# Trustworthiness of systematic review automation

## An interview at Coventry

Xiaorui Jiang xiaorui.jiang@coventry.ac.uk
Centre for Computational Sciences and Mathematical Modelling, Coventry University

### Abstract

[TO_BE_ADDED]

### 1. Introduction

[TO_BE_ADDED]

### 2. Methods

We conducted a series of interviews with researchers in Coventry University who do systematic reviews (SR) regularly. After searching the PURE portal of Coventry University with "systematic review", "systematic literature review", "meta-analysis", "qualitative review" in the titles of the publications, we identified more than 50 regular systematic reviewers with at least one published systematic reviewer papers, most with two or more. We sent out invitation emails to 50 researchers, including PhD candidates who were close to graduation at the time of invitations. We received about 15 positive responses and finally 10 were successfully invited to interviews within the required timeframe.

We used interviews to collect and investigate three categories of information: (1) the visibility of systematic review tools and systematic review automation (SR automation) to human reviewers (Q1-Q2 in Sect. 3); (2) the acceptance of SR automation to regular systematic reviewers, i.e. whether human reviewers trust automated tools, and the factors impacting such acceptance (Q3-Q4 in Sect. 3); and (3) what impacts the explanability of SR automation tools can have on their acceptance and what expectations end users have on explainable SR automation solutions.

The demography of participants shows a vert skewed distribution across fields, mainly because systematic reviewers mainly come from medicine, health, and life sciences domains (The "Field" column in Table 1). However, we tried our best to invite participants at a diverse range of career stages. The "Career stage" column in Table 1 shows the number of years of post-doctoral research experience in the academia, and "0" means PhD candidates close to graduation. Therefore, we have much universal distribution of career stage. This guaranteed a more comprehensive and diverse set of opinions about systematic review automation.

Our data analysis was partially rooted in thematic analysis. We manually corrected and aligned the auto transcripts generated by the meeting software, here Microsoft Teams, we used. Because we had predefined an initial set of questions to answer (detailed in Sect. 3.1), we first aligned the transcripts segments to these questions. Thematic analysis was applied to the non-aligned transcripts, allowing new themes to emerge. Then, we manually identified the codes used to perform content analysis of the interviews. In total, we had three themes as described above, each having two subthemes (detailed in Sect. 3.1), and one coding scheme to analyse each subtheme (detailed in Table 1.)

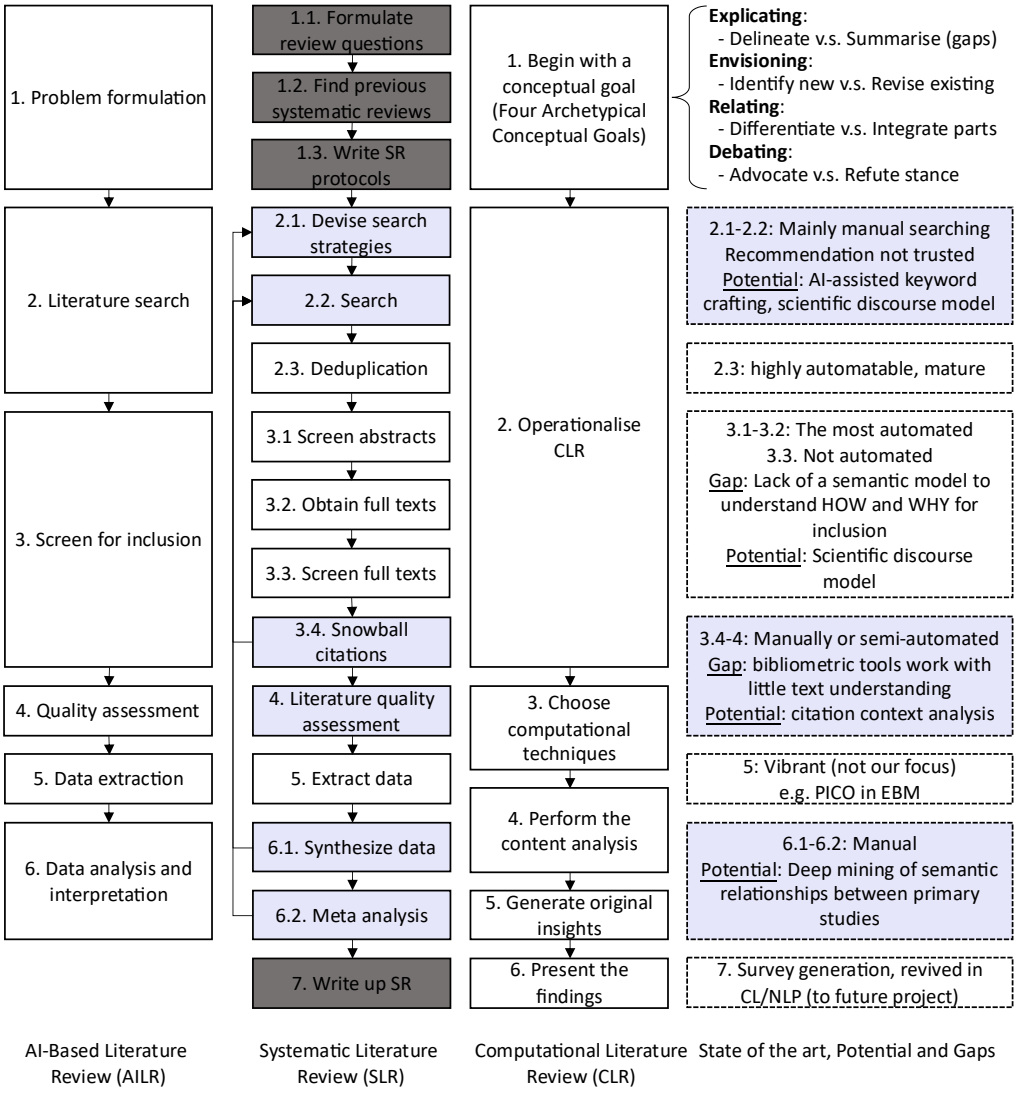


Figure 1. Flow chart of systematic review and gap analysis of two related AI-bases paradigms of review automation. Grey boxes represent the steps that are mainly done in a manual way.

Table 1. Summary of Content Analysis of Interviews.

| No. | Career Stage | Subject | Experience with SR tools | Experience with SR automation | Expected steps to automate | % Time on searching & screening | Trust | *Why trust/not trust* Factors impacting trust | After given explanations | Expectations from explanations (given to v.s. *to give*) |
|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 7+ | Bibliometrics | NVivo | No | Thematic analysis | n/a | NO | *Not conclusive*<br>Low recall, low precision | Better | Detailed report<br>*n/a* |
| P2 | 4-6 | Medicine | Ryyan, Covidence | No | Search, de-duplicate, screen | n/a | Hesitant | *Blackbox*<br>Piloted*; As safety check*; | Increased | Reasons/Rationales; Degree of certainty; PRISMA flowchart |
| P3 | 4-6 | Medicine | Refworks, NVivo | No | Search, de-duplicate, screen, quality assessment, meta analysis | 50% | Perhaps | *Accuracy, reproducibility*<br>Blended (human participation); Endorsed (using it in publications); Piloted | Increased | Reasons based on eligibility criteria<br>*n/a* |
| P4 | 1-3 | Pharmacy | Endnote, Rayyan, Covidence | No | Search (across databases) | 50% | YES* | *Move forward*<br>Endorsed (using it in publications); safety check; Piloted | n/a | n/a<br>*n/a* |
| P5 | 7+ | Life sciences | Covidence, Rayyan | Yes? | n/a | 40-50% (>30-40%) | Tend to | *n/a*<br>Piloted; Blended; Experiment/re-run | Even better | Reasons<br>*feedback (drop list based on criteria)* |
| P6 | 0 | Medicine | Covidence, NVivo | No | Screen, quality assessment, PRISMA, summary | 10%+40% | Tend to | *Machine dictates*<br>Safety-check; piloted; endorsed (by publications); Experiment/re-run | Even better | Reasons<br>*Feedback (Yes/No)* |
| P7 | 3-6 | Health | NVivo | No | Search, screen (abstract plus full text), quality assessment | 50% | Perhaps | *Black box; "*see it *in action"*<br>Inspect decisions; Piloted; Experiment/re-run | Increased | Reasons; highlights; association between highlights and reasons<br>*Feedback based on criteria* |
| P8 | 0 | Medicine | Covidence, Rayyan, NVivo | No | n/a | 40% | Perhaps | *Black box; error*<br>Piloted (full recall, replicability); | Increased | Reasons aligned with criteria<br>*Feedback (checkbox based on criteria)* |

* P8 responded with a strong YES but with conditions that the performance of SR automation tools were proved by pilot use

## 3. Interview Analysis

### 3.1. Focuses of Analysis

The following six questions were the focuses of analysis.

- **Q1. The visibility of the systematic review automation to academia.** Here we mainly as
- **Q2. The visibility of the systematic review automation tools to academia.** We
- **Q3. The willingness of adopting/trust in systematic review automation tools.**
- **Q4. The factors that impact the trust in systematic review automation tools.**
- Q5. The impact of explanability on the trust in systematic review automation tools
- Q6. End users' expectations on the explanability of systematic review automation tools

Some research questions were predefined in the Interview Instructions (Appendix A). They are RQ1, RQ2, RQ3, and RQ5. For these questions, I manually picked out the longest segments of the interview transcripts and aligned the segments to one RQ. The coding scheme for the interview results of each question (here RQ3 and Q5) was developed by clustering interview participants' expressions that had a similar meaning.

The remaining two research questions, RQ4 and RQ6, were gradually emerged from thematic analysis of the remaining transcripts that followed the answers for RQ3 and RQ5 respectively. Corresponding transcript segments were manually aligned to each research question. The coding schemes for RQ4 and RQ6 were also developed by analysing the corresponding transcripts segments in a similar way as above.

### 3.2. Interview Results

Answers to each research question that we intend to answer were coded in Table 1. Note that, two participants were finally found not regular systematic reviewers. They both came from the medical engineering subfield and had a stronger engineering background rather than medical and life sciences background. They shared a lot of valuable opinions about comprehensive literature review using in-house protocols and criteria that were not mappable to the common practice of systematic reviews. Therefore, we excluded their interview results because more than half of the columns in Table 1 could not be filled based on their interviews.

#### Q1. The visibility of the systematic review automation to academia

The results are presented in the "Experience with SR automation" column. Unfortunately, although been developed for nearly two decades, systematic review automation is almost totally invisible to human reviewers. Note that we are unable to conclude such invisibility to the whole academia as data was only collected from Coventry University. We also want to emphasise that it was the concept of "systematic review automation" that was invisible to human reviewers due to the knowledge barrier between distinct fields like medicine and AI. However, some systematic review tools, even some with partial automatability, have already been used by many participants, without explicitly noticing such automation functionalities (detailed in the next subsection).

There was only one participants aware of such automation in the systematic review tools. This was participant P5, the most senior researcher among all participants. Although P5 was less inclined to fully trust SR automation tools, this participant was one of the most positive and confirmative on the role of SR automation in systematic review in practice. An important opinion from P5 was the more SRs you do and the more frequently you do SRs, the more inclined you are to accept SR automation. P5 said,

> *"So these people (i.e., PhD students), most of them that's they have experience in that level they you prefer do things manually. I feel like the colleagues that I've work, they they've done a systematic review here and*

*there, …, They have a completely different view when you do like a proper systematic review like huge ones, and you do with a group of researchers. If I talk to any of my colleagues that they do this professionally. It's like their job. There is no way that they will not use a software. … you cannot even think about."*

Witnessing such a knowledge gap in the interviews, we had to introduce the concept of "systematic review automation" to all human reviewers, explain how AI or machine learning does so in certain SR steps, and advertising some existing SR automation tools that were used for nearly a decade or recently published. Interestingly, all participants were quick at accepting the concept and responded overall positively to such tools. This finding corroborates with conclusions of previous studies that lack of awareness and knowledge was a hinder of user trust in ST automation tools (O'Connor et al., 2019a; O'Connor et al., 2019b).

**Q2. The visibility of the systematic review automation tools to academia**

The answers for Q2 can be seen from the "Experience with SR tools" column. As stated in the last subsection, in fact most participants had already used certain SR automation tools, or at least certain SR tools that have some automation functionalities. For example, Rayyan is a tool that has the functionality to learn from human reviewers' initial annotations to mark the relevant or irrelevant studies and highlight them in green/red colours. Studies were also ranked using colour intensity. In other words, Rayyan has the functionality to automate the selection of primary studies, by automating the abstract screening step (Step 3.1 in Figure 1). Covidence is tool that can take in relevant and irrelevant studies annotated by human reviewers and automatically generate a PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses; Page et al., 2020; Rethlefsen et al., 2021) flowchart for reporting the systematic review process. In fact, reporting the systematic review process is a mandatory requirement (mentioned by two participants) for publishing a systematic review, so the automatic generation of the PRISMA flowchart is an appealing feature.

We also notice that the range of tools used by the participants were quite limited, almost confined to Rayyan, Covidence, and NVivo (for thematic analysis). Not a single automation tool for data extraction, data synthesis and meta analysis was mentioned. This corroborates with the SR steps that our participants expected to automate with the help of AI/ML (see the "Expected steps to automate" column). These steps were confined mainly to literature searching and citation screening (Step 2.1-3.3 in Figure 1). Participant P1 mentioned thematic analysis, which basically means giving an explainable name to the cluster of words generated by NVivo to help perform content analysis. From another side, most participants agreed that literature searching and citation screening will roughly take up 40-50% of the total time needed to finish a systematic review (the "% Time on searching & screening" column). If we take into consideration the fact that usually at two annotators are needed for screening, a lot of time can be saved even if only automated citation screening is used.

Two participants explicitly mentioned that data extraction is "fun" for human reviewers so they were bored by doing it manually. Two participants expressed their concern in the feasibility of automating data extraction and data synthesis. They thought these two steps required very high-level cognitive capabilities that they did not believe current AI technologies could achieve. Finally, one participant (P3) pointed out the high potential of automated meta-analysis because this step only involves statistical analysis. The argument was right, but unfortunately no participant was aware of any such automation tools on the market. Actually, automated meta analysis is the most mature tool in the systematic review automation tool stack. Similarly, no participant was aware of any SR automation tools from SR Step 4 to Step 7 in Figure 1, except the automatic generation of PRISMA flowchart, which partially belongs to Step 7.

**Q3. The willingness of adopting/trust in systematic review automation tools**

From the "Trust" column, we were surprised to see that the level of willingness to adopt SR automation tools (or in other words the level of trust in such tools) was higher than our initial expectation. Originally, the coding scheme included "NO", "Partial", and "YES". However, to reflect the more subtle differences between different responses, the coding scheme was improved to kind of a 5-grade system: NO, Hesitant, Perhaps, Tend to, YES. "Hesitant" means "tend not to trust" even after deeper discussion about the ways to improve trust and the impact of explanations for machine learning decisions. "Perhaps" is weaker than "Tend to" because the former requires interactions between SR automation too and user, while "Tend to" means trust will be established if SR automations tools are piloted on certain human-written systematic reviews for large-scale evaluation and proved their good performance on them.

There was only one "NO" and one "Hesitant". There were three "Perhaps" and two "Tend to". More interestingly, there was one "YES" (P4). Participant P4 had an important comment on the acceptance of or trust in automation tools,

> *"To be honest, I would trust it. … I think if if we want to move forward in our systematic review, we have to have a solid automated system that do screening for us that we need to trust as a researcher.*
>
> *…*
>
> *And then if it's been advised to use it as a researcher, I think I'm obligated to do it through the system because that will help."*

Although P4 was so confirmative in answering the question about trust, a follow-up comment revealed some important and common ways of improving user trust – through piloting and endorsement (see the next subsection),

> *"But I think to be able to trust it, I think we have to read a lot of publication around that the system and how accurate and sensitive it is."*

Participant P6 left a similar comment although P6's answer was classified as "Tend to",

> *"There are many things that's been dictated by computers, and we don't even question that. Why would we question this then?"*

**Q4. The factors that impact the trust in systematic review automation tools**

The reasons why participants did not trust automation tools (or AI/ML substituting human decisions) and the ways to improve user trust in SR automation tools were summarised in the "*Why trust/not trust* – Factors impacting trust" column. The coding scheme included the following codes:

- Blended: Needs human participation in the SR process, e.g., SR automation tools return reasons for machine-made decisions, or users give feedback to machine-made decisions, etc.
- Piloted: Needs large-scale evaluation and demonstrated good performance at or close to human level.
- Endorsed: Needs real-world use by organisations or in publications (i.e., published systematic reviews using SR automation tools exist)
- Experiment/re-run: This is a very interesting feature suggested by three participants (P5, P6, and P7). It was a high-level interaction (human being blended with machine) where users may check machine-made decisions, make some changes, re-run the automation algorithm, and see what changes happen.
- Safety-check: On the contrary, safety-check was defined as a weaker way of being "Piloted" or "Blended" as only a sample of instances are required to be tested or checked by user.

An important finding, though likely to be biased towards the certain population of participants, is that Users' trust will be much increased and most participants' level of trust will be increased to a standard that trust in SR automation tools will be established. For some participants, it was enough to establish high trust in SR automation tools if such tools had been piloted on a large number of human-written systematic reviews and proved near-human-level performance. 100% trust would be established if SR automation tools had been endorsed by other users in real-world practice, i.e., either used by large organisations or used by other researchers in publishing systematic reviews. For some participants, the necessary requirement for them to establish such trust was the ability to interact with SR automation tools, by checking the reasons for machine-made decisions. In some way, they take SR automation tools as a partner to replace the role of the second annotator in citation screening or quality assessment. Finally, participant P6 raised a very interesting point. Different from most interviewees who wanted to teach SR automation tools to improve the performance by giving feedback, P6 also believed that SR automation tools could teach users in the early stage of doing an SR,

> *"If they provided the explanation to me and then I see this article to, maybe, I can even reflect on my eligibility criteria. So what is not well defined and what may be confusing. It would be a mutual learning process in the at the beginning."*

**Q5. The impact of explanability on the trust in systematic review automation tools**

The coding scheme for the impact of the measures for improving trust included three codes:

- Increased: Trust increased, but not necessarily to a level of trust. For example, P1 still said he needs a "detailed report" of all explanations (see the "Expectations from explanations" column) and needs to "go through all of them". P2 was still "hesitant" to use such automation tools.
- Better: Trust will be increased to a level that the user is willing to adopt such tools in their own SRs and trust the result of SR automation tools if they are piloted or endorsed.
- n/a: This only applies to the few participants who have strong trust in SR automation tools based on AI/ML.

We could see that three participants (P2, P7, and P8) mentioned that machine learning being a "black box" was a reason for them to not trust SR automation tools. In the interviews some participants pointed out the importance of ability to interact with SR automation tools (P3 and P7). We further asked what such interaction was purposed for. In the direction from machine to human, the answer was explanations of the machine-made decisions. From the "After given explanations" column, we see all participants agreed that explanations would increase their trust in SR automation tools. We guess that, although not explicitly mentioned, other participants were also aware of the black box nature of automation tools, and this was the reason for increased trust if the black box was opened by giving explanations.

**Q6. End users' expectations on the explanability of systematic review automation tools**

The results were summarised in the "Expectations from explanations" column. The coding scheme for the machine-to-human interaction included five codes:

- Degree of certainty: About the confident levels of machine-made decisions.
- Highlights: Highlighted text snippets, usually keywords, that are coloured to visualise whether they are indicative of inclusion or exclusion.
- Reasons (same as Rationales): An advanced type of explanation detailing the reasons for decision-making, usually based on eligibility criteria, i.e., which inclusion criterion is matched and which exclusion criterion is matched.

- Association between Reasons and highlights: A more advanced type of explanation which provides the evidence for making decisions.
- PRISMA flowcharts: It is a mandatory requirement of publishing a systematic review in a peer-reviewed journal or a systematic review database. This means generating the explanation of the whole systematic review process, which is an advanced feature.

The coding scheme for the human-to-machine interaction, i.e., user feedback, included three codes:

- Feedback by YES/NO: Binary indicators of the correctness of machine-made decisions
- Feedback based on eligibility criteria: An advanced type of user feedback detailing why a machine-made decision is right or wrong. Two formats of feedback were mentioned: drop-down list (single selection) or check box (multiple selections).

In overall, we ca conclude that users, more precisely systematic reviewers from fields that are far from AI/ML or computer science, mainly expect easy-to-use ways of feedback/interaction. For giving feedback to SR automation tools, they do not expect complex forms of feedback. Instead, they prefer simple feedback methods by button clicking. Because eligibility criteria are the rules human reviewers carefully craft and 100% rely on, so the forms of feedback are better to be aligned with eligibility criteria. This is technical feasible because all well-written SRs must define a rigid protocol and register the protocol with a systematic review database before submission for review. The protocol must also be published together with an SR and is publicly accessible.

In the other way round, human reviewers did not expect complex forms of explanations: most expected highlights and reasons aligned with eligibility criteria. Most of them did not expect the complex reasoning process behind a decision, although some participants mentioned that a decision is usually made by checking the criteria one by one and thus the decision process works like a decision tree in machine learning. However, we noticed that, it was not because human reviewers did not want complex explanations, such as a full sentence explaining the reasoning process. Instead it was because human reviewers were unable to imagine that existing AI technologies are able to generate complex formats of explanation in human language, due to their lack of knowledge in AI, as one participant (P7) commented below.

> *"I can't imagine how you'd. I imagine that's a huge undertaking to then take those decisions and then get a natural, you know, a sentence that people would understand for every single thing, for every single project."*

## 4. Conclusions and Discussions

[TO_BE_ADDED]

## References


Antons, D., Breidbach, C. F., Joshi, A. M., & Salge, T. O. (2021). Computational literature reviews: Method, algorithms, and roadmap. *Organizational Research Methods*, First Published March 9, 2021. https://doi.org/10.1177/1094428121991230

Arno, A., Elliot, J., Wallace, B., et al. (2021). The views of health guideline developers on the use of automation in health evidence synthesis. *Systematic Reviews*, *10* Article No. 16. https://doi.org/10.1186/s13643-020-01569-2

Hoang, L., & Schneider, J. (2022). Opportunities for computer support for systematic reviewing – a gap analysis. In: Chowdhury, G., McLeod, J., Gillet, V., Willett, P. (eds) *Transforming Digital Worlds (iConference'2018), Lecture Notes in Computer Science(10766)*. Springer, Cham. https://doi.org/10.1007/978-3-319-78105-1_40

Gates, M., Gates, A., Pieper, D., et al. (2022). Reporting guideline for overviews of reviews of healthcare interventions: The Preferred Reporting Items for Overviews of Reviews (PRIOR) Explanation & Elaboration. *British Medical Journal*, Preprint. https://kclpure.kcl.ac.uk/portal/files/173860204/PRIOR_Manuscript_R1_19May2022.pdf

O'Connor, A. M., Tsafnat, G., Gilbert, S. B., et al. (2018). Moving toward the automation of the systematic review process: a summary of discussions at the second meeting of International Collaboration for the Automation of Systematic Reviews (ICASR). *Systematic Reviews*, *8*, Article No. 3. DOI: https://doi.org/10.1186/s13643-017-0667-4

O'Connor, A. M., Tsafnat, G., Gilbert, S. B., et al. (2019a). Still moving toward automation of the systematic review process: a summary of discussions at the third meeting of the International Collaboration for Automation of Systematic Reviews (ICASR). *Systematic Reviews*, *8*, Article No. 57. DOI: https://doi.org/10.1016/j.infsof.2021.106589

O'Connor, A. M., Tsafnat, G., Thomas, J., et al. (2019b). A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? *Systematic Reviews,* 8, Article No. 143. https://doi.org/10.1186/s13643-019-1062-0

O'Connor, A. M., Glasziou, P., Taylor, M., et al. (2020). A focus on cross-purpose tools, automated recognition of study design in multiple disciplines, and evaluation of automation tools: a summary of significant discussions at the fourth meeting of the International Collaboration for Automation of Systematic Reviews (ICASR). *Systematic Reviews*, *9*, Article No. 100.

Page, M. J., McKenzie, J. E., Bossuyt, P., et al. (2020). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *The BMJ*, *372*, No. 71. https://doi.org/10.1136/bmj.n71

Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., et al. (2021). PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. *Systematic Reviews*, *10*, Article No. 39. https://doi.org/10.1186/s13643-020-01542-z

Tsafnat, G., Glasziou, P., Choong, M. K., et al. (2014). Systematic review automation technologies. *Systematic Reviews*, *3*(74). https://doi.org/10.1186/2046-4053-3-74

van Dinter, R., Tekinerdogan, B., & Catal, C. (2021). Automation of systematic literature reviews: A systematic literature review. *Information and Software Technology*, *136*, pages 106589. DOI: https://doi.org/10.1016/j.infsof.2021.106589

Wagner, G., Lukyanenko, R., Peré, G. (2021). Artificial intelligence and the conduct of literature reviews. *Journal of Information Technology*, First Published October 8, 2021. https://doi.org/10.1177/02683962211048201

Xiong, P., Buffet, S., Iqbal, S., et al. (2022). Towards a robust and trustworthy machine learning system development: An engineering perspective. *Journal of Information Security and Applications*, *65*, 103121. https://doi.org/10.1016/j.jisa.2022.103121

# Appendices

## A. Interview Instructions

Total time: 60 minutes to 90 minutes.

Part 1. Show the **diagram** about the alignments between SLR (Systematic Literature Review), AILR (Artificial Intelligence based Literature Review), and CLR (Computational Literature Review), ask the interviewee whether this matches his/her own experience in doing systematic review. Ask him/her to describe the **way** he/she does systematic review and the **difficulties** he/she met.

Ask the participants to share their estimation of how much time they spent on each stage, and their need for automating or semi-automating some steps of them using AI/ML.

Part 2. Ask the interviewee to talk about he/she did **literature search**, and what he/she thinks about the potential of improvement and anticipate the way AI/ML can help. Similarly, ask the interviewee to talk about he/she did the **selection of primary studies**, i.e., **screening**, and how he/she **assess the quality** of study.

Part 3. Based on his/her response, stimulate the interviewee to discuss about the **trust issue** of AI/ML tools and their generated results, e.g., how they trust the AI-suggested search keywords, how they trust the returned papers and the rankings of the papers, etc.

Part 4. Based on his/her response, let the interviewee further discuss what is the most comprehensive way of **presenting explanations** to the AI/ML results. Is it in short natural language explanations? Should a group of results be given the same/quasi-same explanation based on semantic clustering?

Part 5. Ask the interviewee to talk about his ideas about some other aspects of SLR automation and explanation, such as the aspects listed below, if they have not been touched and time permits:

- explaining the **relationships** between selected primary studies,
- explaining the domain **evolution** among selected primary studies,
- explaining the **reproducibility** of SLR automation, i.e., by automatically generating PRISMA statement diagram and the textual descriptions of the PRISMA statement.
- Explaining how well the SR automation tools perform by **large-scale evaluation**, for the purpose of persuading more people into using such tools
- Etc.