Report of the AI at Stanford Advisory Committee

January 9, 2025

Russ Altman, Kenneth Fong Professor and Professor of Bioengineering, of Genetics, of Medicine, of Biomedical Data Science, and Senior Fellow at the Stanford Institute for HAI *(Committee Chair)*

Yi-An Chen, Senior University Counsel

Michele Elam, William Robertson Coe Professor of Humanities and Senior Fellow at the Stanford Institute for HAI

Zephyr Frank, Gildred Professor of Latin American Studies and Professor at the Stanford Doerr School of Sustainability

Steve Gallagher, Chief Information Officer, University IT

Stephanie Kalfayan, Vice Provost for Academic Affairs

Serena Rao, Senior Associate Dean for Finance and Administration, VPDOR

Mehran Sahami, Tencent Chair of the of the Computer Science Department and James and Ellenor Chesebrough Professor

Dan Schwartz, Dean of the Graduate School of Education

Zack Al-Witri, Assistant Vice Provost for Academic Affairs *(Staff to the Committee)*

Introduction

On March 18, 2024, the Provost charged the AI at Stanford Advisory Committee to assess the role of AI at Stanford in administration, education, and research to identify potential policy gaps and other needs to advance the responsible use of AI at Stanford. The committee met seven times between March 18 and June 19 to assess potential policy gaps in the areas of administration, education, and research at Stanford. They were informed by reports, policies, and resources from peer institutions that have assessed similar questions over the last year (e.g. Cornell, Michigan, Harvard, Yale, and Princeton).

Stanford and the Rapidly Changing Landscape of AI Use

There are great opportunities and great challenges associated with the newest generation of AI technologies, particularly large language models (LLMs). They represent powerful transformative tools for productivity because of their ability to generate/draft, summarize, and analyze text, images and other media. At the same time, they are new and have not systematically been assessed (for example, with respect to biases within the data used to train them, and their performance on critical tasks that are fact-based). Their attractive output may lull users into a misplaced sense of quality, precision, and accuracy. Most importantly, these technologies are rapidly evolving in both their capabilities and the degree to which they can be inspected, tested, verified and validated. And, while such models continue to improve, their already-known shortcomings require that care be taken when using these systems in many contexts.

Stanford is a leader in the development and application of AI. The university community includes experts in the creation, validation and extension of the core capabilities of AI. Across all schools and units, many colleagues are already using AI to advance teaching and research in their disciplines. Therefore, it seems prudent to articulate a core set of policies, standards, and best practices for the use of AI in the administrative, educational and research functions of the university. A fine balance is required. The committee wishes to encourage experimentation with the technology as it evolves, to maximally benefit the university, the research enterprise, and teaching and learning, while ensuring that key principles of the university are honored. Thus, our group approached its charge through the lens of providing guardrails to support productive uses of AI and not to stifle innovation, creativity or exploration.

Per its charge, the committee sought to identify potential policy gaps. We list these, along with other considerations and recommendations for university activities in administration, education and research. Given the pace of change in the AI field, the committee cautions against creating fixed, rigid policies (except where required by law or other compelling considerations). Rather, we recommend the adoption of approaches that are flexible and do not unduly limit the creative use of AI to support the university's mission. In some cases, policies are not required—guidelines and best practices will suffice and provide more flexibility. At a minimum, any best practices, standards or policies promulgated by the university should be evaluated and updated regularly based on their relevance and effectiveness.

The committee also arrived at a set of general principles to help guide the approach to AI, since many situations and challenges may not be anticipated in advance. These principles should help guide the *ad hoc* management of new issues as they arise. Our group learned that there are hundreds of uses of AI each day conducted by faculty, staff, students/trainees. General principles to help guide these "experiments" may be more useful to individual users and practitioners than any policy to cover the specific situations and use cases that we were able to anticipate and describe below. Of course, consultation with the Office of the General Counsel (OGC) is important when potential legal issues arise.

Guiding Principles for AI at Stanford[1]

We encourage the university to view goals in this space in line with the philosophy of Stanford's Institute for Human-Centered AI (HAI) to make sure that the use of AI is centered on improving the human condition. Thus, we prefer to focus on "augmenting human capabilities" versus "replacing humans." Many of us at Stanford aspire towards a

[1] Examples of other related guiding principles and resources:

- Stanford Research Policy Handbook research principles 1.1
- Whitehouse Executive Order on Safe, Secure and Trustworthy AI
- NIST Risk Management Framework for AI 1.0
- National Academies of Science, Engineering and Medicine publications on AI

goal where AI positively impacts our community. We must take into consideration the fair distribution of benefits that come from AI. We seek to remove or repair AI systems that tend to harm individuals or groups disproportionately. For critical systems affecting human health and welfare, we aspire for a high bar for validation and verification of AI accuracy, applicability, fairness, and equity before deployment. Moreover, we should resist the tendency to assume that existing laws, regulations, and university policies are not applicable in the context of AI use, a tendency we call "AI exceptionalism." In short, we believe that the use of AI at Stanford should be human-centered.

We modify from a Cornell report a review of the elements of human-centeredness to formulate the following guiding principles that we can strive towards in our deployment of AI at Stanford:

Human Oversight. Humans should take responsibility for the AI systems used or created at Stanford. Each system should aim to have a clear line of authority and responsibility for system procurement/creation, maintenance, monitoring and sunsetting. There should be plans for re-engaging human control of systems when they do not work, and there should be contingency plans for management of high-risk AI systems.

Human Alignment. AI systems should be built/procured to support Stanford's mission and core values. When appropriate, systems should be adopted and deployed in consultation with diverse stakeholders to identify risks and mitigation strategies–particularly for disparate negative impacts. The scope of use, system capabilities, and limitations of systems should be documented in an ongoing process and evaluated regularly for consistency with community standards.

Human Professionalism. All members of the Stanford community should adhere to standards and aim for high levels of rigor and quality in their work. We expect that they will exercise their best judgment and critical thinking in the use of AI tools. All community members are responsible for the content of our work and must be willing to take responsibility for that work, regardless of whether it was assisted by AI. They are also responsible for the quality of reasoning and veracity of assumptions that underlie the work. Community members are also responsible for abiding by applicable laws, university policies, and must consider the expectations and norms within the multiple professional communities to which they belong.[2]

Ethical and Safe Use. AI should be used to improve university functions. Decision makers should aim to understand the full implication of using AI systems in university functions and should embark upon evaluations when these implications are not fully understood. All community members should aspire to ensure that AI services promote justice and individual autonomy and are free from bias and unlawful discrimination.[3] The

[2] These policies and expectations, as we recommend below, ought to be clearly communicated to the Stanford community and accessible through an online portal.

[3] HHS recently issued rules for healthcare providers to follow to make sure decision-making systems (such as AI tools) are non-discriminatory. eCFR :: 45 CFR 92.210 -- Nondiscrimination in the use of patient care decision support tool. While the rules are intended to apply to patient care settings, it may

community should have processes for identifying and decommissioning previously deployed AI systems that are not safe and/or ethical.

Privacy, Security, and Confidentiality. When AI systems use personal data (especially sensitive data, as defined by law or common practice), the legality and impact should be appropriately assessed. If personal data is used for decision-making by AI systems, Stanford should reasonably aim to be transparent about how the decisions are made, whether they are inspectable and the degree to which they have been validated and verified independently. Some uses of data or application of AI systems may require consent, such as attorney-client privileged information, medical records and psychotherapist-patient information. Generative AI tools (such as Open AI and Chat GPT) save, reuse, and share the information entered with their affiliates including confidential or sensitive information. Any confidential or legally privileged information of Stanford or a third-party may not be provided to generative AI tools.

Data Quality and Control. All data used to create new AI systems with university resources should be collected in legal and ethical ways, and data provenance should be explicitly documented and managed as part of the system design. The university should consider building and maintaining a world class data-centric infrastructure that prospectively plans for data use in AI across the research lifecycle. AI systems should be tested and their risks documented. Technical and organizational controls should mitigate the risk of interference or exploitation by bad actors.

An AI Golden Rule. During a time of great change in the capabilities and uses of AI, it may be useful to "use or share AI outputs as you would have others use or share AI output with you." One test would be for an individual to consider if they would be comfortable if the roles of AI user and recipient were reversed. Another test might be for an AI user to consider whether they would be willing to transparently disclose the details of their use with those affected by the AI output. These assessments would likely change over time, and would be based on individual judgements as well as evolving community norms for use of AI. They would be combined with the other principles to inform decision making about the use of AI.

Policy Areas

While surveying the current uses and activities related to AI at Stanford, the committee found some critical areas that may require further guidance and support from the university. Appropriate groups (indicated where possible) may wish to evaluate the need for, and content of, additional policies related to the use of AI in conducting Stanford activities.

AI in university administration processes

provide a helpful framework to consider for evaluation of responsible use of AI in decision making generally.

- **Hiring.** There are concerns of bias and risk of litigation around the use of AI to review, screen or filter applicants for jobs or roles at the university. There is also concern about bias in the use of AI to prepare job descriptions and job ads. *(University Human Resources.)*

- **Performance Reviews.** The use of AI in the generation of performance review materials may raise issues of trust and staff morale, and should be managed very carefully. In addition, there may be legal implications to AI-generated materials in performance reviews. *(University Human Resources.)*

- **Admissions.** Like hiring, the use and influence of AI in admissions decisions and application review at the undergraduate and graduate level raise concerns around bias, risk of litigation, and reputational harm. The committee recommends not using generative AI in the admissions process without careful and documented assessment of the performance of these systems. We recognize that AI may be helpful in creating communication materials for recruitment, but it is important to use it carefully in ways consistent with the guiding principles articulated above. *(Faculty Senate Committees on Undergraduate Admissions and Financial Aid and Graduate Studies.)*

- **Communications.** Staff in communications roles across the university may need guidance in the appropriate use of AI for generating university content. *(University Communications.)*

- **Surveillance.** Any data collection about the activities of members of the Stanford community is a potential threat to privacy, confidentiality, personal autonomy and the trust between the university and the members. At the same time, some of these technologies may be useful for tracking student classroom attendance. Therefore, there should be guidance on the procedure for deciding on surveillance, the expectations for transparency (including potentially consent), the expectations for data management, maintenance and deletion. *(University Privacy Office and the Faculty Senate Committee on Academic Computing and Information Systems.)*

Other recommendations and considerations:

- **Education and training on sensitive data.** Existing data security and privacy policies for computing resources cover many of the uses of AI services. However, there may be a sense that AI tools are exceptional or are not covered by these rules. For example, it is easy to paste text into ChatGPT and then forget that it might be inappropriate to provide sensitive university data to an external entity (like OpenAI) except under clearly prescribed conditions. The university should consider methods to inform university community members with access to these data of these risks.

- **University-provided access to LLMs**. There are equity issues in access to LLMs, because they can be expensive and thus access across faculty, staff, students/trainees may be uneven. At the same time, most of the major vendors of LLMs are not transparent about the data used to train their models, and the models may not have been evaluated systematically for bias, fairness, precision or accuracy. The committee supports investigation into ways to make access to these tools more balanced, while offering choice and asking vendors to help characterize their products beyond "see what you think." If LLMs are provided by the university, the committee sees an opportunity for education about how to use them and how to be wary of their outputs.

- **Vision for AI and staffing.** It is our view that the best use cases of AI are where they help and enable staff in their work. The university should consider describing its vision for AI systems to augment and not replace staff work, and to emphasize ethical issues and ways that these should be mitigated. Staff may be especially concerned about job loss in the context of AI, and so the university leadership may want to consider how a vision of "augment do not replace" might be adopted. Such a policy should be complemented by readily available education in how staff (and others) can use AI to augment their work. *(University Human Resources.)*

- **Streamlined procurement process.** It is clear that Stanford community members are using Stanford resources to purchase AI systems for a variety of purposes. This is consistent with the vision of experimentation and testing of AI. However, it also may create risks for the university if the procured software is not consistent with the guiding principles we outlined above (e.g. professional activities of staff are captured and saved, student information is captured, surveillance data is used for non-Stanford purposes). The committee recommends that high risk software be identified and evaluated, so that the community can learn about AI without subjecting community members to risks that violate our guiding principles.

- **A website to communicate guidelines and resources to the community.** The university should continue to expand web resources that provide the Stanford community with easy access to information about AI resources, policies, and other supporting information. UIT's GenAI website currently provides a central site with easy navigation to the HAI and Responsible AI websites while presenting an overview of gen AI to the broader Stanford audience. The site's GenAI Evaluation Matrix helps Stanford community members understand the risk classifications for many commonly used GenAI tools. Collectively, pages like these may provide resources to improve AI literacy and present approaches for engaging with gen AI platforms while raising awareness about security and privacy concerns in order to use AI most productively.

- **Letters of recommendation.** Letters of evaluation and recommendation are key to academic advancement, hiring and promotions. These letters ideally contain

honest and original descriptions of candidate activities, and so the use of AI to generate text as a substantial component of letters of recommendation could have unintended impacts—such as recipients recognizing that the letter was produced with AI and thereby diminishing the value and weight of the letter. At the same time, an AI-generated letter as a first draft followed by substantial editing and preservation of original descriptions and detailed (often coded) evaluation language could lead to the more rapid creation of high-quality, accurate, personalized and useful letters. The committee recognizes that faculty have a large number of letters to write annually for a variety of purposes, and so generative AI may have some utility, if used appropriately. The AI Golden rule may be a useful principle in these activities.

Applications of AI to support education

- **Assessment and grading**. The committee sees several potential risks in use of AI software to grade, assess, detect plagiarism, or provide feedback on student work, including (but not limited to) quality, accuracy, fairness, and potential bias. In addition, there are potential privacy concerns when using third-party tools and any FERPA implications. This is another rapidly evolving area, and it seems prudent to recommend or require instructor disclosures and justifications when AI is used for student assessment or student feedback. *(Faculty Senate Committees on Undergraduate Standards and Policy and Graduate Studies.)*

- **Student use of AI.** Students have generally been early adopters, particularly of generative AI for writing–as well as other uses. They are preparing for careers where AI will likely be ubiquitous and increasingly powerful. The committee recognizes that there is an existing Honor Code policy about the use of AI. This should be examined for revision or updating; some faculty have found the guidance unclear, while students and others have also voiced that the policy is unenforceable and therefore not effective or useful. AI should be considered in the context of evolving changes to the Stanford Honor Code. Given predominant concerns about cheating, it may also be valuable to provide finer distinctions in the ways students (and faculty) may use AI to support their work, for example as a source of ideas or feedback versus a wholescale use of AI to take or grade exams or homework. Policies may be quite different for different uses of AI for learning. *(Board on Conduct Affairs.)*

Other recommendations and considerations:

- **Accommodations.** The committee recognizes that AI tools have the potential to be tools to address issues of accessible education. This is a complex topic that goes to the heart of how accessible education should be approached and requires careful deliberation on ensuring optimal educational and learning goals, while considering equity and applicable laws.

- **Sandbox for users.** The committee noted multiple concerns about the lack of university-sanctioned AI services, guided training programs, and best practices

may have deterred Stanford workers from experimenting with AI for their job (administration, education, research). This is particularly acute in administration where efficiencies may be found, and education where instructors and students are interested in learning how to best use the tools. The university should consider offering a makerspace, sandbox, or other setting where community members can experiment with AI tools. For example, the Graduate School of Education has created an "AI Tinkery" (a digital maker space), where educators can explore different AI tools with the help of dedicated staff, who also provide regular workshops. *(Center for Teaching and Learning and Learning and Technology Services could provide a makerspace open to all instructors, utilizing Stanford's AI Playground.)*

- **Access to some LLM tool for all students/campus.** The issue of universal access to one or more LLM tools for student equity came up repeatedly in our discussions and is related to the Sandbox consideration above. Student access to AI and LLMs is a fundamental but difficult issue. At a minimum, access to LLMs may be considered as an educational cost (like textbooks) in financial aid and "cost of college" calculations. There should be a process for determining the process for evaluating, procuring and distributing site licenses for current and future LLMs. Issues of evaluation–accuracy, precision, validation, verification will likely become increasingly prominent in such decisions. *(University IT.)*

- **Resources for teaching with AI.** There is considerable variability across fields and between individual faculty in how best to approach teaching with AI. The university should consider providing frameworks and worked-out examples to help instructors think through all aspects of pedagogy impacted by AI (assessment, 1:1 tutoring, the definition of cheating and its detection). For example, a selection of sample course policies in addition to the existing training resources and workshops on offer through CTL and the Stanford Accelerator for Learning in the GSE is sponsoring seed grants and community events on the development of new AI-infused pedagogies. *(Leverage expertise in the Graduate School of Education together with the Center for Teaching and Learning.)*

- **License for plagiarism checkers.** Access to plagiarism checking software for researchers is increasingly required for author self-checking against inadvertent and unintentional plagiarism. *(Stanford Libraries.)*

Uses of AI in research

- **Authorship.** The university's current policy on authorship, written at a time where multi-disciplinary research was growing, may be due for an update to address potential issues around attribution between individual authors and AI. Although journals, publishers, and the government will no doubt create policies around this, Stanford should aim to have basic statements of authorship consistent with academic values, processes, and assessment of merit. In particular, the risk of quoting text beyond currently acceptable academic standards is present in some

of the current LLMs and requires author professionalism. *(The Vice Provost and Dean of Research and the Faculty Senate Committee on Research.)*

- **Misconduct.** AI tools, in particular AI plagiarism detectors, are already leading to a higher volume of allegations (including spurious allegations), and this creates a huge burden on university resources in adjudicating these allegations. The university's current misconduct policy and rules for investigating individual allegations will need to be re-evaluated and updated in accordance with the new federal policy.[4] *(The Vice Provost and Dean of Research.)*

- **Review and writing of proposals.** There is a growing prevalence of AI content in research proposals and reviews of research proposals. Grant agencies are developing rules around such uses of AI; the university may similarly consider its own policy or supplement existing guidance. Whereas the use of AI in brainstorming, copy-editing or otherwise refining grant proposals may be very useful, the use of AI as a proxy reviewer may violate fundamental academic standards.

- **Training AI on student work.** There are active and exciting efforts to train AI systems to provide tutoring and evaluation of student work. These systems often require large amounts of student work for training, and the use of student work-product without permission and appropriate approval processes may be inappropriate or even illegal in some cases. *(SDOC, the Student Data Oversight Committee, should provide guidelines on appropriate use of student data for AI research.)*

- **Oversight on using data for AI research.** AI research often involves the use of large amounts of training data which may have legal and/or ethical concerns. Given that much of health AI research is exempt from IRB review under federal regulations, the university should clarify the process for reviewing and providing support for responsible and compliant use of data. *(The Vice Provost and Dean of Research.)*

Other recommendations and considerations:

- **Legal issues.** As mentioned above, there may be a tendency to think of AI as an exception to existing laws, rules, and university policies; in the research sphere. Examples of risk areas include copyright and trademark infringement disputes (e.g. when curating and distributing datasets for AI that contains copyrighted work), breach of contractual agreements (e.g. violating our agreement with data owners), potential tort and defamation liabilities (e.g. if the AI tools produce fake, or harmful information or advice), and violation of privacy rights of others and related regulations (e.g. using or sharing PHI or other personal data in violation of existing policies and protocols, or failing to obtain adequate consents to use

[4] The Office of Research Integrity under the US Department of Health and Human Services has recently published the rule on research misconduct.

personal data). Litigation risk will continue to shift to the users of AI (as opposed to the initial focus on developers) as the use of this technology becomes more commonplace. Researchers should be reminded or made aware of risks and obligations in contexts where AI is used for research and updated as the legal risk profile changes.

- **Computing support for campus AI.** The university should consider ways to expand computing resources to enable AI-powered research and experimentation to ensure that Stanford administrators, educators and researchers remain leaders in the productive and human-centered use of these technologies.