Appendix S3 Review Protocol

Section 1: Administrative Information

Title (Item 1a)

Trustworthy AI-enabled automated decision-making in the public sector: protocol for a systematic review

Item 1b: not applicable (not relevant to the planned review).

Registration (Item 2)

The review protocol will not be registered with PROSPERO or any other registry. However, it will be submitted alongside the manuscript as a supplemental document to the intended/targeted Journal.

Authors (Items 3a & 3b)

Contact Information

Olusegun Agbabiaka[1], PhD Scholar - **Affiliation:** Maynooth University, Ireland; **Contact Information:** School of Business, Maynooth University, Co. Kildare, Ireland; olusegun.agbabiaka@mu.ie; Corresponding author[1].

Adegboyega Ojo, Full Professor of Digital Government, AI Governance & Policy Analytics - **Affiliation:** Carleton University, Ottawa, Canada; **Contact Information:** School of Public Policy & Administration (SPPA), Carleton University, Ottawa, Canada; AdegboyegaOjo@cunet.carleton.ca

Contributions

Adegboyega Ojo (AO) is the guarantor of the scientific integrity of the review protocol. Olusegun Agbabiaka (OA) prepared and developed the review protocol. OA and AO contributed to the development of the selection criteria and data extraction strategy. AO reviewed and validated the final version of the protocol.

Amendments (Item 4)

In the event of an amendment to this protocol, such will be documented, the date of such amendment will be provided, the change will be described in details and the rationale for such change will be provided.

Support (Items 5a, 5b & 5c)

The planned systematic review for which this protocol is developed is not sponsored or funded.

Section 2: Introduction

Rationale (Item 6)

Governments globally are increasingly adopting disruptive digital technologies such as Artificial intelligence (AI) to deliver critical public services, and a major area where AI is being applied in the public sector is in the realm of automated decision-making (ADM), where it is being used to deliver services across different sensitive domains and make decisions with life and death consequences. However, as AI gains traction in the public sector, social and ethical issues such as privacy, autonomy, algorithmic bias, discrimination, transparency, etc. have been brought to the fore, creating a trust gap between government and citizens, leading to the publication of numerous documents on requirements for trustworthy AI (see Jobin, Ienca, and Vayena 2019). In spite of the rise in the number of studies on AI, there is a dearth of research in the public sector context (Campion et al. 2020; Sharma, Yadav, and Chopra 2020). Though, some previous systematic reviews have addressed algorithmic fairness (e.g. Starke et al. 2021); application of AI in government (e.g. Valle-Cruz et al. 2019; Reis et al. 2019),

research on AI-enabled automated decision-making (ADM) in the public sector (e.g. Kuziemski and Misuraca 2020) remains fragmented. The proposed systematic review, therefore, seeks to synthesize extant literature on trustworthy AI-enabled automated decision-making in the public sector to provide an overview of the application environment and a greater understanding of what constitutes trustworthiness of AI-enabled ADM in public sector decision-making contexts.

Objectives (Item 7)

The objectives of the systematic review are set out as follows:

1. To identify the contexts in which AI-enabled ADM is used in the public sector;
2. To identify the different artefacts associated with trustworthy AI-enabled ADM in the public sector;
3. To identify the different viewpoints from which researchers address trustworthy AI-enabled ADM in the public sector; and
4. To uncover requirements of trustworthiness in AI-enabled ADM in public sector decision-making contexts.

To this end, the proposed systematic review will answer the following research questions:

RQ1: What are the application contexts, implementation artefacts and study perspectives associated with AI-enabled ADM in the public sector?
RQ2: What are the requirements associated with the trustworthiness of AI-enabled ADM in the public sector?

Section 3: Methods

Eligibility Criteria – Inclusion/Exclusion Criteria (Item 8)

Studies will be selected according to the criteria outlined below:

Study Methodology/Design

- Studies of all research designs namely Qual, Quant and mixed methods will be included. We will include both empirical and conceptual studies; for a more balanced perspective, empirical studies will be included to capture evidence-based data, whilst conceptual studies will be included to reflect scholarly insights on trustworthiness of AI-enabled ADM in the public sector.
- Systematic reviews will be excluded (to ensure studies are not included twice).

Setting

- Only studies conducted in the public sector will be included.
- Studies will not be restricted or limited by geographical location; since the planned systematic review is a synthesis of evidence on a research phenomenon of global interest, we will include studies conducted from every part of the world to avoid selection bias.
- Articles focusing on ADM in other sectors, such as private companies, business, etc., will be excluded.

Date/Year of publication

- Only studies conducted between January 2000 and July 2022 will be included.

Language of publication

- Only studies published in the English Language will be included.

Publication type

- Published articles including peer-reviewed journal articles and conference papers will be included. Conference reviews/proceedings, book chapters, editorials, magazine or news features will be excluded.

Topic

- Studies to be included for review will need to sufficiently discuss automated decision-making powered by AI (machine learning, deep learning, rules-based algorithms, etc.) in the context of the public sector. Additionally, articles that discuss trustworthy AI or its related terms (e.g. lawful, ethical, explainable, accountable, responsible AI) or that contribute to our understanding of trustworthiness of AI/ADM in the context of the public sector will be included. Articles that discuss both solely automated decision-making (with no human judgment) and automated assisted decision-making/decision aid or support (assisting human judgment) in the context of the public sector will be included.

Information Sources (Items 9 & 10)

Sources (Item 9)

Trustworthy AI-enabled ADM being a subject at the intersection of many disciplines and fields, we will search the following two major databases:

- **Web of Science (core)** – it covers top quality journals in the social sciences and a range of other fields;
- **Scopus (core)**- it provides access to abstracts and citation databases across a wide range of disciplines;

Search Strategy (Item 10)

To ensure that the planned systematic review will include a broad range of relevant articles, we will do the following:

- Use a combination of both keywords and index/subject terms where applicable to conduct the search;
- Derive keywords/terms to be used in searching the literature from the research topic and domain;
- Use the AND/OR Boolean operators and wild card (where "*" is designated to capture possible wordings after an original term/phrase;
- Carry out a preliminary/pilot search using the resulting search strings on one of the databases, and based on the results, we will further refine the search strings accordingly for better and more focused results or use the strings as initially formulated; this process will be reported in the planned systematic review.

Based on the research topic and domain, and to enable us retrieve relevant articles, we will use the following terms:

For the "Trustworthy AI-ADM" component, the following terms will be used: "trustworthy" OR "trust*" OR "ethical" OR "robust" OR "lawful" OR "fair" OR "transparent" OR "responsible" OR "accountable" AND "AI" OR "artificial intelligence" AND "Automated decision making" OR "Algorithmic decision making" OR "ADM" OR "decision making" OR "decision*"

For the "public sector context" component, the following terms will be used: "public sector" OR "public administration" OR "public sphere" OR "public organisation*" OR "public organization*" OR "public service" OR "public agenc*" OR "public institution*" OR "government"

We will combine the two components with the "AND" operator, and accordingly use the resulting string combination to perform our search on the two databases:

("trustworthy" OR "trust*" OR "ethical" OR "robust" OR "lawful" OR "fair" OR "transparent" OR "responsible" OR "accountable") AND ("AI" OR "artificial intelligence") AND ("Automated decision making" OR "Algorithmic decision making" OR "ADM" OR "decision making" OR "decision*") AND ("public sector" OR "public administration" OR "public sphere" OR "public organisation*" OR "public organization*" OR "public service" OR "public agenc*" OR "public institution*" OR "government")

Study Records (Items 11a, 11b & 11c)

Data Management (Item 11a)

We will use Mendeley Reference Manager, a free web and desktop-based software programme to store, organise and sort articles generated from the search results, as well as facilitate collaboration between reviewers (the authors). For the purposes of the planned review, we will create and name a folder appropriately for the review on the Mendeley software to manage and sort articles by author, title, year parameters. We will use the software to import PDFs and citations/references (in BiBTex EndNote XML, RIS formats) directly from electronic databases for screening and inclusion. We will also use the software to sync articles with the Mendeley Web, which gives an option to download references with available full texts. We will use the software to share documents and references related to the planned review.

For data extraction, we will create an excel database (data extraction form) to manage extracted data (see Item 12).

Selection Process (Item 11b)

After searching the literature based on our search string, we will select articles (studies) using two levels of eligibility screening:

- **First Step:** (First level of screening - initial filtering for inclusion and exclusion): We will screen the titles, abstracts and keywords of identified articles against the eligibility criteria of *field; study design; language; year of publication; publication type and setting* to perform an initial filtering. Duplicate studies will also be removed at this stage. This step of the process will be performed by one reviewer (the first author);

- **Second Step:** (Second level of screening - assessing eligible articles for quality): At this stage, we will further screen full-texts of articles against the eligibility criterion of *topic* to determine if the articles sufficiently fulfil the requirements as provided in Item 8. This step will be performed by two reviewers (the first author) and subsequently validated by the second author. This step will also include two levels of reading of articles' full texts – first-level reading will be a skim/scan through (a speed read) and the second-level reading (a thorough read) to determine if articles should be included in review which will ultimately culminate into data extraction.

- We will present the study selection process with a PRISMA flow diagram.

Data Collection Process (Item 11c)

- Data extraction form will be created using Excel. The format will be in line with data items to be collected. Data extraction will be carried out by the first author and verified by the second author.
- In addition to the data items listed in Item 12, we will apply standard coding strategy to label and cluster extracted data and text fragments.
- The studies will be coded independently by the first reviewer (first author), but the labels and clustering process will be discussed from time to time with the second author. Every extracted text fragment that is difficult to code or label or categorize will be discussed between the two authors to address the issue, arrive at a common ground and ensure quality of the review. Interpretation of extracted data will be done by the first author and validated by the second author. Differences in opinion will be resolved during meetings.

Data Items (Item 12)

For each publication that will be included in the review, the following data items will be extracted (guided by the Mendeley article metadata template):

- Author(s);
- Year of publication;
- Title;
- Type of publication (Journal or conference article);
- Methods;
- Text fragments that relate to the systematic review's overall objectives and research questions – (see Item 11c for more details).

Outcomes and Prioritization (Item 13)

Based on the overall objectives and research questions, data will be sought for the following outcomes:

- **Outcome 1:** application contexts of AI-enabled ADM in the public sector;
- **Outcome 2:** implementation artefacts associated with trustworthy AI-enabled ADM in the public sector;
- **Outcome 3:** the study perspectives associated with trustworthy AI-enabled ADM in the public sector;
- **Outcome 4:** Requirements for trustworthiness of AI-enabled ADM in the public sector.

Risk of Bias of Individual Studies (Item 14)

To avoid "re-reviewing" articles all over again, we plan to ensure that the source title for journal articles included in the review should be at least **ABS2 or Q2.** For conference articles, the peer-review process will be our first basis for quality assessment. This is in addition to ensuring strong relevance of both journal and conference articles – in terms of aims, design methods and results (where applicable).

Data Synthesis (Items 15a, 15b, 15c & 15d)

Because studies that will be included for evidence synthesis are not likely to be sufficiently similar or homogenous in terms of research design or methodology, synthesis will be done at a level of abstraction that the information collected about the articles permit.

Meta-bias(es) (Item 16)

One way we have guarded against 'meta-bias' is in the choice of articles to be included in the review (see Item 8). To avoid bias, we will not be selective in the reporting of our findings from the reviewed articles. Additionally, both authors will review the coding, categories, etc.

Confidence in Cumulative Estimate (Item 17)

To assure and increase confidence in the cumulative body of evidence, we will adhere strictly to the review protocol, and abide with the specific guidelines outlined in Item 8 (inclusion/exclusion criteria); Items 9 and 10 (data sources and search strategies); Item 11b (selection process); Item 14 (assessment of quality and risk of bias in included studies), to significantly reduce selection bias, publication bias, reporting bias, etc.

Activity Outline

	Lead reviewer(s)
Preparation of draft protocol	Olusegun Agbabiaka (OA)
Internal review & validation	Prof. Adegboyega Ojo (AO)
Searching and study selection	Olusegun Agbabiaka
Quality Assessment	OA & AO as Lead Assessor
Data Extraction	OA; AO (for validation of data)
Draft report for peer review	OA; AO
Finalize manuscript & submit for publication	OA; AO (for validation of manuscript)
Celebrate publication	OA; AO

References

Campion, A., Gasco-Hernandez, M., Mikhaylov, S., & Esteve, M. (2020). Overcoming the Challenges of Collaboratively Adopting Artificial Intelligence in the Public Sector. *Social Science Computer Review DOI: 10.1177/0894439320979953*, 1-16.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 389–399.

Kuziemski , M., & Misuraca, G. (2020). AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications Policy 44 (6). https://doi.org/10.1016/j.telpol.2020.101976*, 1-13.

Okoli, C. (2015). A Guide to Conducting a Standalone Systematic Literature Review. *Communications of the Association for Information Systems*: Vol. 37 , Article 43. DOI: 10.17705/1CAIS.03743 at the 20th Annual International Conference on Digital Government Research (dg.o 2019), Dubai, June 18–20.

Page, M., McKenzie, J., Bossuyt, P., Boutron, I., Hoffmann, T., Mulrow, C., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMC 10: 89 https://doi.org/10.1186/s13643-021-01626-4*, 1-11.

Reis, J., Santo, P., & Melão, N. (2019). Impacts of Artificial Intelligence on Public Administration: A Systematic Literature Review. *14th Iberian Conference on Information Systems and Technologies (CISTI)*, (pp. 1-7 doi: 10.23919/CISTI.2019.8760893.). Coimbra, Portugal.

Rethlefsen, M., Kirtley, S., Waffenschmidt, S., Ayala, A., Moher, D., Page, M., et al. (2021). PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. *BMC 10: 39 https://doi.org/10.1186/s13643-020-01542-z*, 1-19.

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., et al. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ*, 1-25 BMJ 2015;349:g7647 doi: 10.1136/bmj.g7647.

Sharma, G., Yadav, A., & Chopra, R. (2020). Artificial intelligence and effective governance: A review, critique and research agenda. *Sustainable Futures 2*, https://doi.org/10.1016/j.sftr.2019.100004.

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2021). Fairness Perceptions of Algorithmic Decision-Making: A Systematic Review of the Empirical Literature. *Computer Science*, 1-55.

Valle-Cruz, D., Ruvalcaba-Gomez, E., Sandoval-Almazan, R., & Ignacio Criado, J. (2019). A Review of Artificial Intelligence in Government and its Potential from a Public Policy Perspective. *In dg.o 2019: 20th Annual International Conference on Digital Government Research (dg.o 2019),June 18–20, 2019.* (pp. 91-99). Dubai, United Arab Emirates: ACM, New York, NY, USA.

Xiao, Y., & Watson, M. (2019). Guidance on Conducting a Systematic Literature Review. *Journal of Planning Education and Research 39(1)* , 93–112.

Methodology to guide the planned systematic review

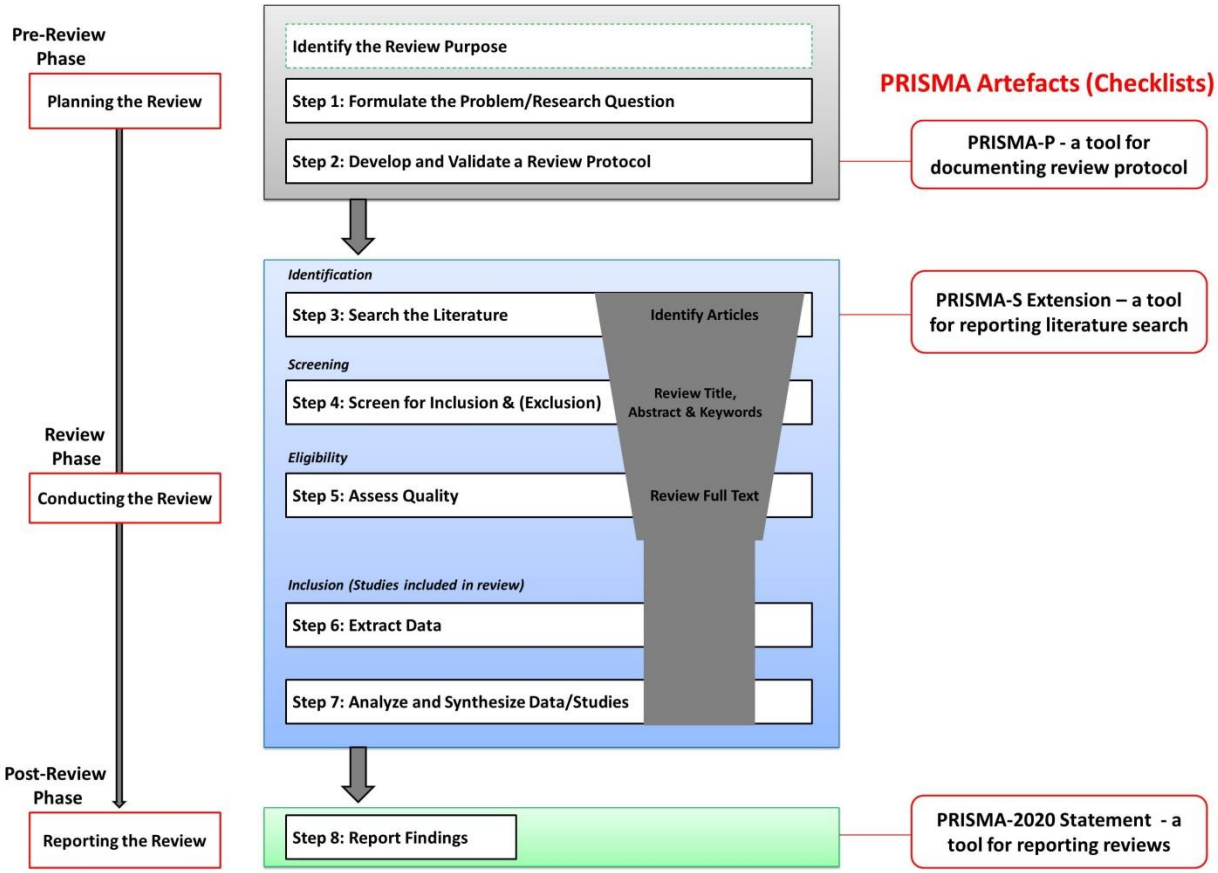

Figure 1. Methodology Guiding the Review
Agbabiaka and Ojo (2022)
Adapted from Okoli (2015) and Xiao and Watson (2019)

NOTE: the conduct of the systematic review will be guided by the 3-stage approach we adapted from Okoli (2015) and Xiao and Watson (2019), alongside the review protocol, and relevant PRISMA artefacts for greater rigour and integrity of the planned systematic review.