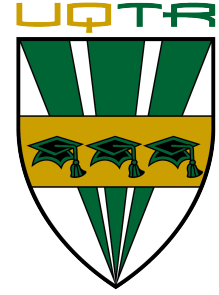
AI Governance in Academia

Guidelines for Generative AI

Clayton Peterson
Marie-Catherine Deschênes

Université du Québec à Trois-Rivières
clayton.peterson@uqtr.ca

38^{th} International FLAIRS Conference



Université du Québec
à Trois-Rivières
**Chaire de recherche UQTR
en éthique de l'intelligence artificielle**

Outline

Université du Québec à Trois-Rivières

Motivation

People tend to rely on generative AI for tasks that should be accomplished by experts.

⇒ e.g. grade papers, define academic objectives within a syllabus, generate ideas for funding proposals

Objective

Reflect on the governance of generative AI in academia and establish principles that should guide and restrain its use.

Plan of the presentation

1) Understanding governance
2) Judicial issues
3) Epistemic issues
4) Governance principles

Understanding governance

Understanding governance

Université du Québec à Trois-Rivières

Current guidelines are presented as if they could...

...ensure an ethical use of AI

Government of Canada

...ensure a responsible use of AI

Government of Quebec

...provide a basis [...] for the ethical use of AI

United Nations

This is misleading and gives a false impression of ethical legitimacy.

⇒ As if any action respecting these guidelines could be qualified as *ethical* or *responsible*.

★ Respecting guidelines is necessary for ethical behavior, but not sufficient.

Understanding governance

Université du Québec à Trois-Rivières

Current guidelines are presented as if they could...

...ensure an ethical use of AI

Government of Canada

...ensure a responsible use of AI

Government of Quebec

...provide a basis [...] for the ethical use of AI

United Nations

This is misleading and gives a false impression of ethical legitimacy.

⇒ As if any action respecting these guidelines could be qualified as *ethical* or *responsible*.

★ Respecting guidelines is necessary for ethical behavior, but not sufficient.

Why?

Understanding governance

Université du Québec à Trois-Rivières

Current guidelines are presented as if they could...

...ensure an ethical use of AI

Government of Canada

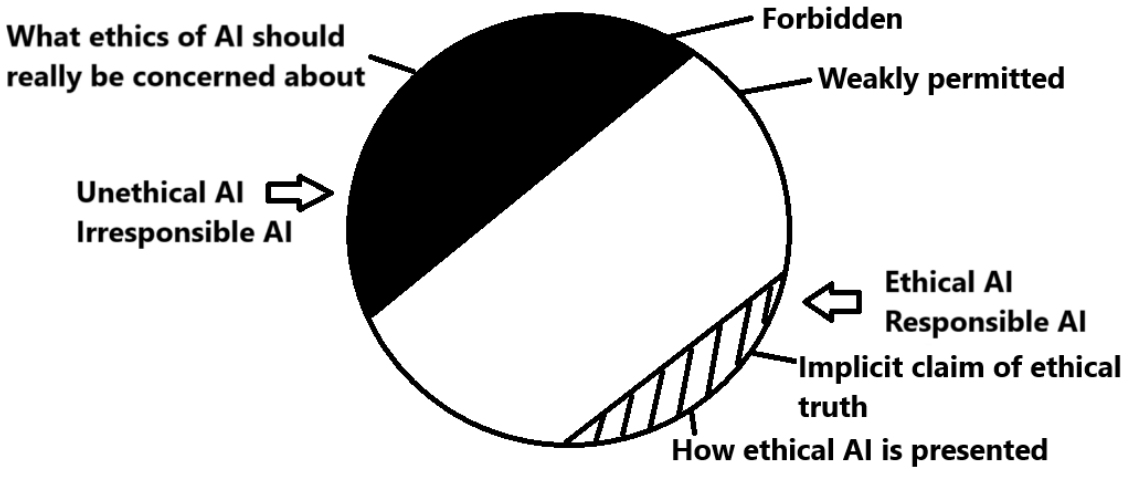
...ensure a responsible use of AI

Government of Quebec

...provide a basis [...] for the ethical use of AI

United Nations

Because this would require taking a stand on what is 'ethical'.



Governance **principles should restrict usage of generative AI**, not promote it.

Judicial issues

Judicial issues

Université du Québec à Trois-Rivières

There is currently no active legislation governing the use of AI in Canada.

There are nonetheless duties stemming from already existing and applicable legislation.

- ⇒ Copyright Act (RSC 1985, c. C-42 [Canada])
- ⇒ Act respecting Access to documents held by public bodies and the Protection of personal information (L.R.Q., c. A-2.1 [Quebec])

Judicial issues

Université du Québec à Trois-Rivières

Unauthorized use of input

Requirement to own the necessary rights on the input when using a generative AI tool.

- ⇒ one cannot use anything as input (e.g. data from patients)
- ⇒ availability or accessibility is not sufficient (e.g. available online)

Unauthorized use of output

Outputs generated by AI tools can infringe copyright if they reproduce a substantial part of an original work.

- ⇒ must be evaluated by its quality rather than its quantity
- ⇒ is not restricted to literal copying
  - ⋆ can result from an abstraction based on the work

★ The probabilistic nature of generative AI tools does not preclude them from infringing copyrights through output generation.

Judicial issues

Université du Québec à Trois-Rivières

Originality

All creations are not necessarily protected under the Copyright Act.

⇒ only applies to 'original work' resulting from an author's exercise of skills and judgment

⇒ usage of generative AI can have an impact on authorship and rights

Authorship

Authors are natural persons.

⇒ software, programs, algorithms and machines are not authors

Personal information

Information is considered *personal information* when it can be used to identify either directly or indirectly an individual.

⇒ e.g. profession, list of applicants, folder number, IP address

⇒ cannot be used without consent and needs to be protected

Epistemic issues

Epistemic issues

Université du Québec
à Trois-Rivières

Are generative AI tools trustworthy?

Trusting a source of information means not feeling a constant need for cross-validation.

Terms and conditions for chatbots and LLMs from OpenAI, Google, Meta and Anthropic, for instance, state that outputs...

- ...may not always be accurate
- ...should not be relied on without cross-validation
- ...should not be considered as substitutes for expertise

Meta goes as far as to write that one 'should not rely upon outputs for any purpose'.

Epistemic issues

Université du Québec à Trois-Rivières

Are generative AI tools trustworthy?

Trusting a source of information means not feeling a constant need for cross-validation.

Terms and conditions for chatbots and LLMs from OpenAI, Google, Meta and Anthropic, for instance, state that outputs...

- ...may not always be accurate
- ...should not be relied on without cross-validation
- ...should not be considered as substitutes for expertise

Meta goes as far as to write that one 'should not rely upon outputs for any purpose'.

So, are generative AI tools trustworthy?

Epistemic issues

Université du Québec à Trois-Rivières

[generative AI] can create original content

IBM

Can generative AI tools create new and original content?

Whether generative AI tools can be conceived as generating new content depends on what one means by *new*.

Epistemic issues

Université du Québec à Trois-Rivières

[generative AI] can create original content

IBM

Can generative AI tools create new and original content?

Whether generative AI tools can be conceived as generating new content depends on what one means by *new*.

1) Prediction of tokens based on input and probability distributions.

- ⇒ LLMs provide us with likely continuations of strings and sentences based on patterns learned from previous data
- ⇒ LLMs repeat information rather than create novel content

Epistemic issues

Université du Québec à Trois-Rivières

[generative AI] can create original content

IBM

Can generative AI tools create new and original content?

Whether generative AI tools can be conceived as generating new content depends on what one means by *new*.

1) Prediction of tokens based on input and probability distributions.

- ⇒ LLMs provide us with likely continuations of strings and sentences based on patterns learned from previous data
- ⇒ LLMs repeat information rather than create novel content

2) 'new' as not previously published

- ⇒ LLMs can provide new content in this sense
- ⇒ this is not, however, how novelty and originality is presented and conceived

Epistemic issues

Université du Québec à Trois-Rivières

[AI can make] sense of spoken language

Government of Canada

[AI can] understand human language [and replace] the need for human intelligence

IBM

3) A content is original not only when it occurs for the first time, but also when it occurs through a **creative process intended by an agent**.

⇒ new and original content is relevant not because of its first occurrence, but given the meaning associated with it through its creation process (intent)

However, generative AI tools are not meant to understand natural language.

⇒ there is no meaning associated with the output they produce
⇒ words do not have meaning in themselves

We, as humans, attribute meaning to words and sentences, and this meaning goes beyond the statistical relationships between words.

Epistemic issues

Université du Québec à Trois-Rivières

[AI can make] sense of spoken language

Government of Canada

[AI can] understand human language [and replace] the need for human intelligence

IBM

3) A content is original not only when it occurs for the first time, but also when it occurs through a **creative process intended by an agent**.

⇒ new and original content is relevant not because of its first occurrence, but given the meaning associated with it through its creation process (intent)

However, generative AI tools are not meant to understand natural language.

⇒ there is no meaning associated with the output they produce
⇒ words do not have meaning in themselves

We, as humans, attribute meaning to words and sentences, and this meaning goes beyond the statistical relationships between words.

Generative AI is **misconceived** as something that can create original content.

★ reinforces the anthropomorphization of AI and misleads people into thinking that creativity is to be understood on a par with human creativity

Epistemic issues

Université du Québec à Trois-Rivières

Are generative AI tools epistemically reliable?

1) Current pre-trained LLMs are not trained on established scientific facts.
   - ★ one should not expect the output of a LLM to come from a probability distribution of scientific facts
   - ★ the probability of the output will be based on the data the LLM was trained on

Epistemic issues

Université du Québec à Trois-Rivières

Are generative AI tools epistemically reliable?

1) Current pre-trained LLMs are not trained on established scientific facts.
   - ★ one should not expect the output of a LLM to come from a probability distribution of scientific facts
   - ★ the probability of the output will be based on the data the LLM was trained on

2) There is, however, a more fundamental epistemic issue relating to reliability.

Epistemic issues

Université du Québec à Trois-Rivières

Are generative AI tools epistemically reliable?

1) Current pre-trained LLMs are not trained on established scientific facts.

- ★ one should not expect the output of a LLM to come from a probability distribution of scientific facts
- ★ the probability of the output will be based on the data the LLM was trained on

2) There is, however, a more fundamental epistemic issue relating to reliability.

One indicator of epistemic reliability is **empirical adequacy**.

⇒ a source providing empirically inadequate information is epistemically unreliable

Epistemic issues

Université du Québec à Trois-Rivières

Are generative AI tools epistemically reliable?

1) Current pre-trained LLMs are not trained on established scientific facts.

- ★ one should not expect the output of a LLM to come from a probability distribution of scientific facts
- ★ the probability of the output will be based on the data the LLM was trained on

2) There is, however, a more fundamental epistemic issue relating to reliability.

One indicator of epistemic reliability is **empirical adequacy**.

⇒ a source providing empirically inadequate information is epistemically unreliable

Generative AI tools are known to yield errors, inaptly dubbed *hallucinations*.

⇒ errors can occur even when these tools are trained on empirically adequate data and prompted correctly

Epistemic issues

Université du Québec
à Trois-Rivières

Are generative AI tools epistemically reliable?

These errors can be understood (among other things) on the grounds of overfitting.

- ★ overfitting happens when one tries to maximize model fit and ends up fitting noise and peculiarities of the training data

Epistemic issues

Université du Québec à Trois-Rivières

Are generative AI tools epistemically reliable?

These errors can be understood (among other things) on the grounds of overfitting.

- ★ overfitting happens when one tries to maximize model fit and ends up fitting noise and peculiarities of the training data

Why do LLMs work? How are they able to provide convincing outputs?

Consider model performance in relation to their complexity.

	number of parameters
GPT-1	117 million
GPT-2	1.5 billion
GPT-3	175 billion
GPT-4	1.75 trillion

Epistemic issues

Université du Québec à Trois-Rivières

Are generative AI tools epistemically reliable?

These errors can be understood (among other things) on the grounds of overfitting.

★ overfitting happens when one tries to maximize model fit and ends up fitting noise and peculiarities of the training data

Why do LLMs work? How are they able to provide convincing outputs?

Consider model performance in relation to their complexity.

	number of parameters
GPT-1	117 million
GPT-2	1.5 billion
GPT-3	175 billion
GPT-4	1.75 trillion

⋆ recent studies have shown that LLMs are not predictively accurate when the data used as input does not share the same distribution as the training data

From an epistemic standpoint, overfitting thwarts predictive accuracy, which in turn is a proxy for empirical adequacy and, thus, for epistemic reliability.

⇒ i.e. more complex does not necessarily mean more reliable

Governance principles

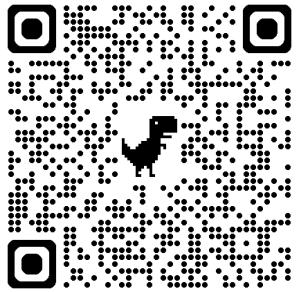
Governance principles

Université du Québec à Trois-Rivières

Scholars should be aware that generative AI tools:

1) Are not built to create theories, hypotheses or solve academic problems;
2) Require constant cross-validation;
3) Produce mistakes, and recognizing these mistakes requires appropriate expertise;
4) Do not provide us with an epistemically sound justification for knowledge;
5) Can be shielded from proper scientific investigation;
6) Can be trained on content that is itself generated;
7) Might have been trained on data used without consent or violating intellectual property;
8) Can affect originality as well as copyrights of scientific production;
9) Can violate legal norms;
10) Can violate universities' regulations.

Governance principles



Université du Québec à Trois-Rivières

As such, the following principles were proposed as guidelines to protect against irresponsible use of generative AI tools:

1) Usage of generative AI needs to be **authorized** by one's immediate superior;
2) Usage **respects legal** as well as **institutional norms**;
3) Usage of generative AI is properly **declared**;
4) Input and output **do not violate intellectual property**;
5) Input **does not contain any personal information**;
6) Research using human-related data has received **approbation from the ethics board**;
7) User understands how the tool works as well as its epistemic limitations;
8) Usage of generative AI tools is consistent with what these tools are (i.e., the objective aimed at when using the tool can indeed be reached by that tool; otherwise usage is instrumentally irrational);
9) Usage respects pedagogical alignment between learning objectives, pedagogical activities as well as evaluation strategies;
10) **Output is carefully and critically analyzed**;
11) User possesses the required expertise and competencies to fulfill the aforementioned principles.

Governance principles

Université du Québec à Trois-Rivières

As such, the following principles were proposed as guidelines to protect against irresponsible use of generative AI tools:

1) Usage of generative AI needs to be **authorized** by one's immediate superior;
2) Usage **respects legal** as well as **institutional norms**;
3) Usage of generative AI is properly **declared**;
4) Input and output **do not violate intellectual property**;
5) Input **does not contain any personal information**;
6) Research using human-related data has received **approbation from the ethics board**;

YOU are responsible!

7) User understands how the tool works as well as its epistemic limitations;
8) Usage of generative AI tools is consistent with what these tools are (i.e., the objective aimed at when using the tool can indeed be reached by that tool; otherwise usage is instrumentally irrational);
9) Usage respects pedagogical alignment between learning objectives, pedagogical activities as well as evaluation strategies;
10) **Output is carefully and critically analyzed**;
11) User possesses the required expertise and competencies to fulfill the aforementioned principles.

If you use generative AI, know what you are doing!

THANK YOU!

www.uqtr.ca/Chaire.Ethique.IA