Teaching and Examining in the age of generative AI

A First Assessment of the Consequences for Higher Education

Thomas Bieger & Martin Kolmar

January 25, 2023

Abstract: *As generative artificial intelligence (GAI) continues to advance, increasingly powerful tools are developed for professionals in various industries, enabling them to economize on time, perform complex analyses and even be "creative." As a result, institutions of higher education (IHE) must prepare students to effectively utilize these tools. However, achieving this goal presents a challenge, as IHEs must navigate the task of defining appropriate teaching methods while also ensuring that students do not misuse GAI tools in their coursework. The integration of GAI tools into higher education poses a significant challenge as traditional forms of teaching and examination may become dysfunctional. Examples of colleagues testifying that GAI-generated essays would get at least a "pass" are abundant by now. To effectively teach the competences necessary to create value with and beyond GAI and prepare students for careers in a the new world of work, IHEs must reassess their models of teaching and examination and, therefore, faculty development. This should be done with a clear strategy regarding the role of IHEs in education.*

1. Introduction

Many educators and managers of IHEs might have undergone a similar experience during November/December 2022. Maybe alerted by their teenage kids, maybe by the massive media coverage, they became aware of advances in GAI that have the potential to disrupt teaching and examination. A chatbot by OpenAI called ChatGPT received most public attention because of its easy-to-use interface and its apparently amazing capabilities to generate text. During that time, one of the authors and his team was preparing a course, the so-called "integration project" for our 1,800 first-year students, in which essays play an import role. We fed the guiding questions into ChatGPT to see what happens, were amazed by the quality of the answers, and started anew, looking for "chatbot-proof" formulations. Conversations with colleagues reveal that we are not alone: on very short notice, the widespread availability of GAI-tools forced us to rethink our approach to teaching, examination, and strategy of our institutions more generally. In this essay we share our preliminary thoughts on these issues, as we are convinced that GAI has the potential to be sufficiently big to challenge our paradigms of higher (business) education.

A typical and intuitive first reaction of educators to new technologies challenging traditional ways of teaching is often to ban them or at least to constrain their use. Like when calculators came up, but students were still forced to solve mathematical problems by hand. This reaction, however, is deeply flawed. If the goal of education is to prepare students for real (work) life, and if GAI promises to be an important part of it, we must find ways to integrate it into our programs even if doing so

challenges our traditions. It almost comes without saying that it is our obligation to help students develop the necessary skills to productively use these tools. But also inside of the classroom, GAI tools promise the potential for creating more new opportunities. But as with all technologies, they come with limitations, and they provide their own set of ethical challenges.

At present, not even the developers of GAI tools fully understand these potentials and risks. Over time, it will become clear what they can do for us and what this potential implies for our programs. The general attitude should be to embrace instead of to ban them. In the same way that innovations like writing, printing, or calculators extended and changed our productivities, GAI will be used as a "cognitive extension" that will—if wisely used—allow to be more autonomous by freeing up time from repetitive, time-intensive, and sometimes boring activities. Take books as an example: instead of laboriously memorizing large texts, the printing press freed our cognitive abilities to focus on other aspects of creative and productive activities; books are a kind of external memory system. From today's perspective, it seems absurd to restrict the use of books by our students, but the transition from "internal" to "external" memory systems required the adaptation of the way one teach and examine.

2. Main challenges

IHEs need to understand what skills, competencies, and attitudes are required—and therefore place them at the center of their curricula, teaching, and examination methods—in order to offer their students a credible promise of a successful career. Before rushing to conclusions, however, we must make sure that we are asking the right questions regarding GAI. It seems useful to analyze three fundamental questions first:

- First, we need an idea about the impact of GAI on the future of work. It will change the way certain professions organize their value chains and while at the same time disrupting others and in this process creates new professions. Hence, program managers must reassess the skills and competencies that are necessary for their students to flourish in their future careers. If there is sufficient risk that the present profile of a program is in an area where GAI tools are getting good at, this profile must be adapted. Humans cannot compete with machines in areas machines are designed to be good at. The value proposition of academic programs depends on the credible promise that graduates will not compete with machines in the foreseeable future. Top schools must equip their graduates with the ability to create value beyond what machines can do.
- Second, despite of the fact that the body of research is growing, we do not yet fully understand how students learn and develop their skills and personalities effectively in environments that blend digital tools with traditional teacher-to-learner formats. However, we need a robust understanding of the optimal mix

of human and technological support to develop programs effectively. As long as we do not have empirically sound idea about how to best integrate these technologies in order to facilitate learning and flourishing, reform ideas are educated guesses at best.

- Third, "rethinking" or "redesigning" management education has become a central theme of business schools over the past decade,[1] and we should link debates about the changes in teaching and assessment necessitated by the GAI with this overarching discussion. One of the main challenges is to reimagine management education to make sure that students as future leaders learn skills that remain relevant over longer periods of time and in new, changing, and unknown contexts and that prepares them to make responsible decisions. Conceptual models envisioning these new objectives often build on a three-pillar model: Business schools need to enable students to develop scientific excellence, integrative thinking capabilities needed to solve practical (and complex) problems, and to foster a process of "becoming" responsible leaders. Hence, curricula need to facilitate *knowing* (theories, models, frameworks), *doing* (literacy, competencies, techniques) as well as *being* (values, beliefs, self-reflexivity).[2] It is important to keep this overarching agenda in mind when thinking about the opportunities and risks that come with GAI.

The appropriate way to adapt to the challenges imposed by GAI on teaching and evaluation methods depends on how we answer the above questions. However, we will argue that some broad guidelines can already be derived. To do so, we must be more specific regarding the technology we are talking about. We focus on technologies that are powering chatbots like ChatGPT that are basically *large-language models* (LLM) augmented by a convenient user interface.[3] LLM are trained on huge quantities of text data to infer the most likely contexts in which phrases are used; what LLM's do is *sequence prediction*. Shanahan (2022) illustrates the implications nicely: "Suppose we give an LLM the prompt 'The first person to walk on the Moon was', and suppose it responds with 'Neil Armstrong'. What are we really asking here? In an important sense, we are not really asking who was the first person to walk on the Moon. What we are really asking the model is the following question: Given the statistical distribution of words in the vast public corpus of

---

[1] Steyaert, C., Beyes, T, & Parker, M. (2016). The Routledge Companion to Re-Inventing Management Education. London: Routledge. Colby, A., Ehrlich, T., & Sullivan, W. M. (2011). *Rethinking Undergraduate Business Education: Liberal Learning for the Profession*. Jossey Bass. George G, Howard-Grenville J, Joshi A & Tihanyi L. 2016. Understanding and tackling societal grand challenges through management research. Academy of Management Journal, 59(6): 1880-1895.

[2] Muff, K., Dyllick, T., Drewell, M., North, J., Shrivastava, P., & Haertle, J. (2013). *Management Education for the World: A Vision for Business Schools Serving People and the Planet*. Edward Elgar Publishing. Muff, K. (2013). Developing globally responsible leaders in business schools: A vision and transformational practice for the journey ahead. *Journal of Management Development*, *32*(5), 487–507.

[3] Shanahan, M. (2022): Talking about large language models, working paper, Imperial College London.

(English) text, what words are most likely to follow the sequence 'The first person to walk on the Moon was'? A good reply to this question is 'Neil Armstrong'."[4]

This technological property has implications: A consensus seems to be settling around the claim that LLMs are especially useful for people who are *qualified to evaluate* the text output of the tool in *supporting them* to set up and start new projects. If you have worked with ChatGTP, you will have realized that depending on the prompt, the answers often seem rather generic, superficial. One reason might be that it has not yet been trained on a large enough data set in this area of expertise. Another reason is the prompt itself: if the GAI finds the closest association with the prompt, generic output is a result of "bad" prompting. Some of these issues can therefore be resolved by learning how to "prompt well." *Prompt engineering* will be an important qualification for students and teachers.[5] However, even if some one-size-fits all rules for good prompting exist, specific knowledge about the topic will most likely play an important role as well. To use the potential of GAIs, the user needs *sufficient expertise to evaluate its output* and to improve the prompts from there.

To create value that goes beyond what GAI can deliver we therefore have to focus on the development of such "meta" competencies like prompting. They encompass the ability to ask GIA tools the right series of questions and to understand and critically reflect on the output. Expressed in standard taxonomic descriptions of learning goals, we must focus on higher levels of comprehension.

2.1 Chinese rooms: what does it mean to understand something?

Which brings us to the hard problem: If GAI is most useful as a support system for people who already have the competencies to evaluate the output generated, how can we make sure and assess if students acquire these competencies if they can fake them by using AI? An example for this problem is the much-hyped debate regarding "the end of the college essay." This problem forces us to think about and reimagine the way we teach and evaluate.

At the heart of the problem of reimagining teaching and evaluation seems to be what we mean by "understanding" something (e.g., a theory). The problem is related to the so-called "Chinese room" argument by John Searle (1980).[6] He imagines himself alone in a room, following a computer program that responds to Chinese characters slipped under the door. Searle understands nothing of Chinese, and yet, following the program, he sends correct strings of characters under the door. Hence, from an outside point of view, it looks as if Searle speaks Chinese.

---

[4] Shanahan, M. (2022): Talking about large language models, working paper, Imperial College London.

[5] https://fourweekmba.com/prompt-engineering/

[6] Cole, D. (2020): The Chinese Room Argument, *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2020/entries/chinese-room/>.

Given that all our texts and arguments are usually supported by "cognitive extensions," a good starting point to address this issue is to distinguish between comprehension-supporting (CS) and comprehension-replacing (CR) functions of these technologies. We want to make sure that they are used as CS instead of CR. For the Chinese-room example, the setup prohibits to empirically distinguish whether Searle is using the computer in a CS- or in a CR-kind of way. The easiest way to solve this issue seems to be to force Searle out of the room and let him translate without support by the computer, but this radical solution may prevent him from developing the skills necessary to use it in a CS-kind of way. Hence, a more balanced approach is needed.

2.2 Mainstreaming, ideological bias, and bullshit

But there is another challenge that is tied to the specific technology. How do we make sure that there is no uniformization of writing skills—and even more problematic—theoretical and empirical interpretations of reality, given that LMMs "mainstream" text in the way described above? How can we distinguish between valid arguments and *bullshit*[7]? Three aspects must be considered:

1. For example, if GAI gives you the most "common" interpretation of the effects of monetary policy on, e.g., unemployment, what makes sure that this specific process of aggregation reflects the scientific consensus? How can we make sure that heterodox views are not "cancelled?" How can we make sure that all credible scientific views on a problem are correctly reflected in the generated texts? Users (especially if they lack the evaluative competencies mentioned above) may tend to go with the first answer generated.

The convenience GAI's offer, and the impressive smoothness of their texts and their authoritative tone can be a real danger. It seems plausible that one of the main challenges of dealing with this technology is psychological, as we are confronted with a technology that mimics thought without possessing it, which creates a powerful urge to anthropomorphize it.[8]

Google scholar or other traditional search engines as well as new aggregator sites like *Consensus* are less convenient but also less biased as they give you at least a list of potential hits (even if the sorting is problematic as well), even it leaves the tedious task of sensemaking to the user. Given the effortlessness and convenience of GAI-generated text, there is always the risk of creating "ideological biases" by accepting the output generated by the model. It seems to us as if this tendency can only be overcome by either giving the program very specific tasks (for which one needs competencies) or by support from human teachers.

---

[7] Frankfurt, H. (2005): On Bullshit, Princeton: Princeton University Press.

[8] Shanahan, M. (2022): Talking about large language models, working paper, Imperial College London.

2. In a sense, the above problem is not the most problematic challenge. As LLMs "guess" text from statistical properties of the data set that has been used to train it, it has by design no sense of correctness or falsehood of the generated text. We can expect an LLM-generated text to be "true" (in accordance with facts, …) only if we can rely on the "accuracy" of the data set and make the bold claim that statistical frequency and "truth" are perfectly correlated. IHE's must therefore take an *epistemic* stance when they decide on the future role of LLMs in research and teaching, as they are to alien to standard epistemic criteria used in science.[9] On a more pragmatic level, we are back at where we have been before: To use the potential of GAIs, the user needs *sufficient expertise to evaluate its output* to be able to separate credible output from fakes and bullshit.

3. Content is one thing, argumentative style another. And this area is where GAI-tools seem to be playing to their strengths: one of the bottlenecks of today's system is to give individual feedback to students because it is very time consuming. GAI may provide a (partial) solution to this problem. GAI-generated texts can be used to teach basic competencies regarding argumentation, rhetoric, etc., in a very time-efficient way. But again, the tendency to "mainstream" as well style may quite rapidly lead to uniformity. This may be better than a situation when students never learn how to argue coherently, but for more advanced skills, human support is again necessary.[10] Ball (2023) puts it nicely: "When we learn good use of language, we are not simply being trained to conform to a model. Those templates for sentence or essay construction do not follow some law of literature demanding particular arrangements of words, phrases, and arguments."

3. What are the likely consequences for IHEs?

3.1 Selection of faculty

If the above conjectures are correct, two challenges become visible. First, academic careers are still mainly built on research credentials. This time-approved model of selection has been adequate for as long as universities were the more or less exclusive access points for knowledge and content was decisive. The way we are teaching did not change very much over the years as its main role was to give students access to knowledge. Hence, selecting professors according to their research potential pretty much solved two problems at once. Digitalization changed that picture, as—except for basic research—access to knowledge became ubiquitous for everyone with access to the internet. GAI furthers this general trend. What becomes more and more important is no longer *what* we teach, but *how* we teach it. But at present, faculty is

[9] The epistemic bullshit-problem is especially relevant for LLMs as they *generate text from other texts*. Other GAI-technologies may be less prone to this problem, like for example pattern-recognizing AI that can help identifying skin cancer.

[10] A solution to "mainstreaming" can, of course, be the technology itself. We will see if it is technologically possible and practical to train GAI to seek for balance in its production of output.

usually not selected to excel in this dimension. Hence, we must reassess the necessary qualifications for academic teachers, train the existing faculty to be able to "teach up" to the new challenges and rethink the criteria for hiring new faculty.

The ability to foster the development of *epistemic, social, and personal virtues* like curiosity, critical thinking, sociability, responsibility, intrinsic motivation, and resilience are key qualities of good teaching in interaction with digital tools. More and more universities offer separate career paths for teaching and research. If our analysis is correct, the teaching track is much more than a second-class alternative for "failed" researchers. Driven by technological progress, excellence of teaching requires qualifications that go far beyond scientific excellence. Given the fast rate of technological progress, we must also continuously reassess which qualifications are necessary and find ways to continuously train faculty. The "teacher"-career path therefore requires the ongoing reassessment of the best teaching and examination formats. Universities therefore not only should engage in financial investments in these tracks but also to actively search for personalities who are qualified in the "how" dimension and to create a culture of learning and critically reflecting the best teaching and examination techniques. Hence, the ideas communicated in this paper are necessarily speculative, educated guesses at best. Relatedly, the need for continuous development of teaching and examining models not only requires the necessary faculty and culture, but also an internal supporting "ecosystem" including organizational support for experiments, labs, staff for technical support, etc.

3.2 Teaching and examining

With the more widespread availability and hype around of chatbots starting last November, a lot of universities felt pressured to develop guidelines and best practices to minimize the risk of students handing in essays generated by GAI tools.

Examples are: "customize" writing assignments, break major assignments into smaller, individually graded chunks, prioritize on-campus exams, test assignments by grading the output generated by a chatbot and modify if necessary, require heavy citations, specify your policies regarding AI explicitly, return to time-honored oral exams, among others.

These "quick fixes" were mostly driven by the fact that the new technology became available during the lecture and examination period, which created the need to act quickly. As a result, one might get the impression that GAI poses a threat and not an opportunity. However, it is rarely the case that we are gravitating towards a good long-run solution if the fire brigade is out. For example, before we rush back to oral exams, we should remember that they are plagued by all kinds of examinator bias.[11] Or, to give another example, it seems clear that the responsible use of GAI as part of academic integrity requires adequate standards of use. In this context, it is sometimes

[11] Joughin, G. (2010): A short guide to oral assessment, in: Leeds Met Press. Yaphe, J., S. Street. (2003): How do examiners decide? A qualitative study of the process of decision making in the oral examination component of the MRCGP examination. Medical Education 37:9, 764-771.

mentioned that plagiarism software is not able to detect GAI-generated text. But even if we would (and over time we are pretty sure we will), we must rethink the whole idea of plagiarism to deal with this new phenomenon.

That "ChatGPT can write a decent essay for the lazy student."[12] points towards a deeper problem than just regulating its use and fixing exams. The fact that even today's GAIs can compile better essays than many students shows that (a) IHEs are apparently not very good at training their students and (b) the type of essay we are expecting from our students seems to be very well defined: *the fact that LLMs excel at generating good essays implicitly reveals that what we are expecting from our students is sequence prediction, is mainstreaming*. We can take this revelation to ask if this is the right kind of objective our students should strive for. And if we conclude that such skills are misplaced, we must ask what we can do about it, which boils down to teaching skills and incentives. Given that we have taught and designed assignments along these lines for so long, we should not rush to conclusions, as we simply do not know how to teach otherwise. However, "this is how we've always done it" should not count as a valid argument for perpetuating the present model.

To summarize, as important and well-meaning these quick fixes are, it is important that we do not stop here but take the new technology as an opportunity to learn how to integrate it effectively. High-quality teaching and evaluation methods will likely be more time consuming, at least for the time being, until the dust of the new technologies will have settled. Even if GAI tools can be used to overcome some of the scaling problems like providing feedback to students, it seems as if we will have to interact with our students in more meaningful and individualized ways on average then we currently do.

3.3 A watershed moment for IHEs?

The increasing use of GAI has the potential to further increase the gap between low cost/low price IHEs which focus on teaching basic knowledge, and institutions that invest continuously in learning innovations to teach the above-mentioned competencies and enable their graduates to deliver value beyond what machines can do. This development has already been driven by the high costs of funding basic research as well as e-learning and other developments which disrupted the traditional academic value chain from research to teaching to executive education and training of junior faculty. To qualify students to make societal contributions that go beyond what GAI and other technologies can produce needs teaching and examination formats that are more interactive, individualized, and focus on personality development, like the ones outlined above. These formats, as a side effect and at least for the time being, rely on human beings as enablers of learning

[12] Ball, P. (2023): ChatGPT Is a Mirror of Our Times, What language AIs make up for in efficiency they lack in humanity, https://nautil.us/chatgpt-is-a-mirror-of-our-times-258320/?_sp=3eab95a5-7fce-41f2-9dd2-3e0fe767219d.1674166639605

processes if the model of education is based on the three pillars mentioned above. Digitalization allows it to support and even replace some traditional teaching and examination formats, and GAI tools will provide additional means to support students. The tools will not replace human beings in education, but they make it necessary to reassess their most productive roles.

The key opportunity presented by the challenges posed by the availability of GAI tools is to challenge IHEs and society at large to discuss what kind of societal contributions they expect from their future graduates and how best to achieve those goals.

Thomas Bieger is professor of business administration and director of the Institute for Systemic Management and Public Governance at the University of St. Gallen. He served as the university's President between 2011 and 2020.

Martin Kolmar is Professor of economics and director of the Institute for Business Ethics at the University of St. Gallen.