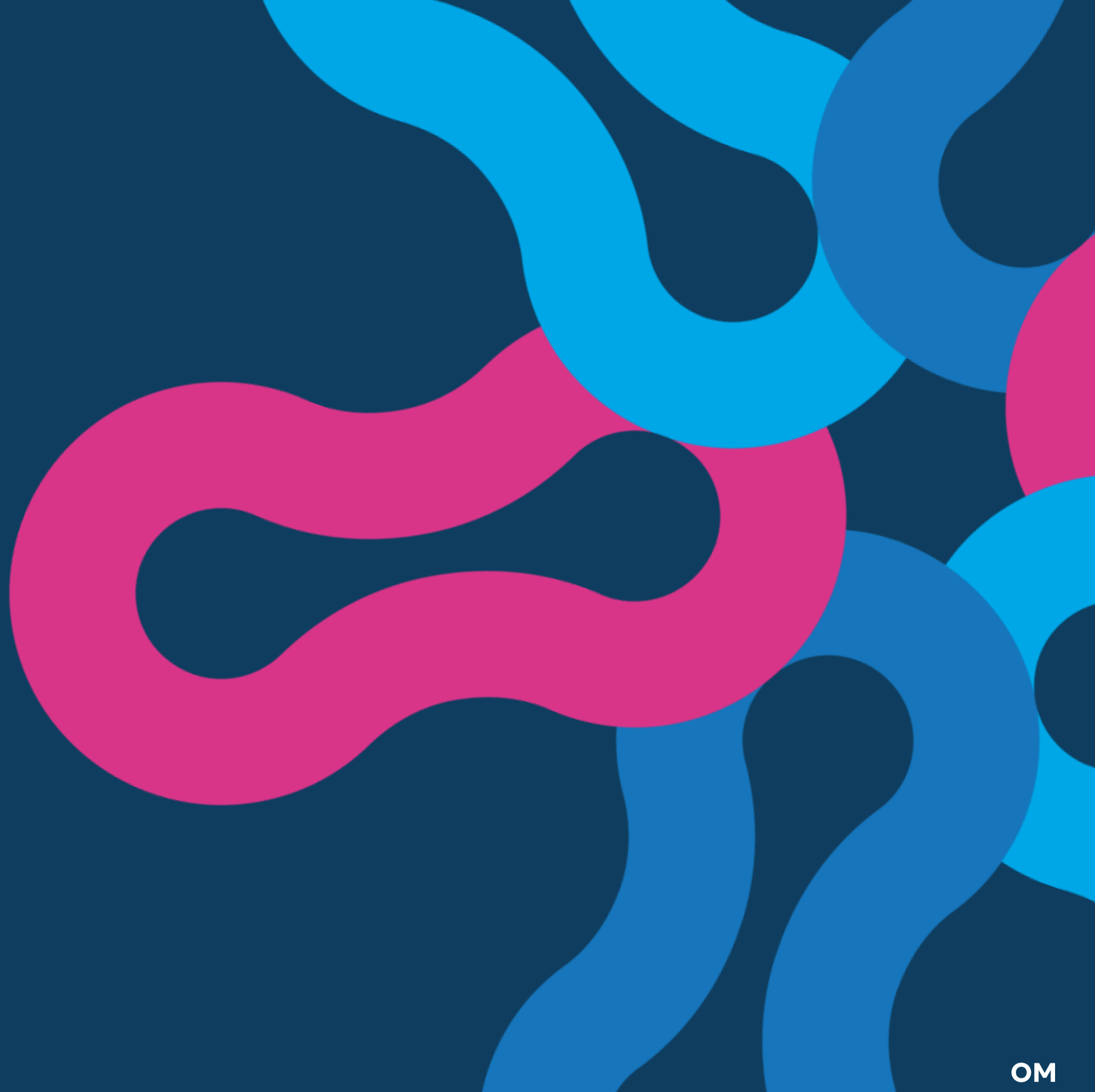
04-06-2026

Decision-making about How and When to use AI in Evaluation

Olivia Melvin, CERE
Tahirah David, UConn
Bianca Montrosse-Moorhead, UConn
Sarah Mason, CERE

Welcome & Introduction



Meet the Presenters

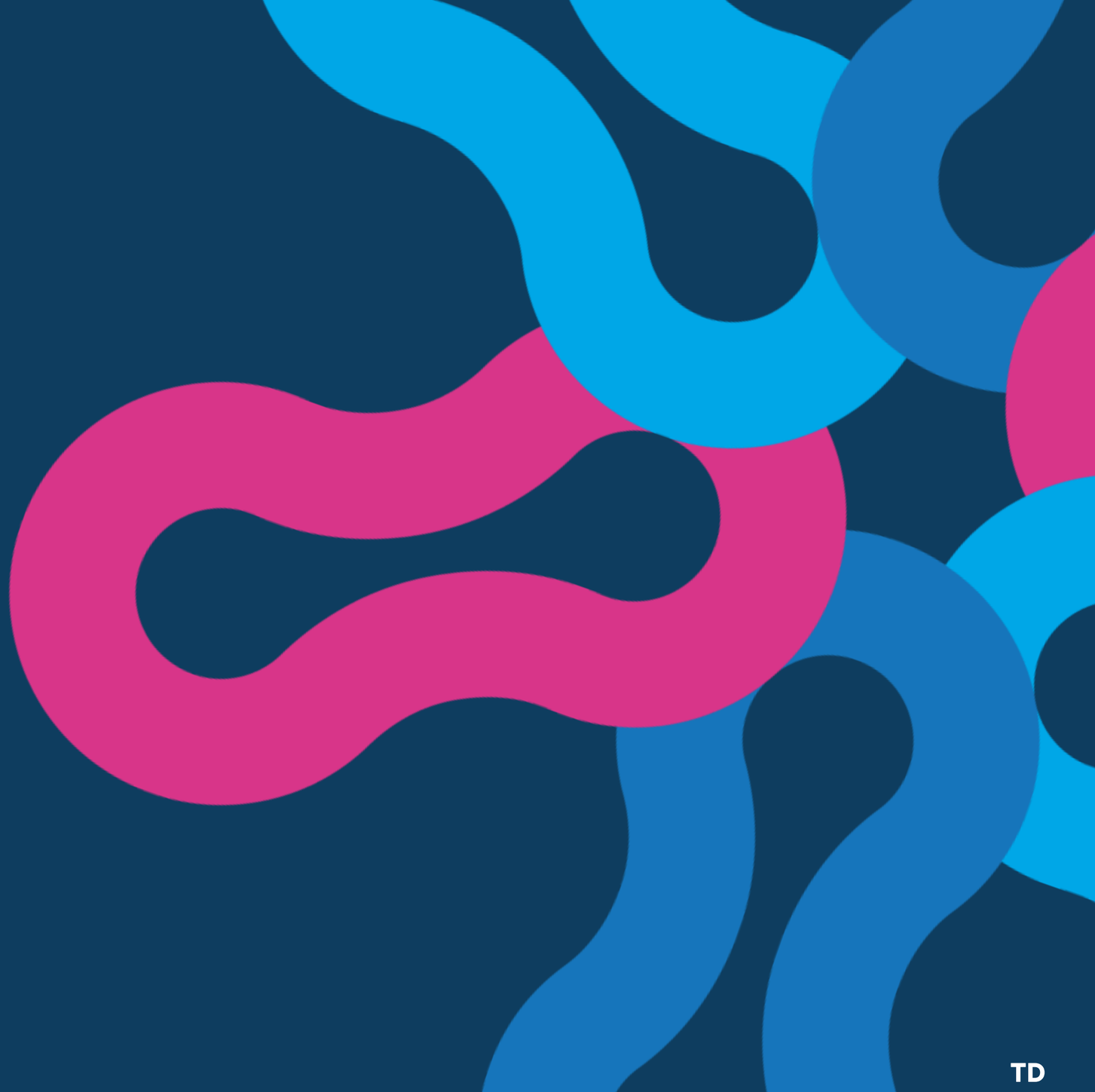
Olivia Melvin is an Evaluation Associate at the Center for Research Evaluation where she leads program evaluation, capacity-building, and learning partnership projects. Her background includes a B.A. in International Studies and Mandarin and an M.A. in Political Science (International Conflict & Peace Studies) from the University of Mississippi.

Tahirah David is currently a Graduate Assistant at University of Connecticut in the Research Methods, Measurement, and Evaluation program. Her current research focuses on AI in evaluation, meta evaluation, and standards. She has prior experience in public health and youth social program evaluations in the Caribbean.

Dr. Bianca Montrosse-Moorhead is Professor and Program Chair for the Research Methods, Measurement, and Evaluation program at the University of Connecticut. She is also an Affiliated Faculty member for the Center for Educational Policy Analysis, Research, and Evaluation at UConn. Dr. Montrosse-Moorhead specializes in evaluation methodology, theory, practice, and capacity building.

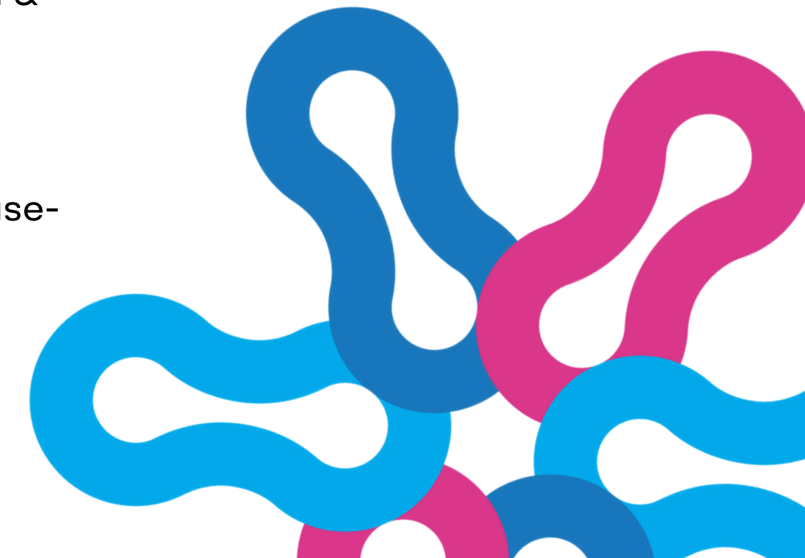
Dr. Sarah Mason is Director of the Center for Research Evaluation at the University of Mississippi where she leads a team of evaluators who oversee utilization-focused, place-based evaluations in the federal, state, non-profit and philanthropic sectors. Over the past 20 years she has conducted research and evaluation projects across the United States, Australia, South, and Southeast Asia.

Rationale



Overview of AI in Evaluation Journey

- ChatGPT launches in late 2022.
- A surge in the release of Artificial Intelligence (AI) models and tools on the market
- Efficiency, scalability and accessibility have become a key features in AI use in the evaluation
- AI is being embraced throughout the evaluation field
- Literature on AI in Evaluation First Generation (2018-2023): early use-case exploration & risk assessment, AI as enhancing human capacity, resistance to adopting non-human processes
- Literature on AI in Evaluation Second Generation (2023-2025): further exploration of use-cases further unpacking moral and ethical concerns of AI early development of frameworks & parameters for use in specific contexts



Where are we?

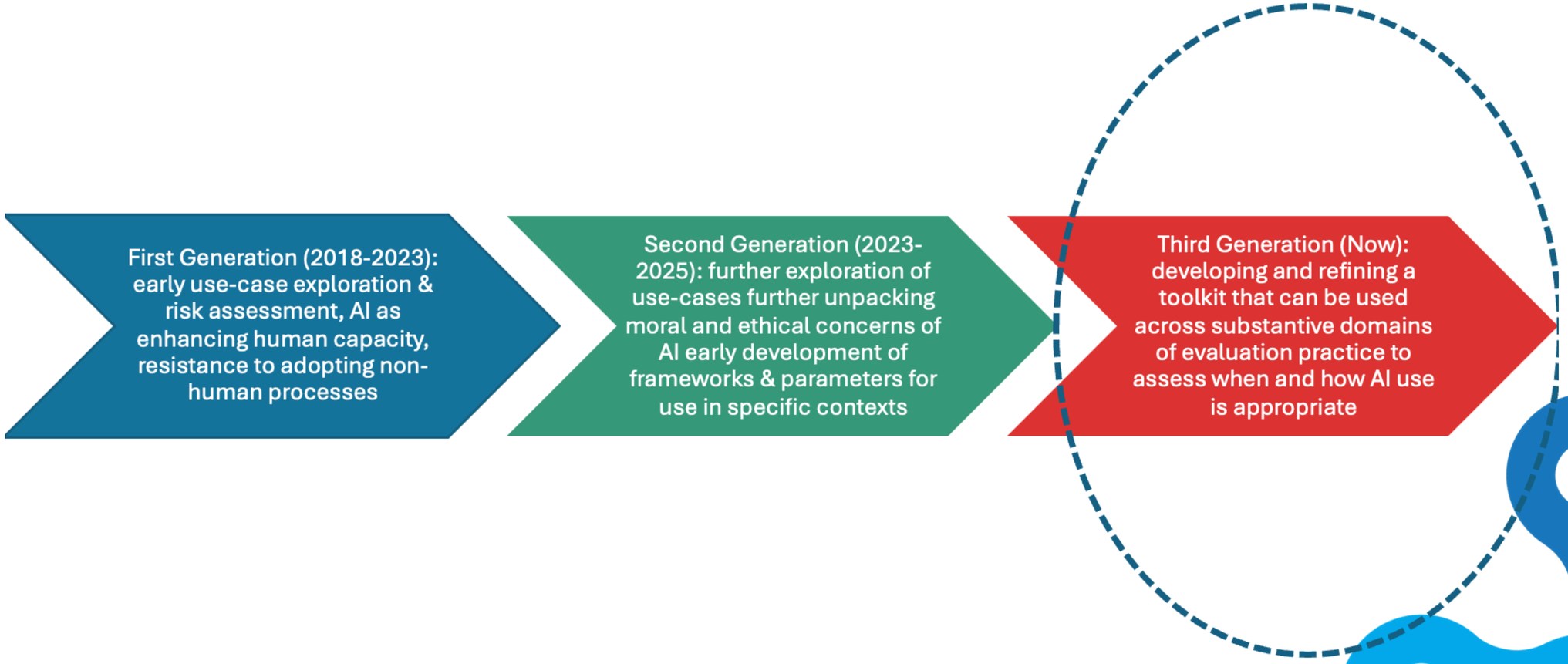
Literature on AI in Evaluation First and Second Generation: AI in Evaluation- AI Use Cases, Strengths, Limitations, and Ethical Concerns

First Generation (2018-2023): early use-case exploration & risk assessment, AI as enhancing human capacity, resistance to adopting non-human processes

Second Generation (2023-2025): further exploration of use-cases further unpacking moral and ethical concerns of AI early development of frameworks & parameters for use in specific contexts

Where are we now?

Literature on AI in Evaluation The Third Generation: AI in Evaluation Decision-Making Framework (AIDEM)



The Rationale for an AI in Evaluation Decision Making Framework (AIDEM)

•Decision-making (DM) (frameworks are tools designed to support individuals in navigating complex decisions (Forester-Miller & Davis, 1995; Manson, 2012; Wilkens, 2011)

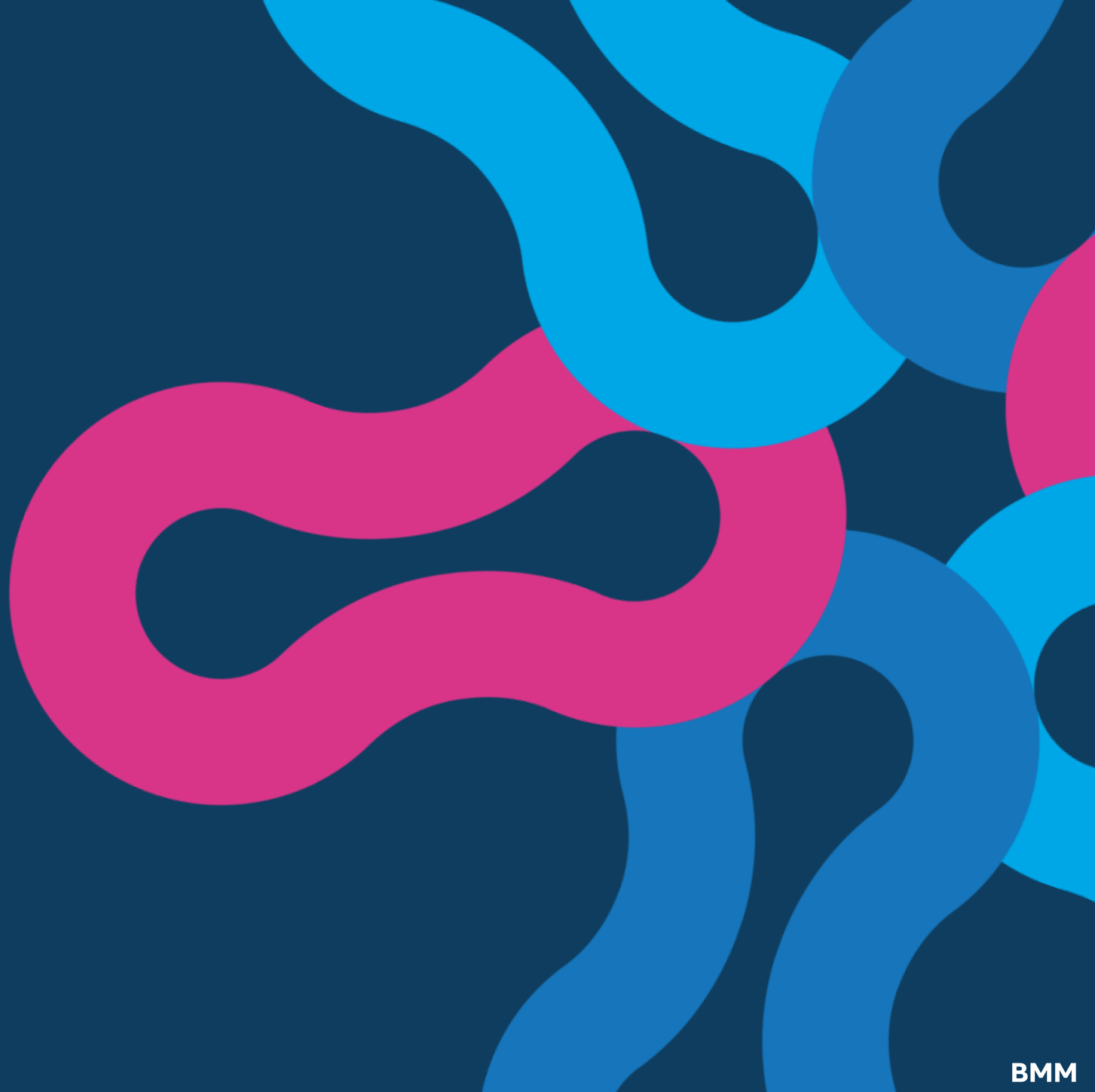
•DM frameworks are often used to aid decision-makers in:

(a) defining the purpose and scope of a decision,
(b) exploring alternate decision pathways,
(c) articulating values or criteria that will be used to guide the decision, and
(d) exploring tradeoffs between the different options

• DM frameworks align with Scriven's (1981) logic of evaluation, with its focus on defining criteria, setting standards, measurement, and synthesis.

•AIDEM shifts the conversation away from the pros and cons of AI use in evaluation, to underlying question should we use AI and the values determining AI use or non-use.

Introducing the Framework



BMM

AI in Evaluation: Decision Making Framework

01

Frame the Question

Make the question as specific as possible. For example: Should we use AI tool A, AI tool B, or no AI at all?

02

Articulate your Values

Set the criteria you will use to guide decision making. Whose values are represented in these criteria?

03

Find the Facts

Investigate your options. What do you know about your AI options? What don't you know? Gather the evidence and data you need.

04

Synthesize the Evidence

How do your AI options align with your values?

05

Final Decision

Meet to review the evidence and come to a final decision

06

Reflect & Follow Up

Check back in a few months after the decision. What have you learned? What might this mean for future decisions?

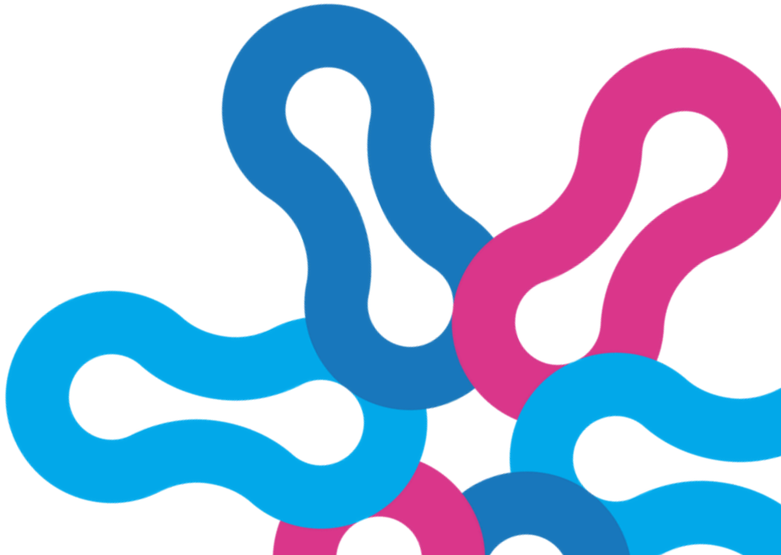
Steps 1 & 2

Step 1: Frame the Question

- Overarching question: Should AI be used? What AI tool should be used?
  - Consideration when framing the question: It should be framed within the context of the operational environment and be culturally responsive.
- Objective: Clearly define the decision to be made "To use or not to use AI"? Why and in what context/capacity?

Step 2: Articulate Your Values

- Overarching question: What criteria for making the decision should be used?
- Objective: Set criteria for decision-making, which includes whose values are being considered in the decision making process.
  - Montrosse-Moorhead's (2023) criteria
  - Other possible criteria examples:
    - Transparency & Accountability
    - Validity & Reliability
    - Cost-Effectiveness
    - Fairness & Inclusivity
    - Data Protection & Privacy
    - Environmental Stewardship



Montrosse-Moorhead's (2023) Criteria

Category	Domain	Essential question(s)	Justification for inclusion from articles included in this NDE issue
Domains that address the conceptualization and implementation of AI in evaluation practice	**Design & implementation**	How well does the evaluation plan for using AI align with evaluation context? Is AI used in evaluation well-conceived and well executed, from inception of idea to planned dissemination efforts?	Thornton's (2023) review of the Hsiao et al. (2021) qualitative framework and Rosenfeld (2021) quantitative framework for evaluating AI. Nielsen (2023) discussion of appropriateness of the technology and nature of the evaluation service, and capability of the evaluator. Ferretti (2023) discussion and illustration of "hacking by the prompt" for evaluators.
	Efficiency of process	Was AI use in evaluation more efficient than alternatives?	Thornton's (2023) review of the Hsiao et al. (2021) qualitative framework for evaluating AI. Nielsen (2023) discussion of competitive strategies, and size and duration of contracts. Head et al.'s (2023) discussion of estimating potential efficiency improvements for evaluation resulting from use of large language models. Ferretti (2023) discussion and illustration of "hacking by the prompt" for evaluators. Sabarre et al.'s (2023) case example of use of three AI tools in evaluation practice.
	Equity of process	Does AI use in evaluation attend to racial, ethnic, gender, and other inequities? Does AI use in evaluation advance equity?	Reid's (2023) ethics and equity framework for AI use in evaluation. Head et al.'s (2023) discussion of limitations and ethical challenges with use of large language models in evaluation. Sabarre et al.'s (2023) responsible AI use manifesto. Tilton et al.'s (2023) discussion of how AI tools to help bridge language access and equity.
Domains that address outcomes from using AI in evaluation	**Effectiveness**	Are the evidence, interpretations, and narratives resulting from AI use in evaluation considered good among key ecosystem actors?	Thornton's (2023) review of the Hsiao et al. (2021) qualitative framework, Rosenfeld (2021) quantitative framework, and Lin et al. (2020) mixed method framework for evaluating AI.
	Trust	Are the evidence, interpretations, and narratives resulting from AI use in evaluation considered trustworthy among key ecosystem actors?	Thornton's (2023) review of the Hsiao et al. (2021) qualitative framework and the Lin et al. (2020) mixed method framework for evaluating AI. Head et al.'s (2023) discussion of limitations and ethical challenges with use of large language models in evaluation. Sabarre et al.'s (2023) case example of use of three AI tools in evaluation practice, and their responsible AI use manifesto.
	Methodological validity & trustworthiness	Are claims generated during an evaluation using AI valid or trustworthy?	Azzam's (2023) discussion of validity types and the ability of evaluators versus AI to support them. Head et al.'s (2023) discussion of limitations and ethical challenges with use of large language models in evaluation. Sabarre et al.'s (2023) case example of use of three AI tools in evaluation practice, and their responsible AI use manifesto. Tilton et al.'s (2023) discussion of how chatbots built performed.
	Understandability	Does AI use in evaluation result in a sufficient level of understanding?	Thornton's (2023) review of the Hsiao et al. (2021) qualitative framework for evaluating AI.
	Equity of resulting information & evidence	Does AI-produced information and evidence attend to racial, ethnic, gender, and other inequities? Does AI-produced information and evidence advance equity?	Reid's (2023) ethics and equity framework for AI use in evaluation. Head et al.'s (2023) discussion of limitations and ethical challenges with use of large language models in evaluation. Sabarre et al.'s (2023) responsible AI use manifesto. Tilton et al.'s (2023) discussion of how AI tools to help bridge language access and equity.

Note. Inspired by and adapted from Teasdale, Pitts et al. (2023) and Teasdale, Strasser et al. (2023).

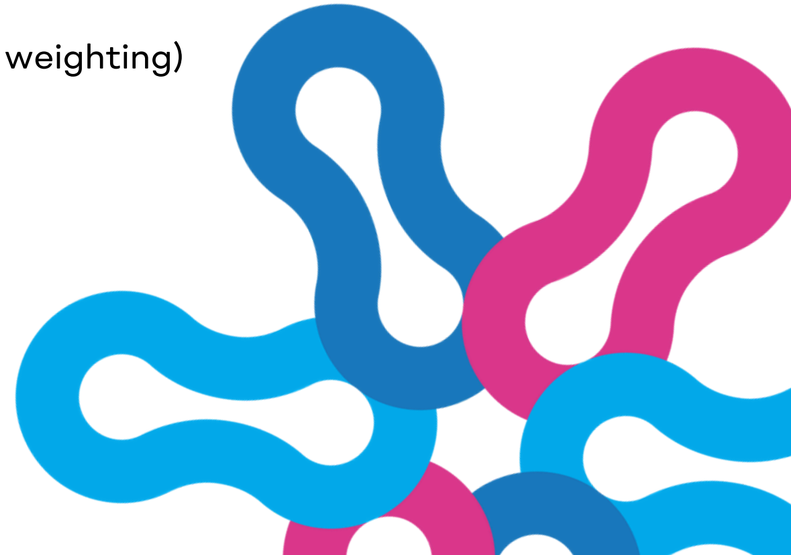
Steps 3 & 4

Step 3: Find the Facts

- Overarching question: What do and don't you know about your options?
- Objective:
  - Investigate the available AI options
    - Features, performance metrics, compliance, environmental impact, use cases.
  - Investigate a "No AI Option"
    - Traditional methods, its effectiveness, how it can be improved to serve the evaluation
  - Questions to Explore when gathering data:
    - What specific functionalities do the tools provide?
    - How do they align with your values?

Step 4: Synthesize the Evidence

- Overarching question: How do your AI options align with your values?
- Objective:
  - Determine the synthesis method to use (e.g., rubrics, qualitative weight and sum, numeric weighting)
  - Compare and contrast AI options and Non AI against criteria.
    - Montrosse-Moorhead's (2023) criteria
    - Other possible criteria examples:
      - Transparency & Accountability: How does each score?
      - Data Protection & Privacy: Compliance and security score
      - Environmental Stewardship: Energy consumption analysis score



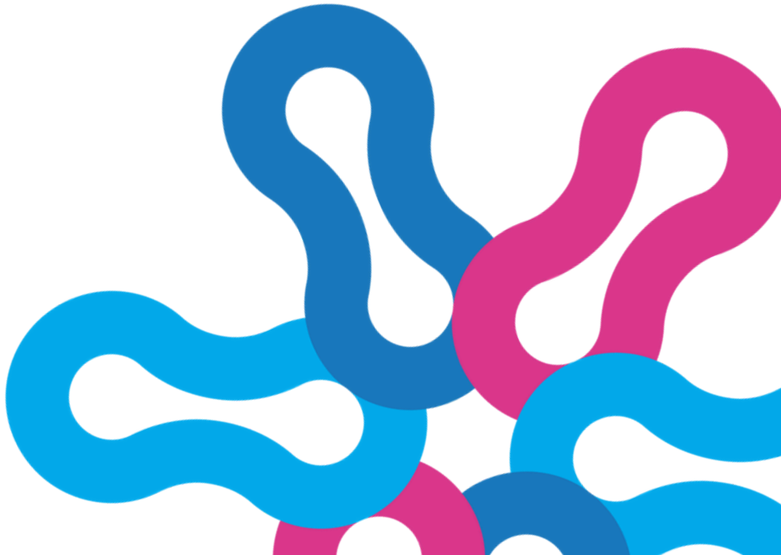
Steps 5 & 6

Step 5: Make a Final Decision

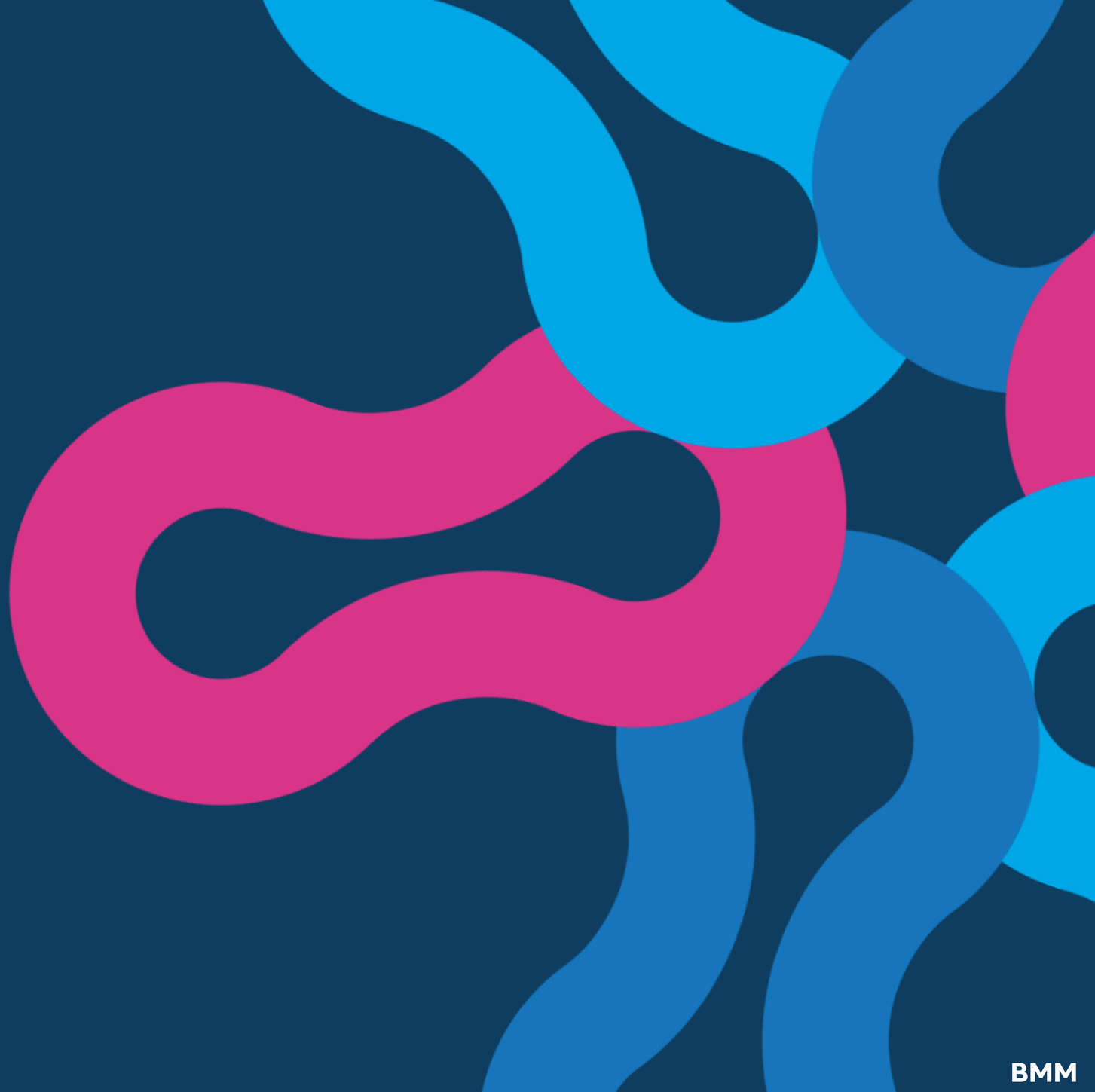
- Overarching question: What decision should we make?
- Objective: Make an evaluative judgement.
  - Review Process should include:
    - Meeting with interest holders, especially those with power and influence to make decisions and those that have power and influence over the implementation of the evaluation
    - Discuss the synthesized evidence
    - Achieve consensus on the final decision. Consensus for decision-making should be clearly defined
  - Documentation:
    - Record all viewpoints for transparency

Step 6: Reflect & Follow-up

- Overarching question: What have we learned? What does it mean for future decision?
- Objective: Reflective practice.
  - Determine a follow-up schedule that will allow sufficient time to review decision
  - Discuss lessons learnt or insights gained.
    - What worked well?
    - Unexpected challenges and learnings?
    - Unexpected external events, e.g., policies that affect the final decision made?
  - Future Implications:
    - What does this mean for future use?
    - How will this decision-making process inform future AI use?



AIDEM in Practice: A Worked Example



BMM

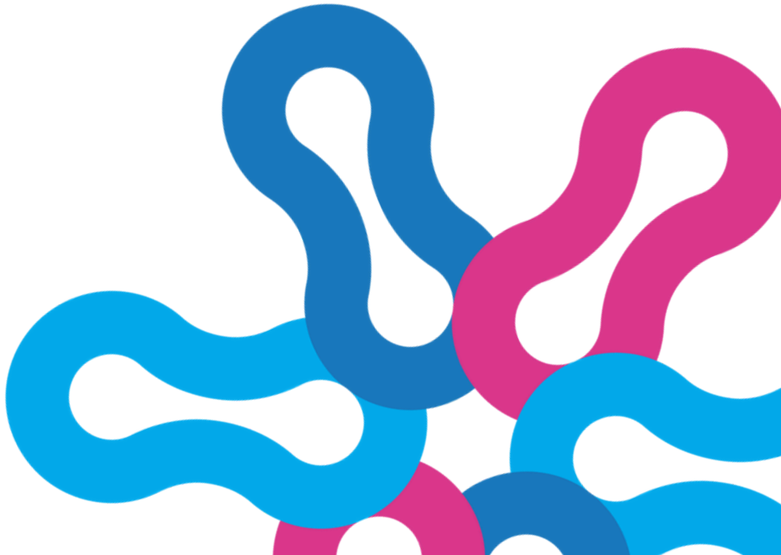
The Case

- 5-year evaluation of a U.S. federally-funded centre
- System-level change
- Centre activities focused on providing three tiers of technical assistance: universal, targeted, and intensive
- Both an internal and external evaluation team

A New Requirement, A Tight Timeline

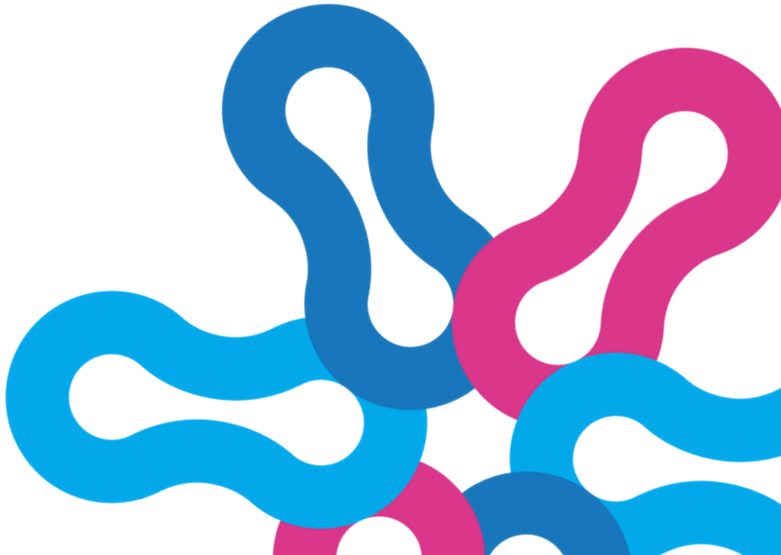
- 2.5 years into the project, the funder adds a new requirement
- Measure pre-to-post changes in capacity for technical and adaptive change
- 2-person external evaluation team
- 4 months to prepare
- No existing team/systems-level instrument
- Survey questions would be added to the state team self-assessments

Should we use an AI tool to help?



Step 1: Frame the Question

- The question: Should we use Microsoft 365 Copilot to help develop questions on systems-grounded technical and adaptive change to add to the state team self-assessments?
- Alternative: The usual evaluator-only survey question development process,



Step 2: Articulate your Values

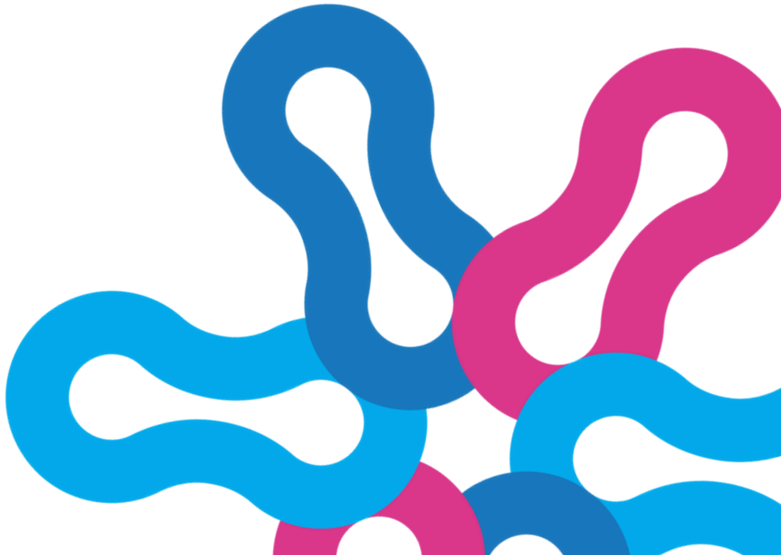
Criterion	Criterion Definition	Which Interest-Holders Hold These Values				
		External Evaluation Team	Funder	Centre	States	Families of Infants & Toddlers with Disabilities
Design & implementation	Use of AI, including prompts, must be fit-for-purpose, meaning that it is well-aligned with the evaluation context	✓	✓	✓	?	?
Efficiency of process/cost-effectiveness	Use of AI for this purpose must be more efficient than not using AI	✓	✓	✓	?	?
Human subjects protections	Use of AI for this purpose must uphold human subjects protections and relevant ethical principles	✓	✓	✓	?	?
Equity of process	Use of AI does not reinforce or make worse existing racial disparities in early childhood personnel systems	✓	?	✓	✓ *Varies by state*	?
Environmental stewardship	Use of AI attends to the environmental impacts on land, air, and water	✓	?	?	✓ *Varies by state*	?
Accountability/ effectiveness	Self-assessment questions resulting from the use of AI are likely to be considered good among ecosystem interest-holders	✓	?	?	?	?
Trust	Self-assessment questions resulting from the use of AI are likely to be considered trustworthy among ecosystem interest-holders	✓	?	?	?	?
Methodological validity & trustworthiness	Self-assessment questions resulting from the use of AI are likely to be considered methodologically valid and trustworthy	✓	✓	✓	✓	?
Understandability/ transparency	Self-assessment questions resulting from the use of AI are likely to result in a sufficient level of understanding	✓	✓	✓	?	?
Equity of resulting information & evidence	Self-assessment questions resulting from the use of AI will not reinforce or make worse existing racial disparities in early childhood personnel systems	✓	?	✓	✓ *Varies by state*	?

? Whether this group holds these values is unknown at the present moment.

(Criteria sources: Higdon & Raftree, 2025; Kwong, et al., 2025; Montrosse-Moorhead, 2023)

Step 3: Find the Facts

- Method: SWOT analysis
- Strengths / Weaknesses = inside the evaluation
- Opportunities / Threats = beyond the evaluation



Step 3: Find the Facts (Strengths & Weaknesses)

Strengths	Explanation	Criterion Alignment
Efficiency	Copilot can rapidly generate and iterate on self-assessment operational definitions and aligned questions, using all the documents identified to date, thereby reducing time and labor costs.	Efficiency of process/cost-effectiveness
Transparency & documentation	Copilot-produced definitions and questions can be logged by the external evaluation team, the funder, and the centre, including the internal evaluation team, to document the process. These same interest-holders can then review this log for transparency.	Understandability/transparency
Rapid question development	Copilot can generate multiple versions of operational definitions and aligned questions, enabling faster iteration and refinement of operational definitions and aligned questions by those involved.	Understandability/transparency, methodological validity & trustworthiness
Data privacy and security	Copilot policies on data, privacy, and security are publicly available (Microsoft, 2026). Moreover, the university contract with Microsoft requires users to sign in with their university credentials. Once this happens, all prompts, information retrieved, and responses generated remain within the university's service boundary and are therefore not used to train the underlying LLMs.	Human subjects protections, trust

Weaknesses	Explanation	Criterion Alignment
Contextual nuance	Copilot may miss or overlook subtle, context-specific dynamics of early childhood systems or state-level variations that might need to be accounted for in the questions it develops (Qamar et al., 2024).	Design & implementation, equity of process
Equity risks	Without careful oversight by the evaluation team, Copilot may replicate existing racial biases in the LLMs and/or training dataset, which could lead to these racial disparities being reinforced in the questions developed (Ashwin et al., 2025).	Equity of process, equity of resulting information & evidence
Perception of trust	If interest-holders value human judgment and distrust AI-generated content, they may not trust operational definitions and aligned questions.	Accountability/effectiveness, trust
Ethical oversight	Without careful oversight by the evaluation team, the process of developing operational definitions and aligned questions may not align with human subjects protections and the AEA Guiding Principles for Evaluators (or, in the Canadian Context, the CES' Guidance for Ethical Evaluation Practice)	Human subjects protections

Step 3: Find the Facts (Opportunities & Threats)

Opportunities	Explanation	Criterion Alignment
AI capacity building in evaluation	Copilot can help the external evaluation team to build AI literacy capacity in the context of evaluation work among the funder and the centre, including the internal evaluation team.	Design & implementation
Case example for the Field	The use of Copilot in this way, if well-executed and shared publicly, can serve as an example from which other evaluators can learn as they work to explore how to use AI in evaluation practice.	Not aligned to any identified value

Threats	Explanation	Criterion Alignment
Permissions	None of the potential documents to be used in developing operational definitions and items contains sensitive, confidential, or embargoed information (such as personally identifying information that cannot be shared under U.S. legislative requirements). Even so, the team did not have pre-existing guidelines for determining the acceptable and unacceptable uses of AI.	Ethics
Whose values are represented	The values assessment has missing information, which leads to uncertainty about the extent to which the final product(s) might be accepted by these groups.	Ethics
Values conflict	For states, there are competing values. For example, some states and organizations within them (e.g., school boards) have implemented laws and policies to restrict the use of equity work. Thus, some states cannot endorse this value, while others can. Such conflicting values may lead to disagreement, lack of use, or even avoidance of the items generated from this underlying value.	Ethics, equity of process, equity of resulting information & evidence
Environment	AI requires a significant amount of energy to run data centres (Rovner et al., 2025). Coal is a primary source of energy and a significant contributor to global warming in the U.S. AI also requires vast amounts of water for cooling its hardware systems, which can strain local water supplies and disrupt local ecosystems (Kwong et al., 2025).	Environmental stewardship

Step 4: Synthesize the Evidence (Crosswalk)

- First move: crosswalk each value (from step 2) to where it appeared across SWOT (from step 3)

Criterion	Strengths	Weaknesses	Opportunities	Threats
Design & implementation	✓ (1)	✓ (1)	✓ (1)	
Efficiency of process/cost-effectiveness	✓ (1)			
Human subjects protections	✓ (1)	✓ (1)		
Equity of process		✓ (2)		✓ (1)
Environmental stewardship				✓ (1)
Accountability/effectiveness	✓ (2)	✓ (1)		
Trust	✓ (2)	✓ (1)		
Methodological validity & trustworthiness	✓ (2)	✓ (1)		
Understandability/ transparency	✓ (2)			
Equity of resulting information & evidence		✓ (1)		✓ (1)

Step 4: Synthesize the Evidence (QW&S)

- Second move: Qualitative Weight-and-Sum (QWS)
- Developed rating scale
- Set weights (using rating scale)
- Finally, rating given for each criterion based on the evidence on the prior slide

Rating scale

- ● Copilot excels on this value (all strengths and opportunities, no weaknesses or threats)
- ◇ Copilot performs well on this value, with few limitations (mostly strengths and opportunities) with no more than 2 weaknesses and threats
- ◆ Copilot performs satisfactorily on this value (some strengths and opportunities) and poses 2 weaknesses or threats in this area.
- ✕ Copilot performs very poorly on this value (few strengths and opportunities) and poses 3+ weaknesses or threats in this area.
- – Value not identified for a particular S, W, O, or T dimension

Criterion	Criterion Importance Weight	Copilot
Design & implementation	●	◇
Efficiency of process/cost-effectiveness	◇	◇
Ethics	●	✕
Equity of process	●	✕
Environmental stewardship	◇	✕
Accountability/effectiveness	◇	◇
Trust	●	◇
Methodological validity & trustworthiness	●	◇
Understandability/transparency	●	●
Equity of resulting information & evidence	●	✕
Sum	● (7) ◇ (3)	● (1) ◇ (5) ✕ (4)

Step 4: Synthesize the Evidence (QW&S)

- Second move: Qualitative Weight-and-Sum (QWS)
- Developed rating scale
- Set weights (using rating scale)
- Finally, rating given for each criterion based on the evidence on the prior slide

Rating scale

- ● Copilot excels on this value (all strengths and opportunities, no weaknesses or threats)
- ◇ Copilot performs well on this value, with few limitations (mostly strengths and opportunities) with no more than 2 weaknesses and threats
- ◆ Copilot performs satisfactorily on this value (some strengths and opportunities) and poses 2 weaknesses or threats in this area.
- ✕ Copilot performs very poorly on this value (few strengths and opportunities) and poses 3+ weaknesses or threats in this area.
- – Value not identified for a particular S, W, O, or T dimension

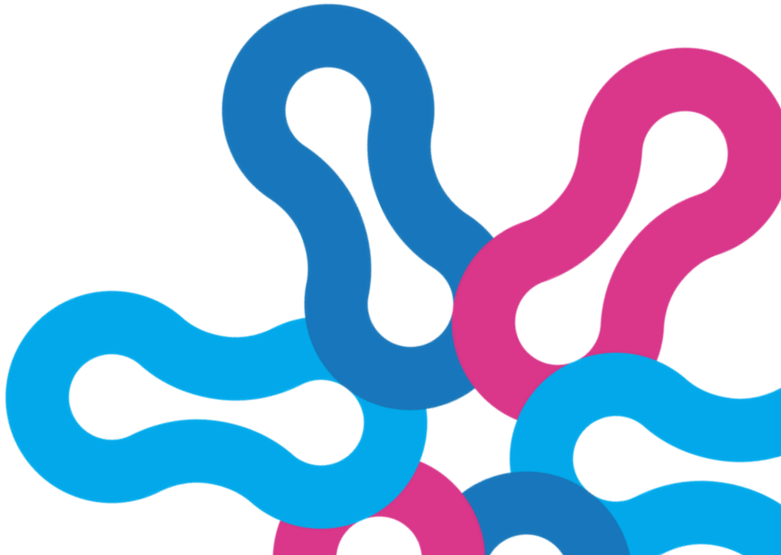
Criterion	Criterion Importance Weight	Copilot
Design & implementation	●	◇
Efficiency of process/cost-effectiveness	◇	◇
Ethics	●	✕
Equity of process	●	✕
Environmental stewardship	◇	✕
Accountability/effectiveness	◇	◇
Trust	●	◇
Methodological validity & trustworthiness	●	◇
Understandability/transparency	●	●
Equity of resulting information & evidence	●	✕
Sum	● (7) ◇ (3)	● (1) ◇ (5) ✕ (4)

Step 5: Final Decision

- Benefits cluster inside the evaluation; harms cluster beyond it
- Decision: do *not* use Microsoft 365 Copilot for this task
- "...low ratings on criteria related to ethics, equity, and environmental stewardship were too significant to outweigh all the areas where Copilot performed well."

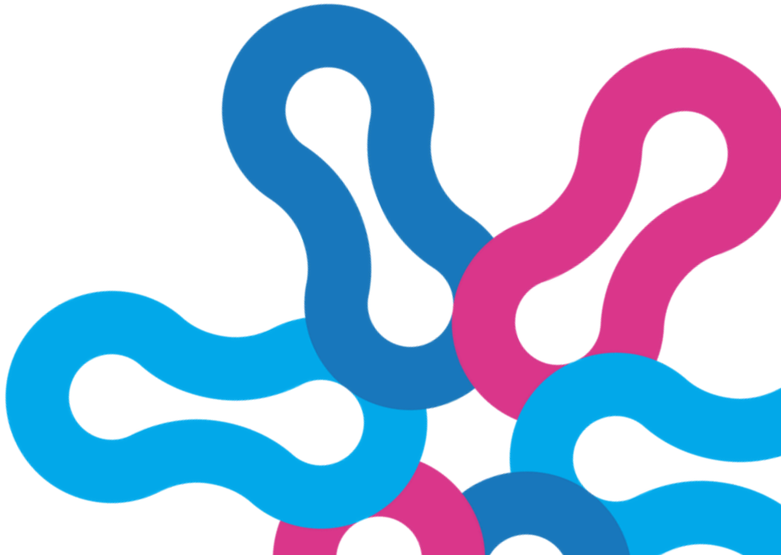
Step 6: Reflect & Follow-Up

- Reflect & Follow-Up: scheduled, not yet possible at time of writing *(decision was current)*
- Added ~10 hours:
  - SWOT ~4h
  - QWS ~4h
  - Decision meeting ~1h
- Without AIDEM: efficiency would have been the default driver

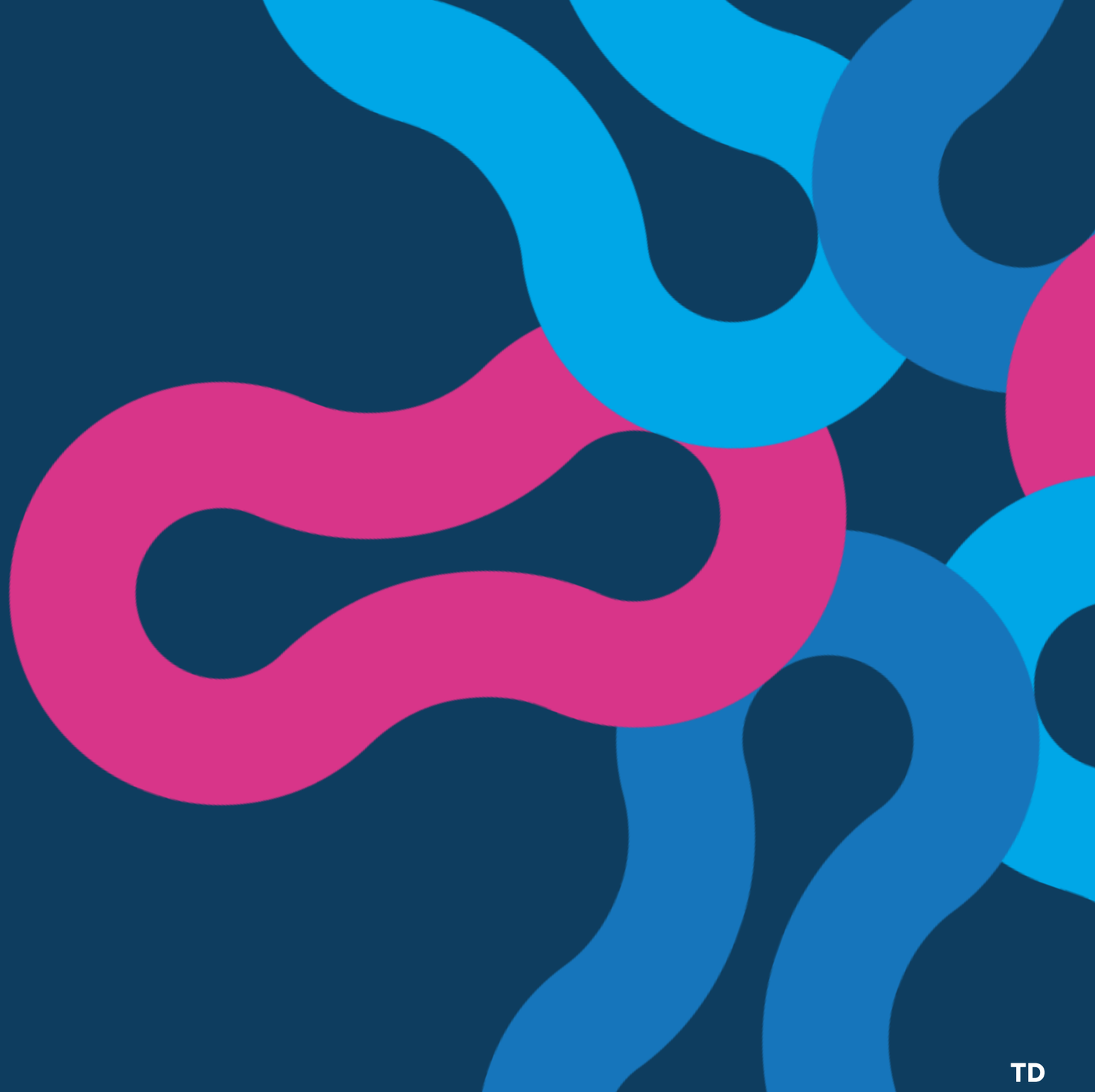


Takeaway's

- A defensible "no" is still a result
- Make the trade-offs explicit and values-grounded
- "...a more values-informed approach to decision-making around AI in evaluation



Q&A and Open Discussion



What questions or reflections do you have on the framework or use case?

Come off mute or post them in the chat!

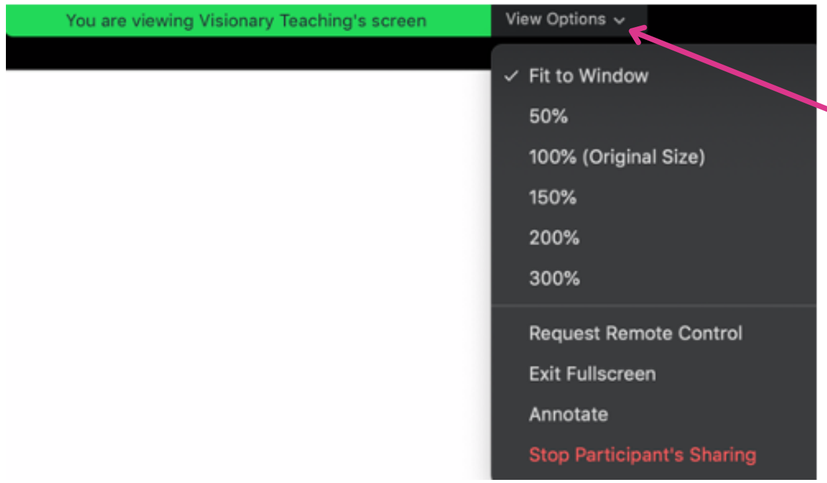
Think about your own AI-supported work…

What criteria (values) are missing from this list?

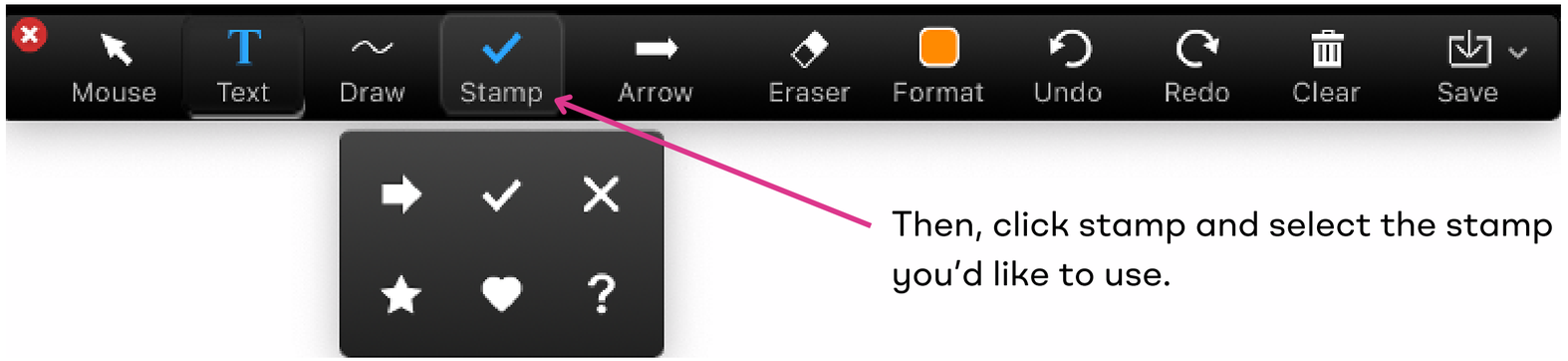
Come off mute or post them in the chat!

Criterion	Criterion Definition
Design & implementation	Use of AI, including prompts, must be fit-for-purpose, meaning that it is well-aligned with the evaluation context
Efficiency of process/cost-effectiveness	Use of AI for this purpose must be more efficient than not using AI
Human subjects protections	Use of AI for this purpose must uphold human subjects protections and relevant ethical principles
Equity of process	Use of AI does not reinforce or make worse existing racial disparities in early childhood personnel systems
Environmental stewardship	Use of AI attends to the environmental impacts on land, air, and water
Accountability/ effectiveness	Self-assessment questions resulting from the use of AI are likely to be considered good among ecosystem interest-holders
Trust	Self-assessment questions resulting from the use of AI are likely to be considered trustworthy among ecosystem interest-holders
Methodological validity & trustworthiness	Self-assessment questions resulting from the use of AI are likely to be considered methodologically valid and trustworthy
Understandability/ transparency	Self-assessment questions resulting from the use of AI are likely to result in a sufficient level of understanding
Equity of resulting information & evidence	Self-assessment questions resulting from the use of AI will not reinforce or make worse existing racial disparities in early childhood personnel systems

Before We Ask the Next Question: How to Use the "Stamper" Feature



First, click "view options".



Then, click stamp and select the stamp you'd like to use.

Think about your own AI-supported work…

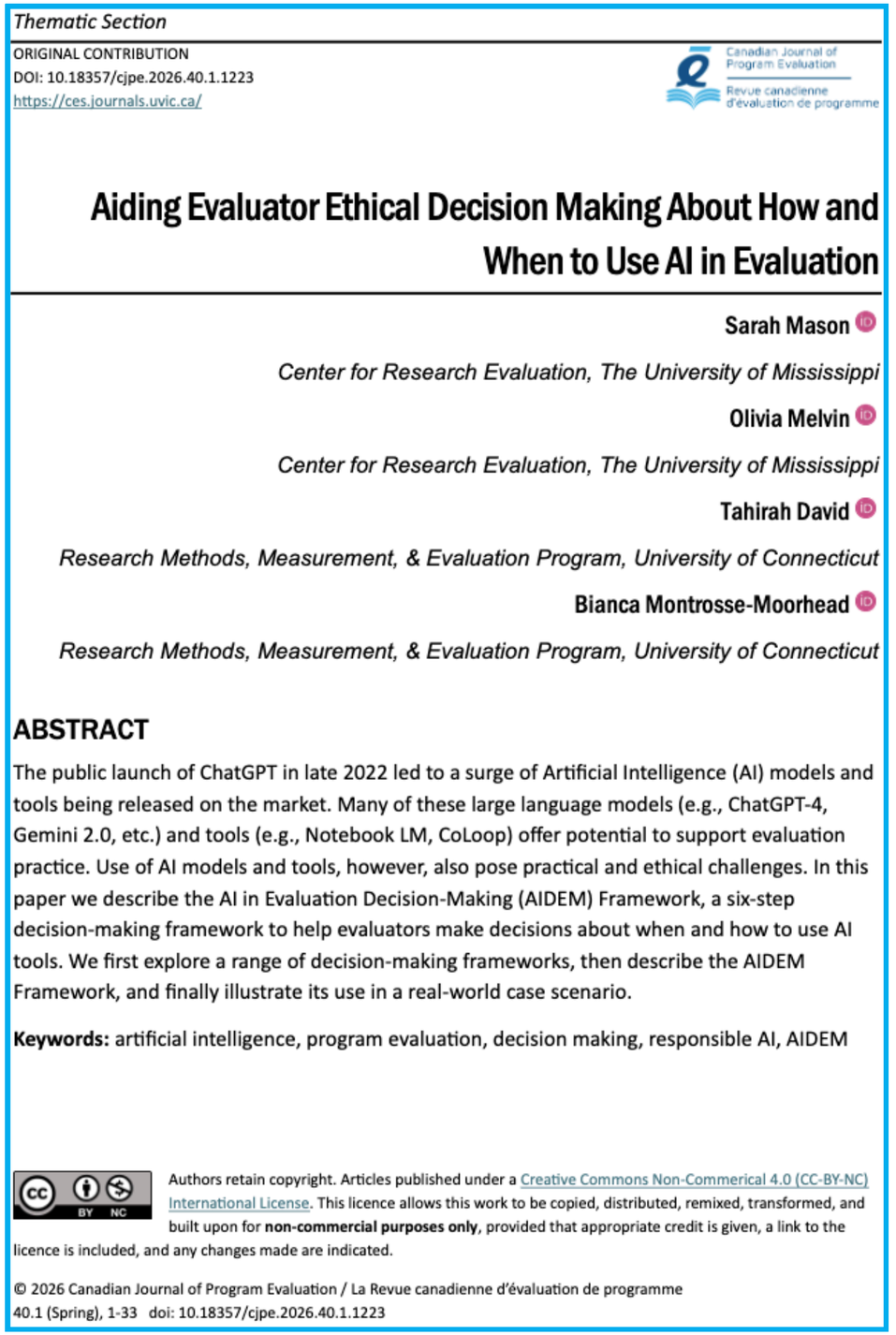
Which criteria (values) are most important to consider?

Use the stamper tool to make your answer visible!

Criterion	Criterion Definition
Design & implementation	Use of AI, including prompts, must be fit-for-purpose, meaning that it is well-aligned with the evaluation context
Efficiency of process/cost-effectiveness	Use of AI for this purpose must be more efficient than not using AI
Human subjects protections	Use of AI for this purpose must uphold human subjects protections and relevant ethical principles
Equity of process	Use of AI does not reinforce or make worse existing racial disparities in early childhood personnel systems
Environmental stewardship	Use of AI attends to the environmental impacts on land, air, and water
Accountability/ effectiveness	Self-assessment questions resulting from the use of AI are likely to be considered good among ecosystem interest-holders
Trust	Self-assessment questions resulting from the use of AI are likely to be considered trustworthy among ecosystem interest-holders
Methodological validity & trustworthiness	Self-assessment questions resulting from the use of AI are likely to be considered methodologically valid and trustworthy
Understandability/ transparency	Self-assessment questions resulting from the use of AI are likely to result in a sufficient level of understanding
Equity of resulting information & evidence	Self-assessment questions resulting from the use of AI will not reinforce or make worse existing racial disparities in early childhood personnel systems

Read & Share Our FREE Paper!

https://ces.journals.uvic.ca/index.php/cjpe/article/view/1223/1150



Thematic Section

ORIGINAL CONTRIBUTION
DOI: 10.18357/cjpe.2026.40.1.1223
https://ces.journals.uvic.ca/

Canadian Journal of Program Evaluation
Revue canadienne d'évaluation de programme

Aiding Evaluator Ethical Decision Making About How and When to Use AI in Evaluation

Sarah Mason

Center for Research Evaluation, The University of Mississippi

Olivia Melvin

Center for Research Evaluation, The University of Mississippi

Tahirah David

Research Methods, Measurement, & Evaluation Program, University of Connecticut

Bianca Montrosse-Moorhead

Research Methods, Measurement, & Evaluation Program, University of Connecticut

ABSTRACT

The public launch of ChatGPT in late 2022 led to a surge of Artificial Intelligence (AI) models and tools being released on the market. Many of these large language models (e.g., ChatGPT-4, Gemini 2.0, etc.) and tools (e.g., Notebook LM, CoLoop) offer potential to support evaluation practice. Use of AI models and tools, however, also pose practical and ethical challenges. In this paper we describe the AI in Evaluation Decision-Making (AIDEM) Framework, a six-step decision-making framework to help evaluators make decisions about when and how to use AI tools. We first explore a range of decision-making frameworks, then describe the AIDEM Framework, and finally illustrate its use in a real-world case scenario.

Keywords: artificial intelligence, program evaluation, decision making, responsible AI, AIDEM

40.1 (Spring), 1-33 doi: 10.18357/cjpe.2026.40.1.1223

global **evaluation** initiative

Thank you!

www.glocalevalweek.org