Developing a Framework for Trustworthy AI-Supported Knowledge Management in the Governance of Risk and Change

Rebecca Vining[1,*[0000-0002-2715-8129]], Nick McDonald[1], Lucy McKenna[2[0000-0002-6035-7656]], Marie E Ward[1,3[0000-0002-6638-8461]], Brian Doyle[1,4[0000-0002-9106-9526]], Junli Liang[2[0000-1111-2222-3333]], Julio Hernandez[2[0000-0003-1347-9631]], John Guilfoyle[4], Arwa Shuhaiber[5], Una Geary[3], Mary Fogarty[3], Rob Brennan[2[0000-0001-8236-362X]]

[1] Centre for Innovative Human Systems, School of Psychology, Trinity College, The University of Dublin, D02 PN40 Dublin, Ireland
[2] ADAPT Centre, School of Computing, Dublin City University, D09 PX21 Dublin, Ireland
[3] Quality and Safety Improvement Directorate, St James's Hospital Dublin, D08 NHY1 Dublin, Ireland
[4] Health and Safety Unit, Dublin Fire Brigade, D02 RY99 Dublin, Ireland
[5] Beacon Renal, Sandyford Business Park, D18 TH56 Dublin, Ireland
*Corresponding author
rvining@tcd.ie

Abstract. This paper proposes a framework for developing a trustworthy artificial intelligence (AI) supported knowledge management system (KMS) by integrating existing approaches to trustworthy AI, trust in data, and trust in organisations. We argue that improvement in three core dimensions (data governance, validation of evidence, and reciprocal obligation to act) will lead to the development of trust in the three domains of the data, the AI technology, and the organisation. The framework was informed by a case study implementing the Access-Risk-Knowledge (ARK) platform for mindful risk governance across three collaborating healthcare organisations. Subsequently, the framework was applied within each organisation with the aim of measuring trust to this point and generating objectives for future ARK platform development. The resulting discussion of ARK and the framework has implications for the development of KMSs, the development of trustworthy AI, and the management of risk and change in complex socio-technical systems.

Keywords: Access-Risk-Knowledge (ARK), socio-technical systems analysis, risk governance, artificial intelligence, trust.

1 Introduction

Safety regulation increasingly calls for a strategy that goes beyond compliance to being proactive, predictive, and preventive [1]. Under such a strategy, effective organisational risk governance relies on evidence-based knowledge, which can be leveraged in support of actions to mitigate risk. A sophisticated knowledge management system (KMS) is needed to oversee this mechanism. While many organisations, particularly in high-risk

domains, are generating large amounts of data from diverse sources, the challenge for risk and safety management is to base operational and strategic decision-making on a coherent, integrated body of data and evidence (knowledge). Our work develops such a system through the case study of deploying an artificial-intelligence (AI)-based software platform that manages risk among three healthcare organisations. There is an ethical obligation to build trustworthiness into AI technology [2], but this obligation must be extended to incorporate issues of trustworthiness in complex socio-technical systems (STS). This paper explores how this extension can be achieved, integrating strategies for building trust in AI and in organisations in order to develop a framework for trustworthy AI-supported knowledge management. This suggests two research questions:

(1) What are the components of a trustworthy AI-supported KMS?
(2) How can these components be achieved in the development and deployment of a software platform for mindful risk governance?

The Access-Risk-Knowledge (ARK) Platform [3] is a software platform that supports the management of risk and change in complex operational systems. The platform deploys the Cube framework for socio-technical systems analysis (STSA) [4-8] along with a risk register, an evidence service, risk mitigation project management tools, analytics, and reports. Risk assessments can be imported from an existing risk register or completed within the platform and are then linked to safety projects. These features enable what we define as mindful governance of risk by leveraging human- and machine-based knowledge to analyse causal relationships. The result of a completed ARK project is an evidence-based analysis of a risk mitigation project throughout the full project management cycle. Projects can also be interlinked in order to synthesise results or to compare results, evidence, or domains. Results can be disseminated to the organisation using the customisable report generation feature.

ARK-Virus is a collaborative project between an academic research team from both the computer science and organisational psychology disciplines, as well as a Community of Practice (CoP) involving quality and safety staff from a 1000-bed urban academic teaching hospital, medical staff from a private renal dialysis service, and management staff from a large urban fire and emergency medical services (EMS) provider. The aim of the project is to develop the ARK platform via a use case relating to infection prevention and control (IPC) in each of the three participating organisations. There are four ARK platform development trials planned; at the time of writing, the project is between the third and fourth of these. A fuller description of each trial and the research activities involved is outlined in a previous paper [7]; the focus of this paper is to develop a framework for trustworthy AI-supported knowledge management, which spans all four of the trials.

In earlier stages of ARK-Virus, our research focused on issues relating to usability, but trust has become increasingly important. Discussions with users centred around a key set of issues: how to make sense of the data, how to do something useful with it, and how to generate a sound basis for engaging others within the organisation. Trust in the platform's ability to deliver this may be a key mechanism for understanding the relationships between the ARK platform, knowledge, users, and the organisation. In this paper, we draw upon the literature on trust in organisations, AI technology, and

data, and upon several decades of research on risk in aviation and healthcare, to outline a framework for the development of a trustworthy KMS that is supported by AI technologies. As the ARK-virus project continues, we aim to apply this framework so that trust can be built into future platform development stages.

Our work is situated at the intersection of technology and people, and there is a clear link between trust in these two domains. Building trustworthy AI involves the full organisational context of implementation, while building trust in the organisation similarly requires taking into account the role of technology supported knowledge as evidence as a rational basis for action. The convergence of knowledge between technology and people inevitably means that technology-based knowledge is a critical resource for human decision-making, as it can generate leverage to address complex problems. Risk and safety management must be based on data and evidence that is integrated into operational decision making. As trust is core to the management of safety and the implementation of change, a unified view of trust that bridges risk management and trust in AI is needed.

Our model of trust incorporates existing theories of organisational trust [9,10], governance of risk [11], and data governance [12]. Drawing upon several decades of research, dialogue with collaborators, and the literature, three core dimensions of trust were identified: data governance, validation of evidence, and reciprocal obligation to act. By supporting improvements in these three dimensions, trust is improved at the level of trust in the organisation [10,13], trust in the AI technology [2], and trust in the data [14]. The framework is outlined in Figure 1.

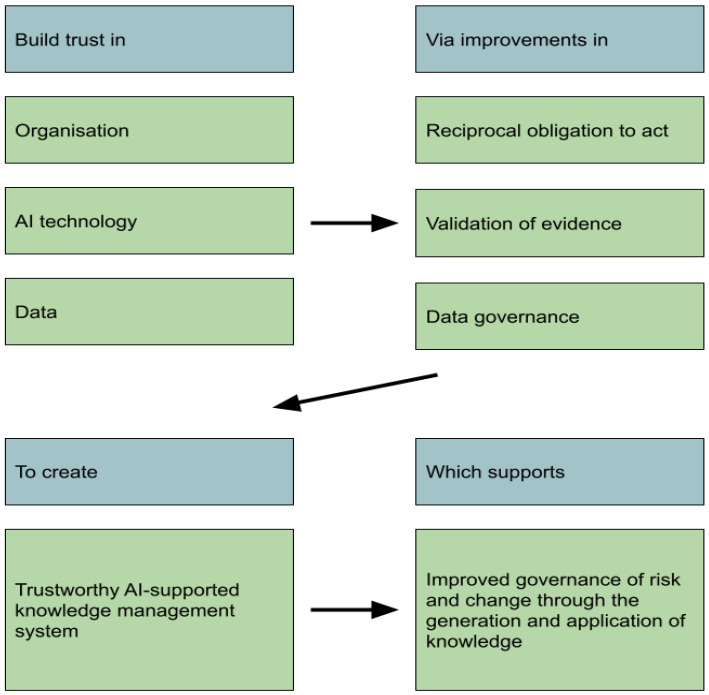

Fig. 1. Framework for developing a trustworthy AI-supported KMS for risk governance.

The ARK platform instantiates this model to support human-directed decision-making and implementation as part of an accountable governance framework. **Data governance** is at the core of ARK's services. **Validation of evidence** is the core activity of STSA Cube analysis, deploying the flexible schemata of Knowledge Graphs to bring together diverse data sources to support analysis, decision-making and project management by quality and safety experts. **The reciprocal obligation to act** is engendered by the mindful governance of a risk project from problem state to verified outcome.

In this paper, five stages are outlined in the development of trust in such a system. The trust model is used to analyse and assess the ARK platform's deployment within each collaborating organisation. Over the course of the previous ARK-Virus trials, trust has been developed through a variety of strategies in each organisation. Using the model of trust as an explanatory concept in this way provides a set of objectives for future development of the project. This suggests the possibility of a capability maturity model (CMM) to provide guidance in development of trustworthy governance of system risk based on verifiable outcomes to demonstrate the effective mitigation of system risk.

2 A Framework for Trustworthy AI-Supported Knowledge Management

Trust has been defined in the literature on trustworthy AI as "(1) a set of specific beliefs dealing with benevolence, competence, integrity, and predictability (trusting beliefs); (2) the willingness of one party to depend on another in a risky situation (trusting intention); or (3) the combination of these elements" [15]. The European Union Ethical Principles for Trustworthy AI [2] outline a set of seven requirements for trustworthiness; our work supports and extends these principles by integrating trustworthy AI, trust in data, and trust in organisations.

Mollering offers a model that helps us build an understanding of the problems with trust relating to our work [9]. Trust is defined as a strategy to cope with the complexity, uncertainty, and risk in the world at large; the necessity to assume a level of certainty projected to the future is based on a combination of reason, routine, and reflexivity. Keymolen applies this model to analyse the relation between trust in other individuals, trust in an organisation and trust in technology [16]. Ward, through a series of case studies in an aviation organisation, illustrates the dynamic nature of the factors that combine to develop trust in an organisational context: understanding and sharing common goals; open communication of information and knowledge; building relationships in resolving conflicts in the process of work; together reviewing and adjusting work-as-imagined based on how work actually happens; it also implies a belief in the future and establishes the basis for future action [10].

The three components of Mollering's model provide a powerful framework to analyse the nature of trust in a data-rich organisational system that is dedicated to

managing risk (achieving certain outcomes) through the deployment of diverse dedicated roles and relationships. Figure 2 illustrates the connections between that model and the trust dimensions identified in our work.

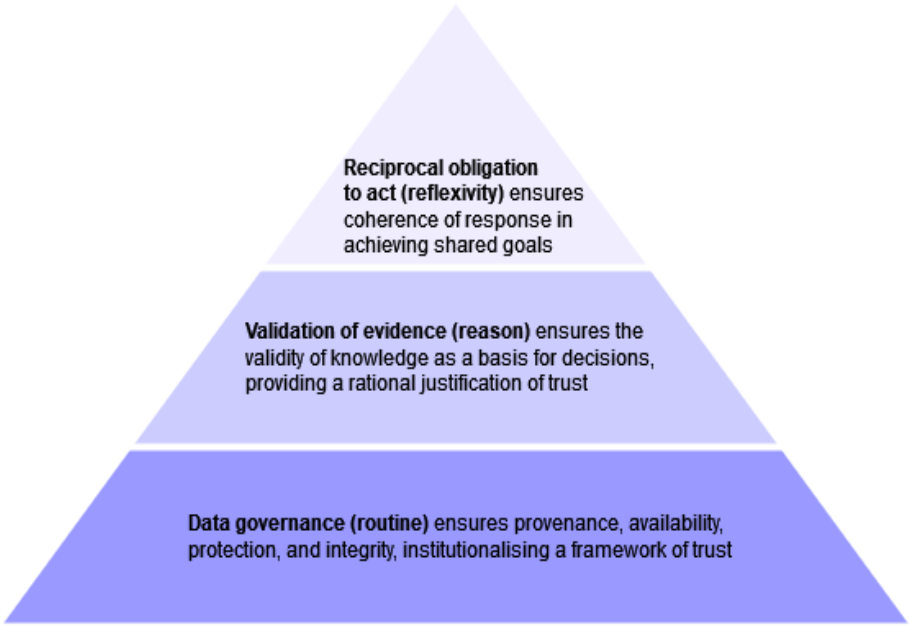

Fig. 2. Mollering's triad and the core trust dimensions.

At a basic level there are three objects of a trusting relationship (trust domains):

- Trust in the data itself.
- Trust in the processing or transformation of data into usable information and knowledge (Trustworthy AI).
- Trust in the sharing of knowledge with colleagues and building trusting relationships, leading in turn to trust in the organisational processes that deploy and use that information and knowledge.

Trustworthy data governance ensures high-quality data and efficient, effective use of the data, thus leading to more meaningful and trustworthy evidence. Validation of that evidence in turn links data governance to reciprocal obligation to act by linking cause to effect. In turn, the obligation to act drives a need for continued collection of high-quality, trustworthy data. What results is a cyclical pathway driving continuous improvement of trust in the KMS.

Table 1 illustrates from a theoretical perspective how each dimension (data governance, validation of evidence and obligation to act) builds trust in each domain (data, AI technology, organisation), explaining the key mechanism by which improvements in the core dimensions will result in the development of trust in each domain. Each column in Table 1 represents the impacts of improvements in that dimension on each of the three trust domains (i.e., the column labelled 'Data

Governance' describes how good data governance improves trust in data, trust in AI technology, and trust in the organisation). The row labelled 'AI Technology' draws directly on the requirements set forward in the European Union Ethical Principles [2].

Table 1. Impacts of Core Dimensions on Trust Domains

Domain	Data Governance	Validation of Evidence	Reciprocal Obligation
Data	Ensures data quality and efficient/effective use.	Generates trust-related metadata and trustworthy data as an outcome of cause and effect.	Action based on data validates the data based on action outcomes - if the outcome works, it increases the confidence in the data.
AI Tech.	Supports human agency and oversight, privacy and data governance, transparency, accountability, diversity and fairness, and technical robustness and safety.	Ensures human agency and oversight, transparency, diversity and fairness, societal and environmental wellbeing, and technical robustness and safety.	Sustains human agency and oversight, accountability.
Organisation	Leads organisational decisions to be data-driven and ensures data decisions are aligned with organisational goals.	Ensures that data-driven decisions are grounded in causal relations.	Sustains coherent response throughout the project cycle, including stakeholder feedback.

3 The ARK (Access-Risk-Knowledge) Platform and Trust

ARK (Figure 3) is a software platform that builds and maintains a Resource Description Framework (RDF)-based unified knowledge graph [17] of risks and projects to link available datasets on practices, risks, and evidence. This bridges traditional qualitative risk evidence and quantitative operational or analytics data, which in turn makes large-scale evidence collection and risk analysis more tractable. Through ARK, human-oriented quantitative risk information is transformed into structured, machine-readable data suitable for automated analysis, querying, and reasoning. A privacy by design approach is taken and data governance principles are followed to ensure support for evidence linkage, classification, and search. The ARK platform is designed to support human-directed decision-making and implementation as part of an accountable governance framework. Data governance, data protection and confidentiality are key features of the design. Knowledge graphs are a natural way to bring together such diverse data sources due to their flexible schemata and through use of uplift to common ontologies, ontology alignment techniques, Natural Language Processing (NLP)-based knowledge extraction and metadata-based integration, e.g., data catalogues.

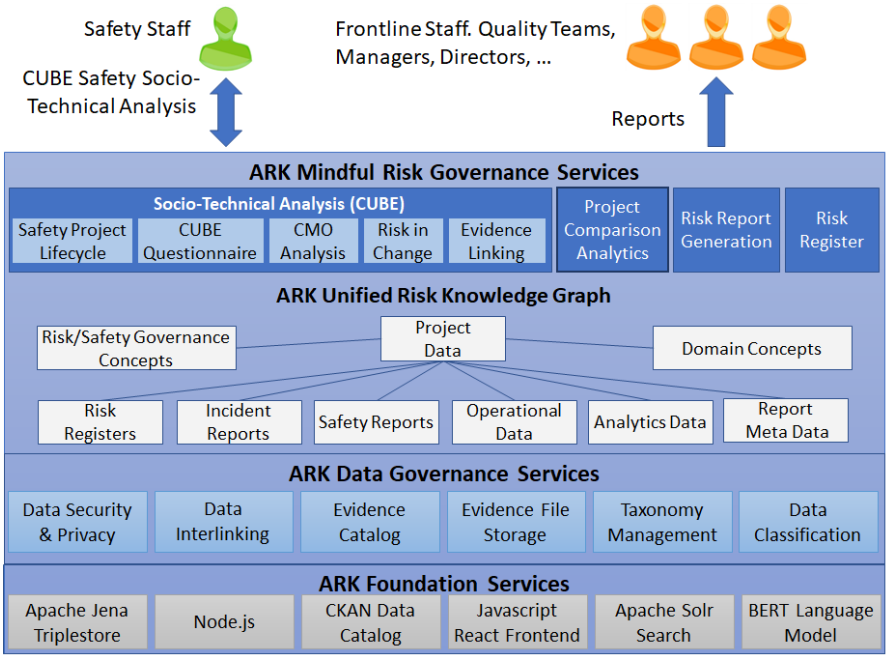

Fig. 3. The ARK Platform risk governance services, risk knowledge graph, data governance services, and foundation services.

ARK supports the development of trust via the key pathway of leveraging data to create knowledge in support of action by embedding the trust dimensions as described below.

Data governance is at the core of ARK's services since it supports Khatri and Brown's data governance decision domains of data principles, data quality, lifecycle, metadata, and access [18] to manage projects, evidence, and risk. The Comprehensive Knowledge Archive Network (CKAN) data catalogue is used to build the ARK Evidence Service. This enables collecting and tracking of extensive metadata on all evidence, relating to provenance, verifiability, reputation, and licensing. Within the Cube knowledge graph, World Wide Web Consortium (W3C) standards for provenance, classification, identity and access control [19] have been used to capture this metadata on all data entities within the graph and a flexible policy-driven, General Data Protection Regulation (GDPR) enabled, context-aware access control system has been implemented to enable federated data sharing within and between organisations [20].

Validation of evidence is the core activity of STSA, where quality and safety experts use ARK to perform a structured analysis of risks and safety projects linking them to a wider range of data sources to support synthesis (with operational data) to give evidence-based assessment of risk and create new knowledge via that synthesis. The structured user interface of ARK exposes multiple views of an underlying ontology that unifies the analysis and enables the combination of traditional qualitative textual analysis fields with structured data in the form of evidence datasets, risk, and domain

classification taxonomies. A natural language processing component based on the BERT language model [21] suggests appropriate taxonomy terms and these are approved by the human expert. Uploading of new evidence (as opposed to evidence linking) is an access-controlled activity and only users with sufficient permissions can do this, to facilitate manual validation of evidence prior to upload.

The reciprocal obligation to act is made explicit in numerous parts of the ARK platform. Firstly, the platform is arranged around the sequence of project stages through verification of the outcome, which gives information about the outcome as well as how the entire sequence works. The Cube summary, project analysis, reporting and synthesis interfaces all contribute to exposing the importance of the problem, the effectiveness of the solution and the viability of the pathway that underlie the obligation to act. Finally, the use of knowledge graphs and feature for linking multiple projects in hierarchies or more general graphs enable a development of a new level of organisational knowledge, facilitating innovative meta-projects rather than reinforcing what's already known. This understanding is leveraged for effective action, responsibility for which can be distributed explicitly to individuals within the organisation.

4 Stages in the Acquisition of Trust

Analysing progress in the three core dimensions provides an enriched understanding of the evolution of trust in ARK-Virus. Understanding the dimensions and the interactions between them develops trust into an explanatory concept, which can be used to inform a set of development objectives. In Table 2 we have outlined five stages in the acquisition of trust, from neophyte to multiple organisations. In the upcoming phase of the ARK-Virus project, the goal is for each organisation to progress up a stage: Organisation 1 from single projects to multiple; Organisation 2 from neophyte to intermediate; and Organisation 3 from intermediate to single projects. This table offers a way of measuring where each organisation is in the trust development process, which will be useful as a point of comparison in the future and support us in determining the key issues to be addressed at that point in time.

Table 2. Stages in the Acquisition of Trust

Stage	Data governance	Evidence validation	Reciprocal obligation
1. Neophyte	Resolve issues of access and privacy.	Plausible interpretation and evidence gathering.	Initial individual use. Potential for collaboration.
2. Intermediate	Assemble and begin integrating relevant data sources.	Gathers evidence and performs effective analysis.	Engages people in real projects that require collaboration.

3. Single projects	Develop knowledge graphs to generate project-level knowledge. Catalogue data source provenance.	Diverse evidence synthesised & validated as representing process & outcome.	Embedded in tactical organisational processes that provide accountable action and outcome.
4. Multiple projects, organisational level	Link data at the level of multiple projects to further develop knowledge graphs and generate organisation-level knowledge. Assure data quality.	Synthesis of evidence provides a basis for policy.	Engage strategic & operational loops of knowledge lifecycle across & beyond organisation.
5. Multiple organisations, sector level	Fully developed private & public knowledge space, routine transformation of private into public.	Evidence provides a basis for guidance, regulation or publication.	Guidance feeds back into the evidence base.

5 Application of the Trust Model to a Community of Practice

In this exercise, we applied our model of trust to the ARK-Virus project within each of the three participating organisations, asking users to reflect on the ways in which trust had been developed to this point and the next steps for further development. The results of this exercise in each organisation are outlined in the subsections below.

Several commonalities emerged in terms of needs moving forward. Firstly, it was noted that many of the more salient issues for the CoP were related to data governance. For Organisation 1, this was the acquisition of data from different stakeholders within the organisation; for Organisation 2, data privacy and obtaining formal permissions to enter information into the platform; for Organisation 3, the resolution of data complexity and organising data from a large number of different sources. Secondly, there is a clear need across all three organisations to extend the user base to encompass the full range of relevant decision-makers. This expansion improves capacity in all three dimensions, but in particular the reciprocal obligation to act. Thirdly, there is a pragmatic need to gather and disseminate evidence showing that actions from ARK projects lead to good outcomes at the organisational level, thus increasing trust in all three dimensions.

5.1 Organisation 1

Organisation 1 developed a project examining personnel compliance with COVID-19 IPC risk management and control measures. At the onset, the organisational representatives hoped to collect data measuring personnel compliance in rest areas, as these were suspected to be a key source of staff-to-staff COVID-19 transmission. However, this was deemed unfeasible as there was a need to develop trust in the project among personnel before such data could be collected. Instead, data was drawn from

what was available in terms of occupational health data, guidelines and control measures over time, impact of limited personnel availability on service provision, and implicit/explicit knowledge about the linkages between the evidence sources from the organisation's ARK-Virus project team. The ARK platform then enabled the project team to analyse a complex and intractable problem for a full project cycle (from problem to embedment). The structured approach to STSA helped frame the problem and identify possible solutions, which were transposed into an implementable operational solution. The platform was also utilised to effectively communicate and implement the solution and verify the efficacy of the solution. Further projects utilising the ARK platform within the organisation have been initiated, indicating the organisation's trust in the platform.

Data governance: Data Protection (DPA) and Non-Disclosure (NDA) Agreements guaranteed a level of data protection that was acceptable to the organisation. However, access to more granular data remained restricted due to concerns about anonymity of personnel. While there were difficulties in acquiring granular data and evidence, the process of seeking this evidence for use on the platform resulted in the acquisition of knowledge from within the organisation which verified the efficacy of the implemented solution.

Validation of evidence: Gathering of evidence was somewhat restricted due to privacy issues, the organisation's work practises, and the organisation's clinical environment. The evidence gathered was done so utilising a top-down/bottom-up approach, with stakeholders from various departments, including operations, health and safety, and logistics, gathering, interpreting, and validating the uploaded evidence.

Reciprocal obligation to act: Organisation 1 has a fairly strict hierarchical rank structure, with a promotional process that means senior managers have fulfilled operational roles, sometimes alongside personnel they now manage. This structure was felt to enhance the level of social trust across ranks in the organisation, contributing to a peer-driven environment where personnel are amenable to the idea of change based on that trust. Initially, there were three personnel from the organisation who engaged directly with the platform, from middle and senior management and health and safety. However, input was also sought from other areas of the organisation, including operations, resources allocation, health and safety, and senior management. To strengthen reciprocal obligation to act, there is a need to involve these stakeholders more formally, in particular by training more personnel as ARK users.

5.2 Organisation 2

Organisation 2 aimed to assess patient compliance with PPE measures upon arrival. Six months of data on patient compliance were collected by front desk staff, a timeframe which covered two different sets of PPE requirements. There were, however, significant issues with obtaining access to the data, with the DPA and NPA taking nearly a full year to complete. In the meantime, users from the organisation were able to fill out sample projects on the platform and participate fully in the other aspects of the project such as the CoP meetings and workshops. In Trial 3, the goal for Organisation 2 is to move from the first stage of trust to the second. The risk is currently being actively

managed at the local level (clinical frontline), but having overcome data governance barriers, a thorough analysis of evidence will enable the organisation to strengthen its management of that risk.

Data governance: Access to data was granted just one week prior to writing of this paper (the datasets remain within the organisation only, while analysis of the data is accessible to others within the ARK-Virus project). Trust in data governance as it relates purely to data has been heightened through the formalisation of data governance procedures via the DPA and NPA, but there is still much progress to be made in terms of data governance and trust in the AI technology and the organisation.

Validation of evidence: The organisation is at the stage of moving from data collection to analysis and use of the data. Moving forward, the organisation is working to identify variables in the data and complete the STSA component of an ARK project, which will allow for further exploration and validation of the predictors and/or outcomes of PPE compliance.

Reciprocal obligation to act: At this stage, operational staff are the primary user group; an important development will be the engagement of a wider variety of users, particularly in more strategic or risk management roles.

5.3 Organisation 3

Early on in the project, it became apparent that a key issue for Organisation 3 was the vast amount of data being produced and reviewed, with no unified structure for tracking all of the data. Over 100 discrete performance indicators are currently monitored in relation to the actions taken for the prevention and control of healthcare-associated infections (PCHCAI), and the processes for capturing, reporting on, collating, and presenting the data can be fragmented and time consuming. As a result of the organisation's experiences completing an ARK project related to environment hygiene and the wider PCHCAI programme, the organisation conducted a data governance mapping exercise. PCHCAI metrics were mapped along dimensions of data governance including the purpose of the metric; type of metric; basis of metric (numerator and denominator); owner; reporting; tools or platforms used for gathering, analysing and reporting the data; whether it could be considered an outcome, process, structure or balancing measure; and the national and international benchmarks and regulatory basis of the data.

Data governance: Progress was made in terms of data governance processes, addressing the issue of the large amount of data and how to turn it into a more manageable data catalogue that provides a clear rationale for management and use. What remains to be done is to expand and embed the data governance processes so that subsequent actions and outcomes can be obtained and measured. The fact that the platform created a strong rationale for compiling and auditing data is an argument in favour of understanding the entire data system prior to initiating a real-world project; in other words, to avoid prematurely structuring an evidence trail without first having agreement on the purpose of each of the metrics.

Validation of evidence: The organisation is moving from the validation and use of individual data sets to the validation and use of knowledge, which will be undertaken

by quality and safety improvement staff and PCHCAI programme contributors using the STSA components of the ARK platform.

Reciprocal obligation to act: To this point, work on this ARK trial has been situated in the core quality and safety improvement team, with some level of engagement via production of the stakeholder report in the previous trial. The results of this trial will strengthen this engagement, forming the basis for drawing additional stakeholders into a collaborative programme and widening the ARK platform user base. Building interpersonal trust within the local team is the first step to engaging a wider stakeholder group and building an organisational basis for trust (and subsequently organisational obligation to act).

5.4 Capability Maturity in Trust in AI and the Organisation

The idea that there are phases in the development of a trustworthy AI-supported KMS suggests the possibility of a CMM that would provide a framework for verifying progress through these phases and provide guidance in development and application. De Bruin, et al. discuss the development of CMMs and provide a relevant example of a Knowledge Management Capability Assessment metric with progressive stages in the sharing, managing, and improving of knowledge assets [22]. An example from safety management in aviation is the Civil Air Navigation Services Organisation (CANSO) model of excellence in safety management for Air Traffic Control Organisations [23]. For the development of the ARK platform, we need a hybrid combination that spans between the technology, the AI, and the organisation.

Table 3 outlines two phases in the development of the platform: Trials 1 and 2, and Trial 3. Trial 1 and 2 measurements were collected in the earlier phase of the project. Trial 3 trust measurements will be collected in the upcoming phase of the project, as will measurements on platform usability and effectiveness. The strategic requirements for achieving advancements in trust are outlined in the middle column, Trial 3 Strategy. Table 3 represents a synthesis of the first two tables and an initial attempt to define and measure progress at this point in ARK-Virus towards the development of a trustworthy AI-supported KMS.

Table 3. Trust-Related Measurements and Development Strategy

Dimension	Trial 1 and 2 Measurements	Trial 3 Strategy	Trial 3 Measurements
Data governance	ISO27001 Security Assessment	Security (advanced access control policies to enable federated sharing) GDPR compliance (privacy by design, compliance reporting, etc.) Privacy-aware data interlinking mechanisms	Trustworthy data metrics to measure provenance, verifiability, reputation, and licensing [14]

Validation of evidence	Develop and distribute stakeholder report on findings	Analyse and better illustrate quality of causal relations Validate sequence of activity and outcome Meta-analysis of multiple projects to support proposal of new guidance	Develop guidance material based on evidence Initiate new projects based on expectation of outcomes of value
Reciprocal obligation to act	Build internal user groups Propose credible solutions to the identified problem	Represent different user roles in platform Represent relationships between reports and their owners in platform Engage stakeholders within and outside of CoP organisations	Build set of expert users and widen user base Engagement with implementation of guidance material

6 Discussion

In order to move the ARK platform along the pathway from development to implementation to embedment, it is crucial that the technology and the system it engenders are trusted by the participating user organisations. Operationally, the ARK platform is for management of risk and change, which involves analysing the issues to do with causal relationships, outcomes, and changing the outcomes. The key mechanism for changing outcomes is the leveraging of knowledge as evidence. A better understanding of this process can help explain the differential success of change projects, impacting at the level of the organisation, sector, and society.

The ARK-Virus project has been a strong stimulus to organise evidence in the participating organisations. Although so far that collection has not been highly sophisticated in terms of AI, and while there has not been the opportunity for in-depth AI supported analysis, there is confidence that the platform will deliver this in the future. The organisation of evidence is a necessary first step. In addition, this exercise showed that the first step is to build trust at the local level; trust is developed in stages, and overestimating the level of trust already achieved within an organisation should be avoided. Trust was built locally by enhancing relationships with working colleagues at the level of the research team, the CoP, and the user groups from each organisation.

Access to data presented key challenges in terms of project progress across the participating organisations. This highlights a need for updated data governance models that enable effective action, rather than solely protecting privacy, aligning with the work of Janssen, et al. [24]. Inter-organisational trust in data governance practices, in particular with regards to protecting anonymity of personnel, appears to play a role in securing access to data, though legal agreements are also necessary.

The ARK-Virus project is a work in progress. This exercise enabled us to develop a structured framework for examining the stages in development of the project and the ARK platform. Analysing trust has helped us to outline a plan for moving forward in the project in a way that supports the embedment of the platform in existing risk

management processes within the participating organisations and led to the selection of key outcome measures relating to the development of trust, constituting the first step in developing a CMM.

7 Conclusions

In this exercise, we outlined a framework for developing a trustworthy AI-supported KMS. In the proposed model, three key dimensions (data governance, validation of evidence, and reciprocal obligation to act) contribute to improved trust in three domains (organisation, AI technology, and data). There are five stages in the development of trust, against which organisations can measure their progress. We then applied the framework to the ARK-Virus project, which deploys a risk management platform in three participating healthcare organisations. This application resulted in a set of objectives that, when achieved, will improve trust in each organisation, as well as a measurement strategy that can be used to track the development of trust. This suggests the possibility of a CMM to provide guidance in development of trustworthy governance of system risk based on verifiable outcomes to demonstrate the effective mitigation of system risk.

Over the course of the previous ARK-Virus trials, trust has been developed through a variety of strategies in each organisation, including participation in the CoP, active feedback loops, engagement of key stakeholders, comprehensive data protection agreements, and building a better understanding of the data. We aim to continue focusing on trust moving forward by measuring the level of trust and developing trial objectives that specifically support its development. There is currently a high level of trust in the platform and its future deployment, particularly in Organisation 1 as evidenced by their selection of the ARK to support additional projects in the coming months. However, there is room for improvement as well. The most salient issues identified were related to data governance, meaning a focus on this area in the coming months will be key. Core needs also included the expansion of the ARK platform user base and the production of a follow-up stakeholder report which consolidates the evidence for beneficial organisational outcomes as a result of ARK projects. These needs will be addressed in subsequent development trials.

Integration of a technology-based knowledge system has social implications, meaning that beyond trust in data or technology, the organisational dimensions of trust must be considered. At the same time, the role of knowledge and evidence is critical for developing trust in the organisation; it is not merely a question of social relationships or expectations. There is a need for frameworks guiding the development of trust in this holistic way. There is also a need to develop guiding principles for AI implementation that support and extend the European Union principles for ethical AI, in particular focusing on the organisational dimension having to do with implementation, action, and outcome. In this exercise, we have contributed to the resolution of this gap by operationalising Mollering's triad [9] to outline a framework for the development of trust in an AI-supported KMS. While our focus has been on a

system that has formal structures for looking at risk and change, any complex STS would benefit from practical examination of a technology-based KMS in terms of trust.

8 Acknowledgments

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 20/COV/8463 at the ADAPT SFI Research Centre at Dublin City University and Trinity College Dublin. This research was conducted with the financial support of Science Foundation Ireland at ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology at DCU [13/RC/2106_P2]. For the purpose of Open Access, the author has applied CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

1. Hollnagel E., Wears, R.L., and Braithwaite, J. From Safety-I to Safety-II: A White Paper. The Resilient Health Care Net. University of Southern Denmark, University of Florida, USA, and Macquarie University, Australia (2015).
2. European Commission High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy Artificial Intelligence. Brussels, Belgium (2019).
3. Crotti Junior, A., Basereh, M., Abgaz, Y., Liang, J., Duda, N., McDonald, N., and Brennan, R. The ARK platform: enabling risk management through semantic web technologies. In Proceedings of the 11th International Conference on Biomedical Ontologies. Bolzano, Italy. (2020, 17 Sept).
4. Ward, M., McDonald, N., Morrison, R., Gaynor, D., and Nugent, T. A performance improvement case study in aircraft maintenance and its implications for hazard identification. Ergonomics, 53(2), pp. 247-267 (2010).
5. Ulfvengren, P. and Corrigan, S. Development and Implementation of a Safety Management System in a Lean Airline. Cognition, Technology, and Work, 17, pp. 219–236 (2015).
6. Corrigan, S., Kay, A., O'Byrne, K., Slattery, D., Sheehan, S., McDonald, N., Smyth, D., Mealy, K., and Cromie, S.A. Socio-Technical Exploration for Reducing & Mitigating the Risk of Retained Foreign Objects. International Journal of Environmental Research and Public Health, 15(4), 714 (2018).
7. McDonald, N., McKenna, L., Vining, R., Doyle, B., Liang, J., Ward, M.E., Ulfvengren, P., Geary, U., Guilfoyle, J., Shuhaiber, A., Hernandez, J., Fogarty, M., Healy, U., Tallon, C., and Brennan, R. Evaluation of an Access-Risk-Knowledge (ARK) Platform for Governance of Risk and Change in Complex Socio-Technical Systems. International Journal of Environmental Research and Public Health, 18(23), 12572 (2021).
8. Geary, U., Ward, M.E., Callan, V., McDonald, N., and Corrigan, S. A socio-technical systems analysis of the application of RFID-enabled technology to the transport of precious laboratory samples in a large acute teaching hospital. Applied Ergonomics, 102, 103759 (2022).

9. Möllering, G. Trust: Reason, Routine, Reflexivity. Elsevier, Amsterdam, Netherlands (2006).
10. Ward, M. Contributions to human factors from three case studies in aircraft maintenance [Doctoral thesis]. Trinity College Dublin (2006).
11. McDonald, N., Callari, T.C., Baranzini, D., and Mattei, F. A Mindful Governance model for ultra-safe organisations. Safety Science, 120, pp. 753-763 (2019).
12. Brous, P. and Janssen, M. Trusted Decision-Making: Data Governance for Creating Trust in Data Science Decision Outcomes. Administrative Sciences, 10(4), 81 (2020).
13. Blomqvist, K. and Stahle, P. Building Organizational Trust. In: 16th ANNUAL IMP-CONFERENCE, Bath, U.K (2000).
14. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehman, J., and Auer, S. Quality Assessment for Linked Data: A Survey. Semantic Web, 7(1), 63-93 (2016).
15. Siau, K., Wang, W. Building Trust in Artificial Intelligence, Machine Learning, and Robotics. Cutter Business Technology Journal (31), pp. 47–53 (2018).
16. Keymolen, E. Trust on the line: a philosophical exploration of trust in the networked era [Doctoral thesis]. Erasmus University Rotterdam (2016).
17. World Wide Web Consortium. RDF 1.1 Concepts and Abstract Syntax. In Cyganiak, R., Woods, D., Lanthaler, M. (Eds.) W3C Recommendation, https://www.w3.org/TR/rdf11-concepts/, last accessed 2022/4/3.
18. Khatri, V., and Brown, C.V. Designing Data Governance. Communications of the ACM, 53(1), pp. 148-152 (2010).
19. World Wide Web Consortium (W3C) Standards, https://www.w3.org/standards/, last accessed 2022/4/27.
20. Hernandez, J., McKenna, L., and Brennan, R. TKID: A Trusted Integrated Knowledge Dataspace for Sensitive Healthcare Data Sharing. In: IEEE 45th ANNUAL COMPUTERS, SOFTWARE, AND APPLICATIONS CONFERENCE, 1855-1860. (2021).
21. Jacon Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. Bert: Pretraining of deep bidirectional transformers for language understanding, https://arxiv.org/abs/1810.04805, last accessed 2022/4/27.
22. de Bruin, T., Rosemann, M., Freeze, R., Kulkarni, U., and Carey, W. Understanding the Main Phases of Developing a Maturity Assessment Model. 16th AUSTRALASIAN CONFERENCE ON INFORMATION SYSTEMS. Sydney, Australia. (2005, 29 Nov-2 Dec).
23. Civil Air Navigation Services Organisation. CANSO Standard of Excellence in Safety Management Systems, https://canso.org/publication/canso-standard-of-excellence-in-safety-management-systems/, last accessed 2022/4/3.
24. Janssen, M., Brous, P., Estevez, E., Barbosa, L.D., and Janowski, T. Data governance: Organizing data for trustworthy Artificial Intelligence. Government Information Quarterly, 37(3), 101493, (2020).