfils public expectations. Other reported facilitators of trust include "a Certificate of Fairness", multi-stakeholder dialogue, awareness about the value of using personal data, and avoiding harm.
Sustainability	To the extent that is referenced, sustainability calls for development and deployment of AI to consider protecting the environment, improving the planet's ecosystem and biodiversity, contributing to fairer and more equal societies and promoting peace.
Dignity	While dignity remains undefined in existing guidelines, save one specification that it is a prerogative of humans but not robots, there is frequent reference to what it entails: dignity is intertwined with human rights or otherwise means avoiding harm, forced acceptance, automated classification and unknown human–AI interaction. It is argued that AI should not diminish or destroy, but respect, preserve or even increase human dignity
Solidarity	Solidarity is mostly referenced in relation to the implications of AI for the labour market. Sources call for a strong social safety net. They underline the need for redistributing the benefits of AI in order not to threaten social cohesion and respecting potentially vulnerable persons and groups

ARC Centre of Excellence for Automated Decision-Making and Society

admscentre.org.au

adms@rmit.edu.au

Professor Paul Henman

T +61 7 3365 2383
E p.henman@uq.edu.au
W uq.edu.au

CRICOS Provider 00025B • TEQSA PRV12080