Proposed Harvard AI Code of Conduct

Since its founding, metaLAB has been experimenting with new technologies, while committing to critiquing those very technologies. For the past seven years, we have been exploring the intersections of AI and philosophy, art and design, and teaching. During the Spring 2023 Harvard course Creativity (taught by Professor David Atherton), course guest and metaLAB Principal Sarah Newman led a unit in which students collaboratively brainstormed a prospective code of conduct for the use of AI tools at Harvard. Under the guidance of Sarah Newman, Kathleen Esfahany (metaLAB Research Assistant and Neuroscience PhD Student), and Professor David Atherton, the students' ideas were refined into the following proposed code of conduct, which was shared with the Harvard administration on July 11, 2023. Given the ongoing swift and visible adoption of AI tools in coursework, we hope that this proposal can assist with the development of policies on the use of AI tools in educational settings.

Contents

Authors

Harvard College Students and Teaching Fellows in Creativity (Gen Ed 1067) (Spring 2023)

David Atherton, Assistant Professor of East Asian Languages and Civilizations (FAS)

Sarah Newman, Director of Art & Education and Project Lead, AI Pedagogy Project, metaLAB (at) Harvard, Berkman Klein Center for Internet & Society (HLS)

Kathleen Esfahany, Research Assistant at metaLAB (at) Harvard, Berkman Klein Center for Internet & Society / PhD Student, Harvard Program in Neuroscience (HMS/FAS)

Contributions from Creativity (Gen Ed 1067) Students

The following proposal was inspired by ideas from students from the course Creativity (Gen Ed 1067), taught in Spring 2023 by Professor David Atherton with teaching fellows Peter Dziedzic and Yitian Li. During the "AI & Creativity" unit, led by course guest Sarah Newman, students undertook an exercise to create a proposed "code of conduct" for how Harvard might govern the use of generative AI technologies in educational assignments. At the end of the semester, a subset of the students and the teaching fellows from the course reconvened to refine and consolidate their ideas, which served as direct inspiration for the content of this proposal. We are grateful to the following students for their contributions to this proposal:

Esther An (Computer Science and Statistics, Class of 2025), Antara Bhattacharya (Linguistics and Computer Science, Class of 2025), Andreea Haidau (Applied Math and Neuroscience, Class of 2026), Annie Miall (History and Science and Molecular and Cellular Biology, Class of 2023), Rares Stefan Avram (Computer Science and Statistics, Class of 2025), Dhriti Vadlakonda (Neuroscience, Class of 2026), Juan Valdez (Psychology, Class of 2026), Michelle Wang (Neuroscience, Class of 2025)

Contact

If you have questions about this proposal, or are interested in discussing it further, please email Sarah Newman, Director of Art & Education at metaLAB (at) Harvard.

Summary

Key Points

1. There should not be a blanket ban on generative AI tools at Harvard. Instead, there should be course-specific policies chosen by faculty, with guidance from the administration.

2. Policies on the use of generative AI tools should include a plan for regular reviewing and updating.

3. In all courses, students should receive an official, written policy clearly stating permitted and prohibited uses of generative AI tools and grading policies related to the use of generative AI tools in the particular course.

Suggestions: Implementation, Framing, & Resources

1. Policies for student use of AI tools in educational assignments should be framed as being a component of academic honesty expectations at Harvard College.

2. The Harvard College Honor Council should be involved in the review of generative AI tool use policies and the review of possible policy violations.

3. We recommend that course policies distinguish carefully between *categories* of generative AI tools and *specific* tools.

4. AI detection tools should be used with caution and should not solely underlie decisions regarding policy violations.

5. Educational resources should be provided to students, faculty, and teaching fellows. These materials should explain the potential uses and pitfalls of generative AI tools in educational settings and help guide students who may be unsure of how to comply with course policies.

6. Canvas template policies for use of generative AI tools should be provided to faculty, to assist with their development of their course-specific policies. In this proposal, we provide a set of template policies.

Proposed Harvard AI “Code of Conduct”

I. Overview

The following policy proposal (written April-July 2023) focuses on the use of generative artificial intelligence (AI) tools by students for producing ideas, outlines, or content in educational assignments at Harvard. Generative AI refers to data-driven systems capable of generating media (including language, code, images, music, and more) in response to prompts or queries. Generative AI systems are created through machine learning methods which learn patterns from input training data, followed by various forms of “fine-tuning” and optimization. Following training, these systems are able to generate new data with similar characteristics to the training data.

II. Motivation

Given the rapid pace of development in the capability and accessibility of generative AI tools, there is an urgent need for policies on the use of AI tools in educational settings at Harvard. In the spring 2023 semester, students observed a swift and visible adoption of generative AI tools in coursework. However, students were often using AI tools in their assignments without any formal guidance from their instructors about how these tools should be used. We hope that this proposal can assist with the development of policies to guide students’ use of generative AI tools.

III. Key Points

1. **There should not be a blanket ban on generative AI tools at Harvard. Instead, there should be course-specific policies chosen by faculty, with guidance from the administration.** The growing prominence and accessibility of generative AI tools, field- and application-specific differences in generative AI tool use, and the observed trends in current student use of such tools together suggest that course-specific policies would be much more beneficial to students than a University or College-wide ban on generative AI tools. With guidance and resources, students can gain proficiency in responsibly using generative AI tools, alongside learning about the pitfalls and challenges of such tools. An understanding of the responsible use of generative AI is likely to be of increasing importance for extracurricular and professional opportunities in the coming years, and a blanket ban would limit opportunities for students to develop such understanding.

2. **Policies on the use of generative AI tools should include a plan for regular reviewing and updating.** Given the ongoing rapid advancements in AI, Harvard policies on the use of generative AI should be reviewed, updated, and communicated to Harvard students and faculty on a semesterly basis. A plan for regular review of any policies for student use of these tools should

be established. The cadence of this review should be made explicit to Harvard students and faculty. Relevant policy making groups (such as the Harvard College Honor Council and groups in Harvard Administration) should be provided with clear deadlines such that coordination between groups can occur.

3. **In all courses, students should receive an official, written policy clearly stating permitted and prohibited uses of generative AI tools and grading policies related to the use of generative AI tools in the particular course.** An official, written policy for generative AI tool use should include, at minimum: (1) a recognition of generative AI tools that might be relevant to the course and the policy for their use in coursework; and (2) an explanation of any grading policies related to the use of generative AI tools: students should be made aware of how, if at all, generative AI tool use will affect grading (for example, if the use of a tool would result in a different grading rubric). If the course policy is that permitted uses of generative AI will be assignment-specific, students should still be informed of this as a policy itself and be provided with a few examples of such assignment-specific policies. Course policies on generative AI tool use should be made easily accessible to students through Canvas and course syllabi. We recommend that faculty discuss their generative AI tool policies on the first day of classes alongside other traditional course policies (such as attendance, participation, and collaboration) and take questions from students. Faculty should also be encouraged to discuss their policies on generative AI tools with their students at the start of the semester to gather feedback before finalizing their policies.

IV. Suggestions: Implementation, Framing, & Resources

1. **Policies for student use of AI tools in educational assignments should be framed as being a component of academic honesty expectations at Harvard College.** Harvard students respect and understand the importance of the Harvard College Honor Code, which delineates the commitment of the Harvard College community to academic integrity and honesty. By framing policies for AI tool use as being a component of academic honesty expectations, students may more quickly understand the importance of policies from Harvard Administration and their instructors, and may be more committed to following such policies. Additionally, referring to the Honor Code (https://honor.fas.harvard.edu/honor-code) in any announcements may help students draw the connection between already-familiar academic honesty expectations (such as "traditional" plagiarism) and unfamiliar expectations on AI tool use.

2. **The Harvard College Honor Council should be involved in the review of generative AI tool use policies and the review of possible policy violations.** The Harvard College Honor Council, created in 2013, reviews possible violations of the Harvard College Honor Code and academic integrity policy. The Honor Council (or another similar group composed of students and faculty) should be involved in the regular review of Harvard's policies for student use of generative AI tools. In cases where course-specific policies on generative AI use may have been violated – constituting cheating or misrepresentation, both prohibited by the Harvard College Honor Code –

the Honor Council should be involved in the review of the possible violation as they would be in cases unrelated to AI.

3. **We recommend that course policies distinguish carefully between *categories* of generative AI tools and *specific* tools.** Course policies may refer broadly to “generative AI tools” and/or include further specification of the *category* of relevant tools (such as “large language models” or “image generation models”). They may also refer to *specific* tools (i.e. “ChatGPT”, “GPT-3”, or “DALL-E”). For example, an instructor may wish to limit students to just one tool (i.e. “Students may use ChatGPT to brainstorm ideas, so long as they cite the use.”) or may wish to encourage students to use any tool (i.e. “Students may use AI tools to brainstorm ideas, so long as they cite the use.”)

We recommend instructors distinguish carefully between these terms in their course policies. This distinction is especially important for policies which intend to *limit* generative AI tool usage. To effectively limit generative AI tool usage, policies should refer to “generative AI tools” or categories of tools (such as LLMs and LLM-powered tools). Such policies should avoid language referring *solely* to specific models (i.e. “ChatGPT”) to avoid students circumventing the intended policy by using similar models not mentioned in the policy. For example, a course policy that requires students not to use “large language models” (LLMs) or “LLM-powered tools” for essay writing effectively conveys that students should not use *any* tools that generate text. This type of policy is recommended over a policy that advises students to “not use ChatGPT” for essay writing, since students may then instead turn to a different large language model tool (such as Bard) for the same purpose.

Faculty may still find it beneficial to name specific tools as an *example* within a category (i.e. “large language models, such as ChatGPT, should not be used for essays”) while still making clear that the entire *category* of tool is affected by their policy.

4. **AI detection tools should be used with caution and should not solely underlie decisions regarding policy violations.** In the face of AI-generated media (especially AI-generated text), a plethora of tools have been released which claim to distinguish between human-written and AI-generated “synthetic” text. However, these tools are highly inaccurate, which can both lead to missed instances of AI-generated text and unfairly implicate students who did not use AI in their writing. Furthermore, trying to “detect” academic dishonesty in this way conveys a lack of trust of students in an academic environment that is already based on an Honor Code.

The challenges of reliably detecting AI-generated content are important to communicate to instructors, given the high stakes involved when attempts to detect such instances go awry. For example, TechCrunch tested seven “AI-text detectors” (using eight types of documents, such as an essay, a cover letter, a resume, and more) and found that none of the detectors were reliably accurate. Existing plagiarism tools, such as Turnitin, have also begun including AI detection tools, which have also caused issues due to their inaccuracy, as reported in the Washington Post. Unfortunately, some educators have inappropriately used generative AI tools to try to detect if student work is AI-generated. For example, an instructor at Texas A&M University incorrectly used a large language model (ChatGPT) to attempt to detect if students had used AI to generate their assignments, immediately giving students a score of zero when the large language model claimed students had used AI (see more in the Washington Post).

Currently, there are no reliable tools for identifying AI-generated text. Until there are reliable AI detection tools, such tools are not recommended. Importantly, decisions regarding whether a policy violation occurred should not be based on the findings of any such tools.

5. **Educational resources should be provided to students, faculty, and teaching fellows. These materials should explain the potential uses and pitfalls of generative AI tools in educational settings and help guide students who may be unsure of how to comply with course policies.** Educational modules about generative AI tools and their use in academic settings will help support students, faculty, and teaching fellows. Some of these modules may be analogous to those which aim to educate students about plagiarism and proper citation, in which both positive and negative examples are provided and explained. An important message to convey in these materials is that current large language models (such as ChatGPT) are not credible sources for facts or citations; as such, these models should not be used as a replacement for proper research and citation. Students should be made aware that improper citation as a result of using a generative AI tool may carry the same consequences as ordinary improper citation (i.e. work generated by someone or something other than the student without attribution to that external source is plagiarism).

6. **Canvas template policies for use of generative AI tools should be provided to faculty, to assist with their development of their course-specific policies.** We have drafted template policies which instructors may find helpful to use – either directly, or as a starting point for further refinement – on their Canvas course pages and in their syllabi. These templates are available below.

Canvas Template Policies

Course syllabi should include policies on the use of AI tools alongside traditional course policies concerning attendance, participation, collaboration, and grading. Students should be provided with clear policies on the permitted and prohibited uses of AI tools for their courses. In courses where policies on AI tool use may vary from assignment to assignment, this policy should be made clear, and assignment-specific policies should be provided.

The following "Canvas Templates" may be a helpful starting place for Harvard faculty, either for direct incorporation or for modification/customization. These templates could be made available directly on Canvas.

Below, we provide template policies organized into two sections. In the first section, we provide a template for how students should acknowledge and cite their use of a generative AI tool, written so as to align with the Harvard College Honor Code. Any such acknowledgement/citation policies should be used in combination with course- or assignment-specific policies on permitted uses of AI. In the second section, we provide templates for course- or assignment-specific policies.

I. Acknowledgement and citation of generative AI tool use

Students are expected to provide proper recognition of the contributions of external sources in their work. As such, any contributions of generative AI tools to student work should be clearly disclosed by the student.

- **Baseline policy:** The Harvard College Honor Code requires adherence to "accurate attribution of sources" and the "transparent acknowledgement of the contribution of others" to "ideas, discoveries, interpretations, and conclusions." As such, all material included in assignment submissions must be properly attributed to sources (using quotations and citations as appropriate). Failure to attribute material to its original source constitutes plagiarism. Students should be aware that generative AI tools often generate incorrect statements, generate fake sources, and/or do not attribute material to proper sources. **Students must acknowledge all instances in which generative AI tools were used in an assignment** (such as in ideation, research, analysis, editing, debugging, etc.). *All* submitted work by a student must either be original work or properly attributed to external sources, as stated by the Harvard College Honor Code. Students are responsible for the *entirety* of their final submission; any inaccuracies or other deficiencies cannot be excused on the basis of originating from an AI tool.

- **Suggested implementation of baseline policy:** For any assignment in which generative AI tools were used (always and only in accordance with the policy specified for the course or the particular assignment), students must include, in addition to a traditional bibliography, a written page entitled "explanation of AI tool use" that contains a description of ***which*** tools were used (such as ChatGPT, DALL-E, etc.), ***how*** each tool was used, (such as in ideation, research, analysis, editing, debugging, etc.), the ***specific prompts*** entered into the model(s), ***how*** model outputs were evaluated, and ***whether and where*** model outputs were used in the work that was submitted.

II. Course- or assignment-specific policies

Listed from 1 (full use) to 4 (no use). Note that in these templates, "coursework" can be replaced with "the assignment" to make the template assignment-specific, rather than course-specific.

(1) **Generative AI tools may be used for any elements of the coursework**, **so long as that use is acknowledged and cited by the student**. The acknowledgement/citation should be provided in accordance with the baseline implementation policy above (an addendum to all submitted work which acknowledges the generative AI tools used and the particulars of the use).

(2) **Generative AI tools may be used for preliminary or exploratory elements of the coursework, including inspiration, ideation, brainstorming, "feedback," summarizing, outlining, etc., but may not be used in the production of final deliverables, such as essays or reports.** All submitted work must be solely produced by the student, and any use of AI tools for preliminary elements must still be disclosed and acknowledged in accordance with the baseline policy above.

Recommendation: For any given course (or assignment), faculty should outline the specific parameters for AI tool use, and students should be encouraged to ask for clarification. In this template policy, for example, faculty should make clear what constitutes preliminary or final work. This is particularly important in the context of different applications of generative AI tools: for example, while language models may not be permitted to generate any text that is submitted by the student, perhaps an image generation AI tool may be used (with citation) to create an illustration for a report.

(3) **Generative AI tools may be used only when (and how) specified for coursework, and otherwise are not permitted.** This may include for a particular assignment or exercise, for a particular module of a course, or for a particular type of AI tool. For any use of a generative AI tool, we recommend that the baseline policy (an addendum to all submitted work which acknowledges the generative AI tools used and the particulars of the use) still be required, unless the course faculty specifies otherwise (for example, the addendum may not be necessary for an in-class assignment on AI tools).

(4) **Generative AI tools may not be used for coursework, in any form, and use of such tools (in ideation, summarization, research, feedback, writing, and/or any other forms of use) is strictly prohibited.** The use of a generative AI tool in this course (or assignment) constitutes cheating. If students are unsure about whether a particular tool or app uses generative AI, students should ask *before* using the tool to complete their work.