Informed consent of this survey

INFORMATION AND CONSENT TO PARTICIPATE IN A RESEARCH STUDY

Study Title: Defining ethically sourced code generation

Lead Researcher: Anonymized for review

Lead Researcher's Contact Information: Anonymized for review

Student Researcher: Anonymized for review

Student Researcher's Contact Information: Anonymized for review

Ethical Approval Certification Number: 30021911

You are being invited to participate in the research study mentioned above. This form provides information about what participating would mean. Please read it carefully before deciding if you want to participate or not. If there is anything you do not understand, or if you want more information, please ask the researcher.

A. PURPOSE

The purpose of the research is to explore how stakeholders (e.g., developers, researchers) understand the **concept of ethically sourced code generation**, and collect various dimensions that they think should be included in its definition. In addition, the study aims to gather examples of situations where participants feel an ethically sourced model is needed, and to understand whether any existing tools are perceived to meet this ethical standard. To understand the potential consequences of unethically sourced code generation, the study also aims to investigate the potential impact of unethically sourced code generation. This study also seeks to assess participants'

willingness to accept potential trade-offs in performance, cost, and efficiency when developing ethically sourced code generation models.

B. PROCEDURES

If you participate, you will be asked to complete an online survey. The survey includes multiple-choice questions, open-ended (free-text) questions, and ranking tasks, and the questions are about background information, open-ended reflections, real-world experience, rating ethical dimensions, and ranking acceptable trade-offs. To assess relevant expertise in answering the survey questions, the background section only asks about professional experience and familiarity with code generation tools. **No identifying information is collected.**

In total, participating in this study will take **10–30 minutes.** This survey will be available until **July 15**.

C. RISKS AND BENEFITS

You might face certain risks by participating in this research.
 These risks include:
 • The survey may take longer to complete than initially expected. If you feel tired or need a break while filling out the survey, you are free to pause and return to it later.
Potential benefits include:
 • Gaining a deeper understanding of the concept of ethically sourced code generation and the ethical considerations involved in building AI-based code generation tools.
 • Reflecting on real-world examples and potential trade-offs may help clarify your own values or practices related to responsible technology development.
 • Understanding the potential consequences on unethically sourced code generation may help in raising awareness of the importance of ethically sourced code generation models.

D. CONFIDENTIALITY

We will gather the following information as part of this research: ·Your responses to survey questions related to your opinions, experience, and judgments about ethically sourced code generation.

The survey includes multiple-choice questions, open-ended (free-text) questions, and ranking tasks. We will not allow anyone to access the information, except people directly involved in conducting the research. We will only use the information for the purposes of the research described in this form. The information gathered will be **anonymous**. That means that it will not be possible to make a link between you and the information you provide.

We ask that you **do not include any personal identifying information in your open-ended responses.** Our research team will review all submissions to check for any accidentally shared sensitive information, and will remove any content that could reveal your identity before analyzing the data.

We will protect the information by hosting the survey on Qualtrics, a secure, university-supported platform. Only the primary researcher has access to the Qualtrics account. After the survey is completed, the data will be downloaded and securely stored on password-protected and access-restricted university-issued devices. Co-investigators will only be granted access to the collected data with the permission of the PI, and solely for the purposes of this research.

We intend to publish the results of the research. However, it will not be possible to identify you in the published results.

We will destroy the information five years after the end of the study.

E. CONDITIONS OF PARTICIPATION

You do not have to participate in this research. It is purely your decision. If you do

participate, you can stop at any time. **Due to the anonymous nature of the survey, however, responses cannot be withdrawn after submission.**

As a compensatory indemnity for participating in this research, you will have the opportunity to enter a draw. For every 50 participants, we will randomly select 5 lucky winners to receive a **$50 CAD Amazon gift card. Participants must provide their email address to enter the draw but the contact information will be stored separately from the responses to maintain anonymity.**

To make sure that research money is being spent properly, auditors from (Anonymized for review) or outside will have access to a coded list of participants. It will not be possible to identify you from this list. There are **no negative consequences for not participating or stopping in the middle.**

If you are a student of (Anonymized for review) University, be aware that the participation in this survey bears no consequence to your course performance and will not impact your grade.

Inclusion Criteria:

· Adults aged 18 and over.

· Participants should have at least one year of programming experience.

· Individuals with at least some familiarity with code generation models (as a user, developer, or researcher).

· Individuals not affiliated with our own research group (Anonymized for review)

F. PARTICIPANT'S DECLARATION

By selecting "**Agree and proceed to the survey**", you confirm that you have read and understood this form. You have had the chance to ask questions and any questions have

been answered. You voluntarily agree to participate in this research under the conditions described.

If you **disagree, do not want your data to be used or do not want to participate**, you should close the web browser before completing the survey and the incomplete surveys will be deleted.

If you have questions about the scientific or scholarly aspects of this research, please contact the researcher. Their contact information is on page 1. You may also contact their faculty supervisor.

If you have concerns about ethical issues in this research, please contact (Anonymized for review).

○ Agree and proceed to the survey.

Background (no identifiable info involved)

Important Note
This survey is divided into multiple sections. Once you proceed to the next section, you will **not be able to return** to or revise your previous responses. A **progress bar** is provided at the top to help you track your progress. Please take your time — we appreciate your thoughtful input.

Background Information
This section aims to understand your professional

background and your familiarity with code generation and ethically sourced products. These insights will help us better understand your responses in later sections, which will focus on defining and evaluating challenges related to ethically sourced code generation.

Reminder

In the context of this survey, **"code generation models"** refer specifically to AI-based or LLM-based systems (e.g., GitHub Copilot, ChatGPT, Codex). The term **"ethically sourcing"** refers to practices of ensuring that products are created responsibly and sustainably.

Which one of the following describes your role in daily work best?

○ Software developer

○ AI / ML engineer

○ Data analyst

○ Researcher (industrial, professor, research assistant, postdoctoral)

○ Software tester

○ Engineering Manager / Tech Lead

○ Product Manager / Designer

○ Funder

○ Student

○ [   ] Other (please specify):

How many years of programming/coding experience do you have?

[   ]

How familiar are you with code generation models?

○ Never used or heard of them
○ Heard of them but never used
○ Used through a chat interface (e.g., ChatGPT)
○ Used through an API or basic integration
○ Participated in training, fine-tuning, or full model development

How many code generation models are you familiar with (e.g., ChatGPT)?

[   ]

How frequently do you interact with code generation

models (using, working with, or developing them)?

○ At least once a day

○ At least once a week

○ At least once a month

○ At least once a year

○ Never

How many years of experience do you have working with AI-based or LLM-based code generation tools/models (including API use, model training, or fine-tuning)?

How familiar are you with the concept of "ethically sourced" products (e.g., ethically sourced coffee, clothing)?

○ Not familiar – I have not heard of this concept before

○ Slightly familiar – I have come across the term, but don't know much about it

○ Somewhat familiar – I have heard of it and generally understand the idea

○ Knowledgeable – I have a clear understanding of what it means and why it matters

○ Expert – I have extensive knowledge and/or have worked with such sourcing standards

How knowledgeable are you in defining "ethically sourced code generation models"?

○ Not familiar – I have not heard of this concept before

○ Slightly familiar – I have come across the term, but don't know much about it

○ Somewhat familiar – I have heard of it and generally understand the idea

○ Knowledgeable – I have a clear understanding of what it means and why it matters

○ Expert – I have extensive knowledge and/or have worked with such sourcing standards

Open questions

Your Understanding and Experiences

In this section, we invite you to share your thoughts and experiences related to ethically sourced code generation. Your open-ended responses will provide valuable insights beyond structured options.

Reminder

In this survey, **"ethically sourcing"** refers to practices of ensuring that products are created responsibly and sustainably, and

"ethically sourced code generation" generally refer to

the auto-generated code (i.e., product) produced by a code generation model through ethically and sustainably process from each point of the supply chain.

In your opinion, what should be considered when defining "ethically sourced code generation"?

[ ]

Have you ever encountered ethical sourcing issues while using, working with, or developing code generation models? Please describe. (If you have not, you may simply write "No.")

[ ]

Dimensions

Dimension Evaluation

In this section, we present a series of dimensions that may

be considered when defining ethically sourced code generation in terms of the whole development process of code generation models.

Each question represents a **dimension** within a broader **area**, and includes one or more specific **aspect**s, each with a brief definition. Your task is to evaluate the overall relevance of each category to the definition of ethically sourced code generation.

Please evaluate each of the following dimensions based on how relevant you think it is to the **definition of ethically sourced code generation**. Use the following scale:
· **Not relevant (1)**
· **Slightly relevant (2)**
· **Moderately relevant (3)**
· **Relevant (4)**
· **Very relevant (5)**
Responses range from **Not relevant (1)** to **Very relevant (5)**, indicating increasing importance in the definition.

Reminder
In this survey, **"ethically sourcing"** refers to practices of ensuring that products are created responsibly and sustainably, and
"ethically sourced code generation" generally refer to the auto-generated code (i.e., product) produced by a code generation model through ethically and sustainably

process from each point of the supply chain.

Source-Subject Right: Whether the model development respects the rights of data subjects or creators to control their source and provide informed consent for its use are respected at each stage and its artifacts.
Including:
-**Infomed Consent**: Whether to inform data subjects or creators and obtain consent to how their source is used.
-**Privacy**: Whether to avoid or remove the private or confidential information when using the source.

Not relevant (1)   Slightly relevant (2)   Moderately relevant (3)   Relevant (4)   Very relevant (5)
○   ○   ○   ○   ○

Source-Equity: Whether the model development reflects principles of diversity, fairness, and inclusive representation at each stage and its artifacts.
Including:
-**Diversity**: Whether to consider diverse social, cultural, geographic, and disciplinary perspectives.
-**Fairness**: Whether to eliminate potential biases and do not contain offensive content.

-**Representation**: Whether to consider meaningful and proportional coverage of different demographic groups in source materials.

Not relevant (1)   Slightly relevant (2)   Moderately relevant (3)   Relevant (4)   Very relevant (5)

○   ○   ○   ○   ○

Source-Access: Whether the artifacts in the model development are accessible and properly controlled. Including:

-**Accessibility**: Whether artifacts are available and usable under appropriate conditions—public resources are openly accessible, while sensitive or consent-restricted materials are properly protected or access-controlled.

Not relevant (1)   Slightly relevant (2)   Moderately relevant (3)   Relevant (4)   Very relevant (5)

○   ○   ○   ○   ○

Source-Accountability: Whether the model development ensures responsibility and traceability at each stage and its artifacts.
Including:

-**Transparency**: Whether the stages and artifacts are made visible and traceable to support accountability.

Not relevant (1) Slightly relevant   Moderately   Relevant (4)   Very relevant
                       (2)   relevant (3)   (5)
   ○   ○   ○   ○   ○

Source-Intellectual Property Rights: Whether the model development respects intellectual property rights at each stage and its artifacts.

Including:

-**Source Acknowledgement**: Whether to acknowledge and credit externally sourced or reused artifacts properly.

-**Licensing**: Whether to comply with relevant licenses, even if they are publicly available.

-**Generation Distinctiveness**: Whether the generated outputs are sufficiently distinct from the original inputs to avoid infringement.

Not relevant (1) Slightly relevant   Moderately   Relevant (4)   Very relevant
                       (2)   relevant (3)   (5)
   ○   ○   ○   ○   ○

Source-Integrity: Whether the model development complies with ethical standards, with no contamination involved at each stage and its artifacts.

Including:

-**Contamination**: Whether to remain uncontaminated and rely only on verified sources. For example, model inputs

should avoid mixing with code that contains bot-generated or model-generated content, especially when the ethical standards or provenance of the bot or model are unknown.

Not relevant (1)	Slightly relevant (2)	Moderately relevant (3)	Relevant (4)	Very relevant (5)
○	○	○	○	○

Which **artifact(s)** of the code generation model do you believe the above dimensions should apply to? (Please select all that apply.)

-Training data: The raw data used to train or fine-tune the model.

-Dependencies: Any third-party dependencies, libraries, packages or frameworks used in training data code repositories and model implementation.

-Model Metadata: Artifacts related to the trained model, including learned parameters, hyperparameters, and related metadata.

-Documentation: Written artifacts that provide essential disclosures about the models, such as training source, limitations, and intended uses.

-Prompt: The input or instruction used to guide the model's response.

-Output: The final generated code or response produced by the model.

	Training data	Dependencies	Model Metadata	Documentation	Prompt	Output	All abo
Subject Right (Informed Consent, Privacy)	☐	☐	☐	☐	☐	☐	☐
Equity (Diversity, Fairness, Representation)	☐	☐	☐	☐	☐	☐	☐

	Training data	Dependencies	Model Metadata	Documentation	Prompt	Output	All abo
Access (Accessibility)	☐	☐	☐	☐	☐	☐	☐
Accountability (Transparency)	☐	☐	☐	☐	☐	☐	☐
Intellectual Property Rights (Source Acknowledgement, Licensing, Generation Distinctiveness)	☐	☐	☐	☐	☐	☐	☐
Integrity (Contamination)	☐	☐	☐	☐	☐	☐	☐

Which **stage(s)** of the code generation model development process do you believe the above dimensions should apply to? (Please select all that apply.)

-Data Collection: The process of acquiring data from various sources.

-Data Annotation: The process of labeling data.

-Data Cleaning: The process of removing duplicates, erroneous, or irrelevant data, and preparing data for model training.

-Model Training and Fine-tuning (if applicable): Training large-scale models on dataset and optionally adapting them to downstream tasks.

-Model Evaluation: Assessing model performance using standard code generation benchmarks.

-Model Deployment: Integrating the trained model into systems for real-world use.
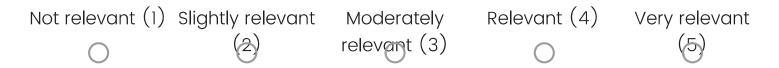
-Post-deployment (Monitoring/Inference): Serving the model in real-world environments and monitoring its outputs and performance during inference.

	Data Collection	Data Annotation	Data Cleaning	Model Training and Fine-tuning	Model Evaluation	Model Deployment	Po deplo
Subject Right (Informed Consent, Privacy)	☐	☐	☐	☐	☐	☐	☐
Equity (Diversity, Fairness, Representation)	☐	☐	☐	☐	☐	☐	☐
Access (Accessibility)	☐	☐	☐	☐	☐	☐	☐
Accountability (Transparency)	☐	☐	☐	☐	☐	☐	☐
Intellectual Property Rights (Source Acknowledgement, Licensing, Generation Distinctiveness)	☐	☐	☐	☐	☐	☐	☐
Integrity (Contamination)	☐	☐	☐	☐	☐	☐	☐

Society-Social Responsibility: Whether the development of code generation models takes on broader social responsibilities, such as contributing to community development at each stage and its artifacts.

Including:

-**Community Development**: Whether communities involved in or affected by code generation benefit from the process, rather than being treated solely as sources of data or labor.

Not relevant (1)   Slightly relevant (2)   Moderately relevant (3)   Relevant (4)   Very relevant (5)

○   ○   ○   ○   ○

Society-Social Acceptability: Whether the development process of code generation models respects for different culture and religious beliefs at each stage and its artifacts. including:

-**Religion**: Whether religious beliefs, practices, and sensitivities are acknowledged and respected.

-**Culture**: Whether diverse cultural identities and values are considered, and cultural dominance, marginalization, or erasure is actively avoided.

Not relevant (1)   Slightly relevant (2)   Moderately relevant (3)   Relevant (4)   Very relevant (5)

○   ○   ○   ○   ○

Society-Labor Rights: Whether the development of code generation model respects basic labor and human rights

of workers involved in each stage and its artifacts.
Including:
-**Fair Wages**: Whether human workers involved are compensated with reasonable and fair pay.
-**Working Conditions**: Whether mentally and physically safe conditions are provided throughout the development process.
-**Legal Employment**: Whether the labor of vulnerable communities is not exploited, and that all employment practices comply with legal and ethical standards.

Not relevant (1) Slightly relevant (2) Moderately relevant (3) Relevant (4) Very relevant (5)

○ ○ ○ ○ ○

Environment-Sustainability: Whether the development of code generation models is conducted in an environmentally sustainable manner at each stage and its artifacts.
Including:
-**Energy Consumption**: Whether energy use and computational resource demands are measured and taken into account.
-**Emissions and Pollution**: Whether the environmental impact—such as carbon emissions and potential pollution—is properly considered during development.

	Not relevant (1)	Slightly relevant (2)	Moderately relevant (3)	Relevant (4)	Very relevant (5)
	○	○	○	○	○

Which **artifact(s)** of the code generation model do you believe the above dimensions should apply to? (Please select all that apply.)

-Training data: The raw data used to train or fine-tune the model.

-Dependencies: Any third-party dependencies, libraries, packages or frameworks used in training data code repositories and model implementation.

-Model Metadata: Artifacts related to the trained model, including learned parameters, hyperparameters, and related metadata.

-Documentation: Written artifacts that provide essential disclosures about the models, such as training source, limitations, and intended uses.

-Prompt: The input or instruction used to guide the model's response.

-Output: The final generated code or response produced by the model.

	Training data	Dependencies	Model Metadata	Documentation	Prompt	Output	All o abov
Social Responsibility (Community Development)	☐	☐	☐	☐	☐	☐	☐
Social Acceptability (Religion, Culture)	☐	☐	☐	☐	☐	☐	☐
Labor Rights (Fair Wages, Working Conditions, Legal Employment)	☐	☐	☐	☐	☐	☐	☐

	Training data	Dependencies	Model Metadata	Documentation	Prompt	Output	All o abov
Sustainability (Energy Consumption, Emissions and Pollution)	☐	☐	☐	☐	☐	☐	☐

Which **stage(s)** of the code generation model development process do you believe the above dimensions should apply to? (Please select all that apply.)

-Data Collection: The process of acquiring data from various sources.

-Data Annotation: The process of labeling data.

-Data Cleaning: The process of removing duplicates, erroneous, or irrelevant data, and preparing data for model training.

-Model Training and Fine-tuning (if applicable): Training large-scale models on dataset and optionally adapting them to downstream tasks.

-Model Evaluation: Assessing model performance using standard code generation benchmarks.

-Model Deployment: Integrating the trained model into systems for real-world use.

-Post-deployment (Monitoring/Inference): Serving the model in real-world environments and monitoring its outputs and performance during inference.

	Data Collection	Data Annotation	Data Cleaning	Model Training and Fine-tuning	Model Evaluation	Model Deployment	Pos deploy
Social Responsibility (Community Development)	☐	☐	☐	☐	☐	☐	☐
Social Acceptability (Religion, Culture)	☐	☐	☐	☐	☐	☐	☐
Labor Rights (Fair Wages, Working Conditions, Legal Employment)	☐	☐	☐	☐	☐	☐	☐
Sustainability (Energy Consumption, Emissions and Pollution)	☐	☐	☐	☐	☐	☐	☐

◄ ──────────────────────────── ►

Since you answered "None of above", what other kinds of **artifacts** do you think should be considered?

[ ]

Since you answered "None of above", what other kinds of

stages do you think should be considered?

```
```

real-experiences

Real-World Experiences and Reflections

In this section, we invite you to share any experiences or thoughts related to ethically sourced code generation in real-world settings.

Short answers are welcome. If nothing comes to mind, you may simply write "No."

After reading our dimensions, do you think you have encountered any situation where an ethically sourced code generation model was needed—for example, when the code generated by models is from unknown sources, you were unsure if it was safe or legal to use? If yes, please briefly describe the situation.

What do you think could be potential impacts or consequences of code generation models that are not ethically sourced (from the perspective of both model developers and users)?

Can you share any examples or experiences that illustrate why you think certain dimensions are relevant to the definition of ethically sourced code generation? If yes, please share your examples, experiences, or just your personal reasoning.

Based on the dimensions you have just considered, do you believe that any currently available code generation tools

align with the definition of ethically sourced code generation? If yes, please list the name(s) of the tool(s) and briefly explain why.

[ ]

Which of the following(s) do you think could be potential impacts of code generation models that are not ethically sourced? (Please select all that apply.)

- [ ] model developer impact: Security risks related to the misuse of personal data or lack of user consent
- [ ] model developer impact: Lawsuits issues due to intellectual property violations
- [ ] model developer impact: Harmful outputs that reflect bias, toxic language, or offensive stereotypes
- [ ] model developer impact: Reduced user trust and willingness to use the model
- [ ] model developer impact: Exclusion of certain users or communities due to limited access or high costs
- [ ] model developer impact: Negative public reputation
- [ ] model developer impact: Environmental impact
- [ ] user impact: Criticized for plagiarism due to using generated code that too similar to existing code
- [ ] user impact: Lawsuits due to using copyrighted code generated by models
- [ ] [ ] Others (Please specify)

Trade-offs

Trade-offs and Acceptable Costs

While making a model ethically sourced could potentially improve its overall quality, it may also involve trade-offs. In this section, we ask about your views on the acceptable costs of making a code generation model ethically sourced. Please answer based on what you would find acceptable in a real-world development or usage context.

How much decrease in output accuracy performance would you consider acceptable if a code generation model were made ethically sourced?

○ No decrease is acceptable

○ Up to 10%

○ Up to 25%

○ Up to 50%

○ More than 50%

○ Not sure / It depends

How much increase in response time (e.g., during generation or inference) would you consider acceptable if a code generation model were made ethically sourced?

○ No additional time is acceptable
○ Up to 10%
○ Up to 25%
○ Up to 50%
○ More than 50%
○ Not sure / It depends

How much additional time cost (e.g., in filtering data) would be acceptable for making a code generation model ethically sourced?

○ No additional time is acceptable
○ Up to 10%
○ Up to 25%
○ Up to 50%
○ More than 50%
○ Not sure / It depends

How much additional financial cost would be acceptable for making a code generation model ethically sourced?

○ No additional cost is acceptable

○ Up to 10%

○ Up to 25%

○ Up to 50%

○ More than 50%

○ Not sure / It depends

What kind(s) of source contamination do you consider acceptable in the training dataset of code generation model? (Please select all that apply.)

☐ Contamination with unknown or unrevealed data sources

☐ Contamination with data from an unclear or unknown dependencies

☐ Contamination from data with incompatible licenses within the dependencies

☐ Contamination with automatically generated content

☐ Contamination from using closed-source models

☐ Contamination with bot submissions

☐ [   ] Others (please specify)

☐ No contamination is acceptable

If trade-offs are unavoidable, which types of costs should be prioritized (i.e., kept as low as possible) when developing ethically sourced code generation tools? Please

drag the following items to rank them in order of importance (1 = most important).

Accuracy loss

Response time / latency

Data preparation time

Financial or resource cost

Data contamination

Block 7

Reflections and Generalization

In this final section, we ask you to briefly reflect on your experience with this survey.

After completing this survey, do you feel that your understanding of ethically sourced code generation has improved?

Strongly disagree	Disagree	Neutral	Agree	Strongly agree
○	○	○	○	○

Has this survey made you feel that the issues of ethically sourced code generation are more important than you previously thought?

Strongly disagree	Disagree	Neutral	Agree	Strongly agree
○	○	○	○	○

Which of the following dimensions do you realize that you have ignored before you taking this survey? (Please select all that apply.)

- ☐ Source-Subject Right (Informed Consent, Privacy)
- ☐ Source-Equity (Diversity, Fairness, Representation)
- ☐ Source-Access (Accessibility)
- ☐ Source-Accountability (Transparency)
- ☐ Source-Intellectual Property (Source Acknowledgement, Licensing, Generation Distinctiveness)
- ☐ Source-Integrity (Contamination)
- ☐ Society-Social Responsibility (Community Development)
- ☐ Society-Social Acceptability (Religion, Culture)
- ☐ Society-Labor Right (Fair Wages, Working Conditions, Legal Employment)
- ☐ Environment-Sustainability (Energy Consumption, Emissions and Pollution)

- ☐ All of above
- ☐ None of above
- ☐ [   ] Others (please specify)

Do you think the dimensions and definitions discussed in this survey could be generalized to other fields beyond code generation (e.g., image generation, voice generation)?

Definitely not	Probably not	Not sure	Probably yes	Definitely yes
○	○	○	○	○

Is there anything about ethically sourced code generation that you still find unclear, confusing, or difficult to define? Are there any other dimensions you think should be included? (If nothing comes to mind, you may simply write "No.")

[   ]