Responsible use of artificial intelligence for research purposes

Table of contents

Note : This page will be regularly updated with new information as the landscape continues to evolve. It is recommended that you always consult the most recent version of the guidelines.

For questions about these guidelines and the use of AI tools in research, please contact AI4research@vub.be .

1. Introduction

1.1 General

Artificial intelligence, often abbreviated as "AI", is a collective name for several technologies that perform tasks or behave in an "intelligent" manner. The organisation for Economic Cooperation and Development (OECD) defines an Artificial Intelligence System as *"a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy"* (1,2) .

Developments in AI have recently gained momentum. The launch of OpenAIs ChatGPT in November 2022 has created high expectations about the possible deployability of generative AI tools (hereinafter " GenAI ") for a variety of purposes. Numerous alternatives to ChatGPT have now been developed, such as Anthrophic's Claude and Meta's LLaMa. Various plug-ins have already been created for some of these AI tools, which further increases the range of functionalities and therefore the large-scale usability of these tools. These tools can process not only text input, but also images (e.g. DALL-E, Midjourney) and datasets (e.g. ChatGPT Advanced Data Analysis) (3,4) .

It remains unclear exactly how the above-mentioned AI tools can be used. However, it is undeniable that this evolution creates many new opportunities to (partially) strengthen and automate processes within research. In the future, researchers can expect that the range of available AI tools (and associated plug-ins) will only continue to expand. In time, these AI tools will also be integrated into commonly used software programs and databases. The benefits are clear: Researchers' resources and time are limited, so using these tools will ensure that more can be achieved with the same resources (in other words, increasing productivity). This is especially true for activities or processes that are time-consuming and repetitive, and require only limited insight from researchers. Outsourcing these tasks to AI tools will ensure better use of scarce research resources, benefiting both researchers and taxpayers.

In addition, these tools can also improve the quality of scientific work. Here, AI tools should be seen as *a tool that promotes researchers' own capacity to create high-quality scientific work.* For example, AI tools can help researchers come up with new ideas and improve self-written texts, especially for non-native speakers of English.

Nevertheless, AI tools also raise questions about what exactly their responsible use entails. With these policy guidelines, the VUB wishes to develop a strategic vision regarding AI tools and to provide more clarity about what our university considers appropriate and inappropriate use of AI tools within research processes. The VUB supports and encourages the use of AI tools by researchers. Our university would also like to point out to researchers the limitations and risks that the use of AI tools in research processes can entail.

1.2 Scope of guidelines

These guidelines apply to the **use of generative AI tools for research purposes.** We mainly focus on the use of commercially/publicly available AI tools regardless of their scope (broadly applicable or with specific purposes) and type of input/output (e.g. written text, figures, data). The guidelines also apply to AI tools that are available in software programs through plug-ins. It

includes both online and offline use of commercially/publicly available AI tools (e.g. within a closed environment).

These guidelines do, strictly speaking, not apply to processes associated with *development of AI tools* as part of academic activities (as well as the reuse of in-house AI tools by VUB colleagues). In addition, the guidelines exempt the *use of AI techniques for data analysis if data are not made available through an external portal.* AI tools for data mining that are formally part of the research method or design are not considered. These guidelines also do not apply to educational activities at VUB, including the organization of master's theses. For more information around AI and education, please visit Generative Artificial Intelligence and Education. Some parts of these guidelines may remain purely informative for these matters.*

**The elements taken into account in this scope will be further examined with VUB researchers. The scope can therefore be adjusted or specified in the future.*

2. General principles

The VUB **is open to using AI tools** for various research purposes and imposes few specific restrictions.*

**Note: If AI tools are made available, for example through open source licenses, applicable legislation and in particular the AI Act of the European Union must also be taken into account. This prohibits AI that poses a threat to humans (e.g. AI with unacceptable risk). With high-risk AI, legislation and applicable safety regulations must be respected in terms of safety, health, well-being and the protection of human rights.*

However, deploying AI tools for research purposes should **always adhere to a set of general principles**. These principles are:

- **Researchers remain fully responsible for AI outputs as well as the appropriateness of the research process in which AI is used.** Responsibility cannot therefore be passed on to AI tools. In practice, this means that researchers can be held accountable for inappropriate use of AI tools, such as failure to comply with the rules regarding scientific integrity (plagiarism, citing sources, etc.). This rule is consistent with the position of the *Committee of Publication Ethics (COPE)* that *"authors are fully responsible for the content of their manuscript, even those parts produced by an AI-tool, and are thus guilty for any breach of publication ethics"* (5) . ALLEA's 2023 *European Code of Conduct for Research Integrity* requires researchers to report the use of AI tools in accordance with the standards of the scientific field to which they belong (6) .
- **The creative ideas and input of researchers remain important.** The interaction between researcher and AI tools can generate new ideas. Researchers can build on their own ideas or gain new insights through AI tools, even if the insight is not explicitly presented by the AI tool. With regard to ideation, AI tools are best regarded as (versatile) instruments that can be an extension of the researcher: They promote the capacity of researchers to generate ideas.
- **Researchers must provide the greatest possible degree of transparency** on the use of AI tools in their scientific activities. This means that the way AI tools are used often needs to be described in detail in the Materials & Methods section and/or Supplementary Data, including prompts, output, name and version of the AI tool. Researchers should also properly reference AI tools in scientific publications (See section *"How to reference AI?"*)

- **Researchers need to check the outputs of AI tools**. For example, various verification steps must be built in. One should always look for the original sources and, where applicable, check licenses for the original work. Original ideas must also be properly attributed to their creator. Training data used by AI tools may also be copyrighted or reuse may be restricted by licensing. In that case, one cannot literally copy the AI output. In addition, it is necessary to provide information (mainly *facts and figures*) to *fact -check*. Researchers should check whether outputs reflect sufficiently diverse views (e.g. conflicting results of studies). In addition, one must check whether scientific statements are sufficiently supported by empirical material (if relevant to the research discipline). The more "autonomous" AI tools work, the greater the degree of verification that researchers have to perform. The greater the risk of potential harm and the amount of harm that could be caused, the greater the degree of scrutiny researchers must exercise.
- Researchers must follow possible **discipline-specific rules regarding the responsible use of AI tools as well as guidelines from scientific journals.** Some journals prohibit the use of AI tools for text production, while others allow it conditionally. Researchers should inform themselves about these guidelines and follow them if they want to publish in a specific scientific journal (See section "Referring to the use of AI tools in research processes").
- Researchers should be aware that **using AI tools can also be a disadvantage in some circumstances.** For example, AI tools can generate text with generic statements. The lack of specificity in the text and/or the use of AI tools itself (about which researchers should be transparent) could also play a role as negative elements in review processes.
- Researchers serve **not to enter the following data in AI tools** :
    - Personal data: data that can (directly or indirectly) lead to the identification of persons
    - Data that is important with a view to future valorization or data that is (or can be) protected by intellectual property law. See also "Text Prompts and User Inputs in AI Models" under the "Limitations and Potential Issues" section. Article 3 of the Valorisation Regulations states that *"Before publication, Valorisable Research Results must be made known to the Knowledge & Technology Transfer Interface of the Institution"*. Contact techtransfer@vub.be with questions or for support.
    - Data whose release could be ethically problematic, such as in the event of a real risk of inappropriate use or group harm (e.g. stigmatization)
    - Data where existing contracts (e.g. *non-disclosure agreements*) or rules from funding organizations impose restrictions on sharing
    - Data subject to contracts with third parties (e.g. companies)
    - Data that is protected by copyright or *database rights* of another person, unless with that person's permission.

  Exceptions: This general guideline may be deviated from for certain AI tools and under certain conditions. Please contact AI4research@vub.be if you have any questions about possible exceptions.
  We recommend that researchers **check the privacy settings of the AI tools** and adjust them if deemed necessary. This is how you can get *Chat History & Training at ChatGPT* to expand.

- It is **not permitted to use AI tools to write the substantive aspects of peer reviews.** In contrast, researchers are allowed to:
    - the *lead author (s)* to use AI tools themselves to generate feedback.

- Use AI tools to better formulate self-written peer reviews linguistically (e.g. Grammarly) if measures are taken to ensure confidentiality.

3. For what research purposes can AI tools be used?

3.1 Research purposes

This section serves to give researchers more insight into how AI tools can be used. The following table describes possible purposes. The purposes listed are non-exhaustive. Each purpose is illustrated with a short case study that also provides the *Dos & Don'ts* when using AI tools. These *Dos & Don'ts* are more specific guidelines that should be observed in addition to the "General Principles". Finally, a series of AI tools are linked to each purpose. This list is also not exhaustive. These AI tools can be applicable for specific or multiple purposes. There can be significant differences in performance of these AI tools, and we recommend always using the *state-of - the -art AI tools*. In practice, this often means using the paid versions (via individual or group licenses). If there are any further questions about these cases or use of AI tools for other research purposes, please feel free to contact us at AI4research@vub.be .

Purposes	Type	Case	Do's	Donts	Examples of AI tools
Identifying new literature	Information retrieval	When conducting a literature study, an AI tool is additionally used to identify relevant literature in lesser-known fields (e.g. in interdisciplinary research).	- Appropriate mention of AI tool in "Method" section - Free use if not within the context of a formal literature study	- Using AI-suggested papers in literature review without content checking	- Elicit (http://elicit.org) - Perplexity (https://www.perplexity.ai/) - Inciteful (http://inciteful.xyz) - Research Rabbit (http://researchrabbit.ai) - Connected Papers (http://connectedpapers.com) - Consensus (http://consensus.app) - Scite (http://scite.ai) - Iris (http://iris.ai)
Summarizing articles	Word processing	The researcher uses an AI tool that screens and summarizes recent literature. The summaries are then reviewed in detail.	- Use for initial screening before reading the full texts	- Citing/referencing articles based on a summary without having read it yourself	- Semantic Scholar (http://semanticscholar.org) - Humata (https://www.humata.ai/) - Iris (http://iris.ai) _ - ChatGPT (http://chat.openai.com) _ - Perplexity (https://www.perplexity.ai/)
Selectively searching the content of articles	Information retrieval	The researcher uses an AI tool to find "Recommendations" and "Perspectives" in scientific publications to keep up with the most recent policy implications of studies within his/her domain.	- Screening for specific parts of articles or terminology and then reviewing them manually	- Summarizing specific parts across articles and not reading these individual parts thoroughly - Lack of references when processing in own academic work	- Iris (http://iris.ai) - Semantic Scholar (http://semanticscholar.org) - ChatGPT (http://chat.openai.com) _
Formulating scientific propositions	Word processing	The researcher enters his or her own scientific statements and hypotheses into an AI tool to receive feedback. The AI tool suggests	- Indicate use of AI tool to rework statement, including prompts and outputs , if hypothesis testing is part of scientific publication		- ChatGPT (http://chat.openai.com) _ - Elicit (https://elicit.org/?via=topaitools)

		possible improvements (e.g. substantive improvements or clarification)			
Criticizing statements or positions taken	Ideation	The researcher uses an AI tool to criticize a statement. This can be used to gain more insight into the counterarguments that can be raised.	- Self-test whether all arguments that undermine the position can be refuted - Possible use to educate students through debate	- Citing very personal stories from one's own life to try to support one's own position	- DebateDevil (https://www.debate-devil.com/en) - Elicit (https://elicit.org/?via=topaitools) - Scite (https://scite.ai/?via=topaitools) - ChatGPT (http://chat.openai.com)
Carrying out a preliminary screening of evidence for and against scientific statements	Information retrieval	The researcher uses an AI tool to screen the results of scientific publications. These results can be positive or negative relationships between certain variables. In this way, the researcher wants to gain insight into the burden of proof for a particular hypothesis.	- An initial screening of evidence and shortcomings of studies, before thoroughly reviewing them later - Additional screening during systematic review and meta-analysis of studies - Take *publication bias* in favor of studies with positive results into account during this exercise	- Use of output from AI tool as "burden of proof" for or against a statement in social debate - Complete dependence on a tool for selecting studies for a *systematic review*	- Consensus (https://consensus.app/search/) - Perplexity (https://www.perplexity.ai/) - System Beta (https://www.system.com/landing) - Scite (https://scite.ai/?via=topaitools) - ChatGPT (https://chat.openai.com/)
Brainstorming	Ideation	The researcher submits his or her own ideas to an AI tool and asks whether they are already covered in existing literature. The	- Mention interactions with AI tool to generate ideas in publications, including prompts and outputs	- Avoid introducing completely new ideas if you anticipate to engage in valorisation later	- ChatGPT (http://chat.openai.com _ - Elicit (https://elicit.org/?via=topaitools)

		tool finds some potentially valuable publications that partly answer the research question.			
Generating new ideas for manuscripts	Ideation/production of new works	The researcher uses an AI tool and receives an output that contains an idea/argument that is suspected to be new. The prompts provided by the researcher did not actively generate this idea.	- The origin of the output is determined via multiple prompts. The researcher also searches through other sources. If necessary, the original source is referenced. The prompt of how the idea was arrived at is added as Supplementary Data in the Method section - The idea is used as an impetus/starting point	- Copying ideas and presenting them as one's own intellectual contribution - No checking whether ideas come from other sources and no acknowledgment of sources	- ChatGPT (https://chat.openai.com/)
Generating an initial structure for writing subparts of a scientific publication	Word processing	The researcher uses an AI tool to create an initial structure for the introduction section. Afterwards, the researcher writes the text himself based on this structure.	- The use of AI tools is mentioned in the Method section of the publication		- ChatGPT (https://chat.openai.com/)
Generating a text by entering the initial structure of subparts of a scientific publication	Word processing/Producing new works	The researcher gives the titles of five paragraphs with the instruction to write six lines per paragraph. The	- The researcher must have read all cited publications and check whether the quotation is an	- Copy-pasting text in your own publications without checking for plagiarism, for	- ChatGPT (https://chat.openai.com/)

		AI tool generates a general text with source information.	accurate representation - The researcher checks whether the paragraphs present sufficiently diverse views or evidence (e.g. studies with contrasting results) - Use of the AI tools is reported in the Method section of the publication and the prompt and output are shown in Supplementary Data	possible licenses that limit reuse	
Correcting spelling errors, grammatical errors and structure of paragraphs based on your own text input	Word processing	The researcher uses an AI tool to improve their own text. The tool makes suggestions to improve the structure of paragraphs and corrects grammatical and spelling errors. In addition, the AI tool suggests replacing vague terms.	- The use of AI tools is mentioned in the Method section of the publication	- Automatically accepting all possible adjustments without checking whether the content has changed	- Writefull (https://www.writefull.com/) - Quillbot (https://quillbot.com/) - ChatGPT (https://chat.openai.com/)
Analyzing research data	Data analysis and visualization	Various AI methods are already available to analyze research data. This doesn't necessarily have to be about generative AI. Analyzes	- Performing analyzes with AI with sufficient substantive control and in accordance with ethical (e.g. checking for bias in	- Performing analyzes based solely on prompts without checking whether the prompt has been interpreted correctly	- Code Interpreter in GPT-4 (https://chat.openai.com/?model=gpt-4-code-interpreter) - AutoML

		can possibly be done based on prompts within existing software.	data sets) and legal principles - Mention of AI tool in Method section	- The *post-hoc* formulation of hypotheses based on a given data set so that these hypotheses are always confirmed	
Code generation	Producing new works	The researcher uses an AI tool to help write parts of software.	- Mention of AI tool in Method section	- Copying entire pieces of code without checking the feasibility of the code and whether the desired result is achieved	- GitHub Copilot (https://github.com/features/copilot) - ChatWithGit plugin in GPT-4 - AskTheCode plugin in GPT-4 - Code Whisperer (https://aws.amazon.com/codewhisperer/) - Code LLama (https://ai.meta.com/llama/)
Generation of data	Producing new works	The researcher uses AI to generate synthetic data while maintaining statistical correlations for exploratory research.	- Publications mention that synthetic data was used and how it was produced in the Method section. - Use of AI tools is only acceptable if it is clear what happens to data inputs	- Falsifying research data (fraud)	
Science communication	Writing assistance for administrative/additional tasks	The researcher uses an AI tool to summarize scientific texts for the general public. These messages are shared on social media and the research group's page.	- Checking AI summaries before they are sent out for correctness of the contents and style - Chatbot for citizens that can answer follow-up questions	- Excessive distribution of science communication (“spamming ”)	- ChatGPT (https://chat.openai.com/)

Writing a Data Management Plan (DMP)	Writing assistance for administrative/additional tasks	The researcher uses an AI tool to ask specific questions about completing the DMP (e.g. certain sections). The AI tool provides targeted answers about the researcher's options.	- Use of AI to better answer specific questions (e.g. which databases are suitable for data type X?) - Writing DMPs by AI if substantive control *and* understanding of DMP implications - Take into account VUB -specific guidelines regarding Research Data Management	- Avoid generating very generic DMPs with little detail	- ChatGPT (https://chat.openai.com/)
Writing project applications	Writing assistance for administrative/additional tasks	The researcher uses an AI tool to write certain parts of a project application.	- The researcher who submits the application has final responsibility for the scientific content and must be sure that it is an original proposal - Publicize the use of AI tool for generating the text - Check whether AI-generated text is sufficiently precise - Check whether the project application meets the conditions of the *project call*	- Completely outsource the writing of project applications to AI tools	- Grantable (https://grantable.co/) - Granted AI (https://grantedai.com/) - ChatGPT (https://chat.openai.com/)

3 .2 Available AI tools for research purposes

find more detailed information on how to get started with AI tools on our "Practical information for using AI tools" page.

3.3 Trainings regarding the use of AI tools for these purposes

If you are interested in helping organize classes around the use of AI tools for specific purposes, please contact us at AI4research@vub.be. If you are already organizing lessons on this topic, please also contact us. In this way we gain insight into the existing training material and so we can further build on your work and the work of others.

4. Benefits of using AI in research

There are several benefits to using AI tools for various research purposes. We can divide the benefits (and consequences) for researchers into different levels:

- **Increasing the productivity of researchers.** We consider productivity here as the amount of (desirable) research outcomes that one produces given a fixed amount of resources. Scientific studies indicate that using GenAI tools can significantly increase employee productivity (7–12) . With regard to research, we foresee that time-intensive (and/or repetitive) tasks that do not require substantial intellectual contributions from researchers to AI tools can be partly outsourced. In addition, AI tools can also strengthen researchers' own competencies (e.g. to generate ideas). AI tools can also be used to identify complex patterns in large data sets, which can be useful in the context of exploratory research. Together these can increase the productivity of researchers.
- **Increasing the quality of researchers' work.** Observations from other sectors indicate that the use of GenAI tools can help perform certain tasks in a more qualitative manner (7–9) . We equally expect that researchers will be able to achieve their stated objectives in the context of research in a more qualitative manner. In this way, a literature study could be made more comprehensive. AI tools can also foster researchers' insights, creativity, and self-reflection, which can lead to more high-quality work.
- **Promote reflection on one's own work.** Just like your fellow scientists, AI tools can cast a new light at your own beliefs, arguments, ideas and works. However, all kinds of factors can complicate the process of giving and receiving feedback. AI tools can therefore have the advantage of quickly taking an "external and impersonal" look at one's own work.
- **Lead to greater agility among researchers.** Within today's research environment, researchers are often confronted with challenges that require new skills. For example, researchers must learn to write project applications, draw up DMPs and deal with data protection legislation and ethical guidelines. A significant barrier here is that the learning process can take a lot of time. We expect that the availability of AI tools can reduce the learning time required for "new" tasks. They can therefore increase the researcher's agility to deal with these challenges.
- **Bringing about a shift in the daily work of researchers.** One study shows that AI, like other technologies, can cause shifts in job roles (9) . For researchers, this range of tasks consists of coordination, core research tasks, teaching tasks and administrative tasks. AI tools can be used for all these tasks. However, we anticipate that, at this time, they

will be more likely to be deployed for repetitive tasks of a non-intellectual nature (e.g. administrative work). They can also be used for more complex tasks, mainly in collaboration with human beings. If we assume that repetitive tasks will require less time (and they will not increase in number), then more complex tasks will account for a larger share of daily work in the future (e.g. core research tasks).

- **For certain skills, the gap between researchers could decrease.** Some studies indicate that less skilled workers benefit more from using AI tools than skilled workers (7–11) . For example, one can imagine that a skilled researcher who is not skilled in writing down his/her findings could use an AI tool to compensate for this deficiency. In other words, AI tools could have a leveling effect on the gap between researchers for certain skills.
- **Narrowing the gap between citizen and scientist.** AI tools can be used to summarize and explain scientific texts at the level of the average citizen. This helps scientists explain their research in a sufficiently understandable manner. In fact, AI tools in this way allow to circumvent an important cognitive limitation of humans (the so-called *" curse of knowledge "*) (13) .
- **Promote interdisciplinary dialogue and research.** AI tools facilitate the exploration and understanding of abundant information, such as diverse academic work across disciplines. They can also make connections across this diverse literature that are not immediately visible to researchers (In fact, this is a form of creativity) (14) . Here, AI tools can also take on the role of *match-makers* to connect researchers with each other, which can promote interdisciplinarity (15) .

At a system level, we can anticipate that these benefits for individual researchers will translate **into greater speed of scientific progress and overall societal impact of research** (given the same amount of resources) **.** This means that the planned resources used in scientific research would generally be better spent.

5. Limitations and potential problems when using AI tools in research

As described above, using AI tools can have clear benefits. However, researchers should be aware that AI tools can also have some serious shortcomings and limitations. These are explained per category below.

Training data and learning for AI models

- **Data bias can occur in training data**
  a. **Training dates may already be influenced by existing structural inequalities between groups.** Differences between groups, such as between different genders, ethnicities or socio-economic classes, can result in over- or under-representation in *real- world data* (e.g. workforce files, healthcare data). Structural inequalities in group participation or outcomes can be highly dependent on the data source and context. They are not limited to the aforementioned categories.
  b. **Training data may contain discriminatory language or racist undertones, incorrect information and argumentation (i.e. *logical fallacies*).** This is especially true when written text is used as training data for Large Language Models (LLMs).
  c. **Training data can exhibit other forms of bias** depending on the type of source. For example, *publication bias in favor of positive results* can still have an impact

on AI tools that perform an initial screening for evidence for scientific statements.

- **Training dates for AI tools may be dated.** AI tools, including LLMs , may not have access to the latest events or publications. The *knowledge cut-off* of GPT-3.5 was September 2021 because sources after this date are not included in the training data. Since September 27, GPT4 (Plus and Enterprise) no longer have this limitation and GPT-3.5 will no longer have it in the future.
- **Training data for AI tools may already be (partly) influenced by LLMs.** If the use of LLMs is widespread, this will influence the content of new texts. These new texts can in theory be reused as "training data" for LLMs . This is a form of "circularity" where there is a risk that, if the degree of human intervention is too low, machines will imitate the mistakes of other machines (16) .
- **Several parts of the training data are not equally informative.** The training data for LLMs may contain sets of information that, under normal circumstances, would have varying levels of credibility. For example, results from a single empirical study would be considered less credible than a meta-analysis of all available empirical studies. Likewise, blogs, newspaper articles and scientific publications are not necessarily given the same level of credibility. LLMs such as ChatGPT can, in principle, take these differences into account, for example through *fine-tuning* or *reinforcement learning* from human feedback (RLHF) steps during development. However, at this time, one can still get links to less "reliable sources", such as blogs, from some LLMs if one does not use plugins or specialized AI tools.
- **Training data may contain personal information or copyrighted material that has been unlawfully released.** Personal information that has been unlawfully released may be contained in training data. The use of this data may therefore constitute a (further) violation of research integrity and/or the GDPR.
- **Training data used by AI tools could be partially reconstructed under certain circumstances.** Various forms of *adversarial privacy attacks,* in which targeted attacks are made to disrupt the functioning of AI models or steal information, could pose a risk. The theoretical possibility of this has been documented for AI-tools that use personal data for data analysis (and, therefore, not necessarily for GenAI). In theory, third parties could, under specific conditions, request the presence of an individual's data in training data ("Membership Inference Attacks"), whether parts of training data could be reconstructed ("Model Inversion Attacks") (17–20) .

Text prompts and user inputs for AI models

- **Entering personal data into AI tools may violate the European Union's General Data Protection Regulation (GDPR).** Personal data entered into AI tools could be stored outside the European Union and third parties could gain access to this data. Please contact AI4research@vub.be before entering personal data into an AI tool.
- **Entering personal data into AI tools may conflict with the informed consent obtained from the research participant.** Ethical guidelines surrounding participation in research may require that consent is only informed if the risks have been made known to research participants in advance. If the use of AI tools produces risks that have not been disclosed, this may mean that the consent was not sufficiently "informed". In addition, the informed consent may include that data will not be released to commercial companies. Please contact AI4research@vub.be before entering personal data into an AI tool.
- **The use of AI tools to provide feedback on completely new research ideas could be equated with the "making public" of these ideas for specific purposes. For example,**

OpenAI employees view user conversations with their AI tool to improve the algorithm (21) . In that case, researchers may lose the opportunity to patent and valorize inventions. So be careful what you enter if you intend to obtain a patent (in the long term).

- **Your inputs into AI tools can impact the outputs others get.** Unlike the paid GPT-4, OpenAI states that the free version of GPT-3.5 " *[…] may use content such as prompts, responses, uploaded images, and generated images to improve our services. (…) ChatGPT, for instance, improves by further training on the conversations people have with it, unless you choose to disable training"* (22,23) . This means that your interactions with ChatGPT, including your prompts, can be used to improve the performance of the model. It is unclear whether this concerns basic training, *fine-tuning* or *reinforcement learning* from human feedback (RLHF). If this is basic training, your prompts/ inputs could reappear as outputs for other users. **However, there is no conclusive evidence for this at this time.** If it's just about further RLHF, then your prompts won't come back to others.
- **Your inputs themselves (e.g. quantitative data) may also be of low quality.** For example, there may have been shortcomings in carrying out the sample, making the data obtained less informative for making statements about the population (in other words, methodological shortcomings). Your data may also contain insufficient "metadata" (i.e. information about experimental conditions) to be processed appropriately by AI models. Finally, AI models cannot always handle metadata appropriately at the moment. We recommend that researchers keep the idea of " Garbage in, garbage out" in mind.

Output from AI models

- **Outputs from AI tools can contain bias.** The forms of bias described under "Training data for AI models" can influence the outputs of AI tools. In this way, Large Language Models (LLMs) can reproduce sexist, racist, or other discriminatory ideas (24–26) . In addition, fine-tuning LLMs for specific tasks that require humans to assess outputs from AI tools can introduce bias (27) .
- **Lack of insight into exactly how AI tools work.** It is often impossible to determine exactly how algorithms arrive at a certain output. Even if it is clear that a certain part of training data had a greater influence, an explanation that is plausible to humans is still lacking. For example, if one observes that a certain region has a greater influence on images, then there is no causal explanation for *why* this is the case. A lack of insight into the exact workings of AI tools can have serious consequences. This is mainly the case if unintended correlations with regard to protected characteristics (e.g. gender, orientation, religion, etc.) or *proxies* for these are identified (28) .
- **Truthfulness.** The outputs of LLMs can be presented as convincing but can be inaccurate, completely wrong or simply irrelevant (29) . A single study already showed that, under some circumstances, the use of high-quality AI can harm the quality of work if users become too dependent (mainly on tasks for which AI tools are not yet suitable). They fall asleep behind the steering wheel, as it were (30) . The self-confidence with which AI tools make statements can therefore be misleading. Researchers must therefore be very critical of text outputs from AI tools.
- **Limited degree of reproducibility.** Generative AI tools create outputs that cannot always be exactly replicated. For example, the same prompt can lead to different text fragments. This also means that complete reproducibility of research processes is not always possible, such as when entering the same prompt for a tool that recommends literature. In addition, algorithms, mainly dynamic-learning AI tools, can themselves also be

unstable in the sense that inputs do not process in the same way over time, which makes it more difficult to develop (reproducible) workflows .

- **Outputs may contain plagiarized elements.** Form elements or ideas from existing texts or images can be copied in their entirety and thus violate research integrity. Not all AI tools are able to generate qualitative references to the sources used. Correct citation in practice may involve some complexities, such as the presence of sources within other sources (e.g. citations within review paper) and attributing a misinterpreted text to a particular source.
- **AI tools can produce new data that conflicts with existing privacy rules, such as the GDPR.** There may also be (general) problems regarding *data linkage* and/or derived information where individually anonymized data can together give rise to new privacy-sensitive data. For example, AI tools can, in principle, create new information that is privacy-sensitive or that can be used as a proxy for protected characteristics (e.g., combination of income class and zip code). This also means that the original data sets were not truly "anonymous" (i.e. irreversibly deidentified), which also raises the fundamental question of whether "anonymity" can actually exist.

6. Refer to the use of AI tools in research processes

4.1 State of affairs

The widespread use of AI tools in research processes creates new challenges for the attribution of scientific contributions. There is currently no consensus as to whether AI tools are allowed and how their use should be made transparent. Scientific journals and publishers have their own positions regarding the use of and reference to AI tools. Note that some publishers impose restrictions on the use of AI tools. **We advise researchers to inform themselves about this.** This can be done before conducting the research if the use of AI tools is formally part of the method or research design. This can also be done before writing the manuscript if AI tools are used to generate text or figures. The way in which references should be made may also evolve further. A regular check is therefore recommended.

4.2 How to refer to AI?

The VUB advises researchers to **first check the guidelines of the journal or publisher** where they want to publish! Follow these guidelines if they are available.

If the journal or publisher does not provide explicit guidelines around the use of AI tools, the following general guideline can be followed:

Researchers should describe the use of AI tools in the **Method section**, including their technical specifications. These specifications include but are not limited to the full name, version, link to a repository , parameters used and configurations. If the AI tool has been described in detail in scientific publications, these papers can be cited. The prompts used and outputs obtained are described in the Method section or included as Supplementary Data. If a text output from AI tools contains text fragments or ideas originating from others, the original sources should be cited.

In **two specific cases,** reference should be made differently:

a. *AI tools are the object of the study (e.g. studying bias in ChatGPT outputs) and/or the output of AI tools is quoted verbatim*

Researchers should accurately present all prompts and cite AI outputs *verbatim*. This can be done, for example, by putting the name of the AI tool in brackets at the end of the quote. So, in principle, this does not require reference to AI tools as "author" of the quotes in the References section. If the journal uses a referencing style that allows this, researchers should follow these guidelines.

b. *AI tools create seemingly new insights or arguments without significant intellectual contribution from researchers*

Researchers should first and foremost check the origin of the output to avoid plagiarism. If there is an existing idea, the source must be cited. If the origin of the idea cannot be found after extensive searching, researchers may quote the idea, **but it must be clear that it does not come from the researcher himself.** The actions taken by the researchers to determine the origin should be described in the Method section. The prompts themselves should be added as Supplemental Data.

7. Reference list

1. OECD. Recommendation of the Council on OECD Legal Instruments Artificial Intelligence. 2022. Available from: https://legalinstruments.oecd.org/public/doc/648/1df51f15-53fc-43ef-9f13-ee9f957076bc.htm
2. European Parliament. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONIZED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS. 2021. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206
3. Midjourney [Internet]. [cited 28 June 2023]. Midjourney . Available from: https://www.midjourney.com/home/?callbackUrl=%2Fapp%2F
4. DALL·E 2 [Internet]. [cited 28 June 2023]. Available from: https://openai.com/dall-e-2
5. COPE: Committee on Publication Ethics [Internet]. [cited 29 June 2023]. Authorship and AI tools. Available from: https://publicationethics.org/cope-position-statements/ai-author
6. ALLA. The European Code of Conduct for Research Integrity REVISED EDITION 2023. Available from: https://allea.org/wp-content/uploads/2023/06/European-Code-of-Conduct-Revised-Edition-2023.pdf
7. Dell'Acqua F, McFowland E, Mollick ER, Lifshitz-Assaf H, Kellogg K, Rajendran S, et al. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. SSRN Electron J. 2023 [cited September 22, 2023]. Available from: https://www.ssrn.com/abstract=4573321
8. Brynjolfsson E, Li D, Raymond L. Generative AI at Work. Cambridge, MA: National Bureau of Economic Research; 2023 Apr [cited 26 September 2023]. Report No.: w31161. Available from: http://www.nber.org/papers/w31161.pdf
9. Noy S, Zhang W. Experimental evidence on the productivity effects of generative artificial intelligence. Science . 2023 Jul 14;381(6654):187-192. DOI: 10.1126/science.adh2586
10. Peng S, Kalliamvakou E, Cihon P, Demirer M. The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. 2023 [cited September 26, 2023]. Available from : https://arxiv.org/abs/2302.06590
11. Kanazawa K, Kawaguchi D, Shigeoka H, Watanabe Y. AI, Skill, and Productivity: The Case of Taxi Drivers [Internet]. Cambridge, MA: National Bureau of Economic Research; 2022 Oct [cited 6 October 2023]. Report No.: w30612. Available from: http://www.nber.org/papers/w30612.pdf
12. Campero A, Vaccaro M, Song J, Wen H, Almaatouq A, Malone TW. A Test for Evaluating Performance in Human-Computer Systems. arXiv ; 2022 [cited September 27, 2023]. Available from : http://arxiv.org/abs/2206.12390
13. The Curse of Knowledge: A Difficulty in Understanding Less-Informed Perspectives – Effectiviology [Internet]. [cited September 27, 2023]. Available from: https://effectiviology.com/curse-of-knowledge/
14. Mollick E. Automating creativity [Internet]. 2023 [cited September 29, 2023]. Available from: https://www.oneusefulthing.org/p/automating-creativity
15. Chubb J, Cowling P, Reed D. Speeding up to keep up: exploring the use of AI in the research process. AI & Soc. 2022;37:1439–1457. DOI: 10.1007/s00146-021-01259-0
16. Kenton Z, Everitt T, Weidinger L, Gabriel I, Mikulik V, Irving G. Alignment of Language Agents. 2021 [cited October 6, 2023]. Available from : https://arxiv.org/abs/2103.14659
17. Rigaki M, Garcia S. A Survey of Privacy Attacks in Machine Learning. 2020 [cited 6 October 2023]. Available from : https://arxiv.org/abs/2007.07646
18. Zhang G, Liu B, Zhu T, Zhou A, Zhou W. Visual privacy attacks and defenses in deep learning: a survey. Artif Intel Rev . August 2022;55(6):4347-401. DOI: 10.1007/s10462-021-10123-y

19. Balle B, Cherubin G, Hayes J. Reconstructing Training Data with Informed Adversaries [Internet]. arXiv ; 2022 [cited 26 June 2023]. Available from : http://arxiv.org/abs/2201.04845
20. Carlini N, Tramer F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, et al. Extracting Training Data from Large Language Models. 2020 [cited October 6, 2023]. Available from: https://arxiv.org/abs/2012.07805
21. What is ChatGPT? | OpenAI Help Center [Internet]. [cited 10 July 2023]. Available from: https://help.openai.com/en/articles/6783457-what-is-chatgpt
22. Data usage for consumer services FAQ | OpenAI Help Center [Internet]. [cited September 29, 2023]. Available from: https://help.openai.com/en/articles/7039943-data-usage-for-consumer-services-faq
23. How your data is used to improve model performance | OpenAI Help Center [Internet]. [cited September 29, 2023]. Available from: https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance
24. ChatGPT and Large Language Model bias. Available from: https://www.cbsnews.com/news/chatgpt-large-language-model-bias-60-minutes-2023-03-05/
25. Omiye JA, Lester J, Spichak S, Rotemberg V, Daneshjou R. Beyond the hype: large language models propagate race-based medicine [Internet]. Health Informatics; 2023 Jul [cited 6 October 2023]. Available from: http://medrxiv.org/lookup/doi/10.1101/2023.07
26. Deshpande A, Murahari V, Rajpurohit T, Kalyan A, Narasimhan K. Toxicity in ChatGPT: Analyzing Persona-assigned Language Models. 2023 [cited 6 October 2023]. Available from : https://arxiv.org/abs/2304.05335
27. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. arXiv ; 2022 [cited September 27, 2023]. Available from : https://doi.org/10.48550/arXiv.2203.02155
28. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science . 25 Oct 2019;366(6464):447-453. DOI: 10.1126/science.aax2342
29. Hosseini M, Rasmussen LM, Resnik DB. Using AI to write scholarly publications. Account Res . January 25 , 2023;1-9. https://doi.org/10.1080/08989621.2023.2168535
30. Fabrizio Dell'Acqua . Falling Asleep at the Wheel: Human/AI Collaboration in a Field Experiment on HR Recruiters. Available from: https://static1.squarespace.com/static/604b23e38c22a96e9c78879e/t/62d5d9448d061f7327e8a7e7/1658181956291/Falling+Asleep+at+the+Wheel+-+Fabrizio+DellAcqua.pdf