Technical, legal and ethical checklist for the use of AI at UvA/VU

> *Status: This document presents a **proposal** for a structured way to manage the ethical issues around the purchase of AI-based systems and services at the UvA and VU. It was written by the* Taskforce for AI in education*. Most of the document serves as an example of the* type *of structure we envision, and as a starting draft, rather than an exact specification. If the proposal is accepted, the exact details will be finalised in conversation with the various stakeholders.*

Section 1: Intended use

Universities aspire to an adequate level of control over the variety of AI models that are deployed in their organisation. Moreover, they have substantial influence over the development of AI systems purely through their purchasing power, an influence which should be wielded responsibly.

To facilitate these aims, this document provides a checklist for the UvA and VU to assess any potential purchase of an AI system or service. The checklist is made up of three parts: pre-flight checks, required checks and optional checks. These are processed in order: the pre-flight check should be completed and passed before the required checks are performed, and the required checks should be passed before the optional checks are performed.

Pre-flight checks These allow the university to evaluate whether a proposed solution should be implemented at the university at all, or whether doing so would clash with the core values of the institution.[1]

Reasons why a particular technology can clash with university values can be that a) the supplier is unacceptable (e.g. they contribute to systematic human rights violations, are financed by a rogue state, are under suspicion of espionage, etc.) or b) the technological solution as such as unacceptable (e.g. real live biometrical identification of students and staff; work place monitoring, exposing students and staff to risk for their safety, health or fundamental rights or banned uses under the AI Act), c) the particular technical solution is unacceptable (e.g. because strong suspicions of bias, trained on unlawful date, etc.) or the conditions under which the solution is offered is unacceptable (e.g. because it threatens the sovereignty of the university or the independence of academic research).

If the pre-flight check fails on any of these items, the university will not purchase that solution. Under exceptional circumstances, the university may consult with a group of representatives of the university and expertise in AI and human rights to find out whether there are circumstances in which the solution should be adopted nevertheless, and if so, under which conditions, or if further investigation is necessary. If needed, this group will do more in-depth investigation. This committee should include at least representatives of the student, teacher and researcher community, as well as one person with expertise in the area of AI and one person with expertise in ethics and human rights. The universities

[1] The core values of the VU are summarized here and those of the UvA in here. For the purposes of this document, we include in "core values" also those that are not explicitly stated. For instance, safeguarding the privacy of students is clearly in line with a core value, even if it is not addressed in any mission statement or strategic frame. See Appendix A for details.

designate a suitable committee. This could be an already existing committee, such as the ethics committee, or a newly erected committee.

In addition to the choice of whether or not to purchase, this process may lead to a clear set of limitations or conditions placed on the use of the product, if it is purchased.

Required checks These are used to evaluate whether a supplier and their product are suitable for the university. If any of the required checks fail, the university should either:

- Check with the supplier if the product may be amended to suit the checklist.
- Choose an alternative supplier.

If the above are not feasible, or there is an urgent need to use the technology when no supplier satisfies the demands of the checklist, a **pilot programme** may be set up. In this case, the technology may be deployed in a limited scope. The following conditions should apply:

- The technology is deployed for a short, fixed term, no longer than two years. The contract with the supplier is also short-term.
- The use of the technology will be empirically evaluated. The evaluation must follow strict guidelines.
- Students and employees will not be forced to use the technology. There is informed consent from those who do, and an opportunity to opt out of the use of the technology.

Detailed requirements for pilot programs are given in the document *Advies randvoorwaarden pilots en exploratief onderzoek over experimenten AI in onderwijs.*

Optional checks If all required checks are passed by multiple suppliers, the optional checks may further assist the university in deciding among them.

The required and optional checklist may be shared with the supplier so that they may offer detailed answers for each check. Failing that, the university may answer the checklist itself, based on the best available information. In this way, the universities are commited to uphold their core values, while also providing clear incentives for suppliers towards the production of more ethical and responsible AI.

A. Scope

This checklist should be used whenever a university (currently the VU and UvA) is considering purchasing, licensing or offering to its students or employees a product that makes substantial use of modern AI technology.

The following are examples of attributes that put a system within scope of the checklist. This is not an exhaustive list.

- **High to limited risk** Any model that falls under the categories *high risk*, *limited risk* and *general purpose AI*, as defined in the European AI act. See Appendix A for definitions. The **Unacceptable risk category** automatically disqualifies a system for university use (see also the pre-flight check).
- **General purpose** A system which contains an open-ended interface in which many different goals can be accomplished using a human mode of communication. For example, a textual chatbot, or spoken-word natural language interface.

- **Model size** The inclusion of a machine learning model of more than 6 billion parameters, a model trained on more than 100GB of textual data, or a model trained using more than 10^{22} FLOPS of computation.[2]
- **Potential bias** The inclusion of an automated component with potential biases that have not been or cannot be fully quantified. For example, a face recognition algorithm which may detect one skin colour less well than others.

Note that it should not be left up to the supplier or their marketing materials to decide whether a system “contains AI.” For example, a face recognition component is common in many products that do not market themselves as AI systems.

[2] These boundaries are necessarily somewhat arbitrary. However, in current architectures, 6B parameters appears to be a relatively clear delimiter between the models that are only usable in low risk applications and models that can generate convincing responses to user queries and commands. For example, the “instruction tuning” behavior of generalizing from one task to another appears at this point . Under the widely used Chinchilla training regime , this model size requires about 100B tokens of training data (approximately 250GB) and 10^{22} FLOPS of computation.

Section 2. The Checklist

A. Pre-flight check

These are questions not to the vendor, but to the university itself. For certain opportunities for automation, the university should first consider whether automating the issue at hand is in line with the values of the university to begin with, independent of which supplier is chosen. Such questions are relevant, for example, in the case of

- (Partially) automated remote proctoring,
- (Partially) automated marking of homework or exams,
- (Partially) automated study recommendation (studieadvies), plagiarism or fraud detection.

Given the scale of modern teaching there is a strong incentive to look for automation. Doing so must never be to the detriment of core values like safeguarding privacy or protecting students from biased decision making.[3]

1. Is the technological solution as such unacceptable?
    a. Example 1: the solution is on the list of banned uses under the European AI Act.
    b. Example 2: the solution risks exposing students and staff to unacceptable risks for their safety, health or fundamental rights
2. Is this particular technological solution unacceptable?
    a. Example 1: there are strong indications of bias in either the training data, the model or the outcome.
        i. Example: A proctoring system based on face recognition software that work less well on black skin than on white skin, dure to the underrepresentation of black skin in the training data.
    b. Example 2: There are strong indications that the AI model has been trained on unlawful data.
    c. Example 3: There are strong indications that the solution has not been trained with respect for the planet and global resources
    d. Example 4: there are strong indications that the solution has been developed in a situation that disregards workers' rights
3. Are the conditions under which the solution is offered unacceptable?
    a. Example 1: Switching to another solution is made disproportionally difficult with the consequence of a lock-in situation that can threaten the sovereignty of the university
    b. Example 2: the conditions limit the ability to criticise or do critical research into that particular solution or its providers
    c. Example 3: the conditions are in conflict with the intellectual property rights or rights to data protection of students and staff
4. Does the proposed solution automate *actions*, which may impact students' or employees' lives, fundamental rights, or the university's core values?

[3] Note that all these examples are considered *high-risk* under the AI act.

    a. If so, does the product require or allow a human-in-the-loop? Is this human safeguard strictly enforced or is it likely to be reduced to a mere formality over time?
    b. For example: a system which automatically grades students' homework.
5. Does the proposed solution make *predictions*, which are likely to be translated to actions which impact students' or employees' lives, fundamental rights or the university's core values?
    a. Example 1: A plagiarism or AI "detector" which uses external features, such as a student's profile, changes in their style of writing, or speed of writing in an editor.
    b. Example 2: A system which predicts whether a student is likely to attain the required number of ECTS in the first year.
        i. This may be translated into very different *actions*: either more support, or an official advice to quit early. The choice of action determines whether there is a positive or negative feedback loop from a biased prediction.
6. Are the actions or predictions that the product makes likely to be biased with respect to protected or otherwise sensitive attributes in a way that is not justified?
    a. Such attributes can include gender, skin colour, social or cultural background, financial status, medical status, etc. Note that they do not need to be explicitly included in the data provided to the model, they may be inferred from other attributes.
    b. Whether such bias is likely to exist, and when it is and isn't justified is subtle to establish. An AI ethics expert should be consulted in case of doubt.
7. Does the *nature* of the proposed solution clash in any way with the fundamental values of academia or science, or fundamental rights?
    a. Example 1: a chatbot trained on copyrighted material which is likely to parrot that material, or to present a synthesis of that material without reference to the sources, can be seen as a form of automated plagiarism, clashing with a core academic value.
    b. This check refers specifically to aspects of a product that are essential to it. That is, they may not be resolved by switching to a different version of the product, since all products of this type share this aspect.
8. Does the supplier itself clash in any way with the fundamental values of academia or science, or fundamental rights?
    a. Example 1: the solution is offered by a supplier known or strongly suspected to be involved in, or aid systematic human rights violations.
    b. Example 2: the solution is offered by a supplier known or strongly suspected to form a security risk.
9. Does the product fully comply with the GDPR?
10. Does the product fully comply with the European AI act?
11. Is there a risk that students or employees will unknowingly expose sensitive or protected information?
    a. Example: A teacher has a ChatGPT license for research purposes, but decides to use the chatbot to get a "second opinion" on a student's essay, or to draft a reply to a sensitive email. The researcher doesn't realise that the chat log is stored on OpenAI servers outside the EU and may be used to train future versions of GPT.

C. Required checks

The checks are mostly framed as open questions. Each should be answered in detail. Answers may be provided by the supplier of the product, or failing that, by the university itself, based on the best available information.

If a question cannot be answered because the supplier is not able or willing to divulge the relevant information, the check fails.

Ideally, answers should be detailed and quantitative in nature. For example, if the question is “How likely is X”, the answer should contain a percentage and a detailed explanation for how this percentage was arrived at.

Whenever a property of AI product is referred to, like training data or model size, this refers to *all* AI products involved in the entire production of the final product, including APIs exposed by other companies, and direct or indirect use of models trained by other companies.

Legal and technical

1. Does the product[4] store user data, specifically *student* data? If so, how is the confidentiality, security and privacy of the data safeguarded?
    - b. For example: in a chatbot product, is the vendor allowed to
        - i. see chat transcripts,
        - ii. train on chat transcripts for future models, and
        - iii. are chat transcripts stored on servers outside the Netherlands and/or outside the EU?
2. Is the product or supplier updated to respond to security concerns, legal requirements, state-of-the-art insights into risks and ethical requirements?
3. Is the product susceptible to *vendor lock-in*? That is, if the university decides later to switch to a competing product, are aspects of the product that make this excessively difficult.
    - a. For example, can user data be easily extracted and converted as needed to use the data in a competing product?
    - b. Does the product depend on particular infrastructural requirements (e.g. use of a particular cloud provider)?
    - c. What guarantees are offered in terms of pricing, support and continued support of the product?
4. Is sufficient transparency and advance notice offered, or does the provider reserve a unilateral right to change, modify or discontinue the service or product at any time?
5. Mutual support: what kind of support and assistance is offered to test the suitability of the product/service for the purpose of the university? What kind of assistance is offered in dealing with legal claims of third parties, particularly of the source of the claims is outside the control of the university?

[4] We use the word *product* as a catch-all for any software, service or anything else that the university may purchase or deploy that may contain or build on AI technology.

User impact

6. Does the product allow users, for whatever reason, to opt-out of using the AI parts of the product?
    a. Example 1: In a word processor, extended text suggestions may be made as the user types. These may come from a large language model, which a student does not wish to use for ethical concerns. Can these suggestions be turned off?
7. Does the product make *predictions*, which are likely to be translated to actions which impact students' or employees' lives, fundamental rights or the universities core values?
    a. Example 1: A plagiarism checker which uses external features, such as a student's profile, changes in their style of writing, or speed of writing in an editor.
    b. Example 2: A system which predicts whether a student is likely to attain the required number of ECTS in the first year.
    c. If so, does the product clearly indicate what features these predictions are based on?
    d. Does the product clearly state, in a quantitative manner, the limitations of its predictions?
    e. Does the product clearly indicate how the predictions should be interpreted, and what actions are justified?
        i. For example: if poor academic performance is predicted based on the fact that a student is first-generation, then offering extra academic support is a justified action, but informally advising the student not to continue, is unjustified and will amplify biases.

Ethics and social impact

8. Is the training data, if any, ethically and lawfully sourced?
    a. Does the training data include copyrighted material, for which the authors did not give explicit permission for the material to be used in training AI models?
    b. What biases were introduced by the selection of the training data, and how are they justified? What languages, cultures and segments of society were prioritized and why?
9. Is the training process ethical?
    a. If manual annotation or feedback was used, what were the detailed labor conditions under which this was done.
        i. Were annotators paid at least minimum wage?
        ii. Were underage annotators used?
        iii. In which countries was annotation performed, and what are the labor conditions in those countries?
    b. If the model was "aligned", for example through manual feedback, how were the annotators instructed. To which values has the model been aligned, and how well do these match the values of the university?
10. How likely is the product to repeat content from the training data without substantial change as an original answer?
    a. How likely is it that this answer contains copyrighted content, which the user is not entitled to use?

11. What is the environmental impact of the building of the model (including training) and what is the environmental impact of its use?
    a. This can include the total energy use, the CO2 produced, but also other impact factors that may depend on the nature of the product.
12. Can the use of the product conceivably lead to personal harm to the user. If so, is the vendor explicit about what liability they accept. If safeguards are in place, how easy are these to circumvent?
    a. Example 1: In conversations with chatbots, some users develop an unhealthy relationship with the artificial agent, which, in rare cases has led to suicide.
    b. Example 2: Biased face recognition used in proctoring software, may lead to critical failures, causing some students excessive stress before exams because of the color of their skin.
    c. Example 3: A chatbot that allows photo-realistic image manipulation by natural language is used by a student to create explicit images of a fellow student, or otherwise harmful material.

Auditing and scrutiny

13. To what extent are the AI parts of the product open to public scrutiny?
    a. This can be made possible through an open-source license, but also through other means, such as audits by reliable independent parties.
    b. For instance, is it possible for independent parties to inspect model weights, run predictions and check for biases?
    c. Does the vendor offer a complete and detailed overview of the training data used for the product? Can the training data be checked by independent parties?
14. Does the product automate *actions*, which may substantially impact users' lives, fundamental rights or the universities core values?
    a. If so, does the product require or allow a human-in-the-loop? Is this human safeguard strictly enforced or is it likely to be reduced to a mere formality over time?
    b. For example: a system which automatically grades students' homework.
15. How was the product tested or audited?
    a. Is the result of the testing publicly available?
    b. Has the model been tested for relevant biases in the output and training data.
        i. For example, if the model relies on face recognition, does this software work equally well on users of all skin colors, genders etc.
        ii. For a model that makes predictions or takes automated actions (possibly with a human in the loop), do these correlate with protected, and irrelevant attributes such as social background. Note that such features do not need to explicitly present in the data: they can be inferred from related features.
    c. What benchmarks were used to evaluate the functioning of the product?
16. To the extent that the product/service produces synthetic content: is that content automatically marked as synthetic? What does the provider do to address concerns around disinformation, discrimination and the generation of harmful or unlawful content?
    a. Example: an AI image generator may include metadata or watermarks (visible or invisible) in the image so that the content can be easily distinguished from non-synthetic content.

17. Does the provider facilitate and/or cooperate with research into the product/service?
18. Does the provider respond negatively or disproportionately to critical discussion of, and investigation into, its service?
    a. Example: Companies like Proctorio have responded to criticism with extremely aggressive litigation against researchers and university employees . Such behaviour strongly clashes with the fundamental values of a university.

D. Optional checks

These checks do not represent hard requirements, but rather properties that align with the values and needs of the university. All else being equal, the university should look at these properties to decide between two alternative suppliers.

1. Is the product, or part of it, released under an open-source licence?
    a. If so, what is the nature of this license according to an independent review body such as the open-course initiative?
2. Is the company behind the product an active contributor to open science?
    a. Do they regularly publish (part of) their innovations in an open and reproducible manner?
    b. Do they publish training data and model details to benefit the community?
3. To what extent are *other products* offered by the vendor ethical and responsible in nature?

Bibliography

Appendix A: Fundamental values of the UvA and VU

While some limited documentation is available for what the UvA and VU consider their mission and fundamental values, a value does not need to be explicitly expressed to be fundamental to an academic institution. The following are particularly relevant in the context of AI and should be self-evidently core academic values.

Academic Honesty. This is the core value that is violated, for example, by the act of plagiarism.

Student safety. Out of all the people in the university community, the students form the most vulnerable group, and their safety should take special priority.

Student privacy. Students entrust us with sensitive, and very private information. We have a responsibility to treat this information with care, and to allow others access to it only when strictly necessary.

Societal responsibility. The university has a responsibility not to engage in conduct directly or indirectly, that causes harm to society. This means, for instance, that if a product is built on exploitative labor practices, using such a product is at odds with our values.

Freedom of investigation. Open and free investigation is probably the primary purpose of a university.

This list is non-exhaustive, and should serve only to illustrate what we mean by the core values of the university against which AI products should be evaluated.

Appendix B: Risk levels of the European AI Act

The European Artificial Intelligence Act defines the following risk levels (quoted from , citations omitted):

> **Unacceptable risk**: AI applications that fall under this category are banned. This includes AI applications that manipulate human behaviour, those that use real-time remote biometric identification (including facial recognition) in public spaces, and those used for social scoring (ranking people based on their personal characteristics, socio-economic status or behaviour).
>
> **High-risk**: the AI applications that pose significant threats to health, safety, or the fundamental rights of persons. Notably, AI systems used in health, education, recruitment, critical infrastructure management, law enforcement or justice. They are subject to quality, transparency, human oversight and safety obligations, and in some cases a Fundamental Rights Impact Assessment is require. They must be evaluated before they are placed on the market, as well as during their life cycle. The list of high-risk applications can be expanded without requiring to modify the AI Act itself.
>
> **General-purpose AI ("GPAI")**: this category was added in 2023, and includes in particular foundation models like ChatGPT. They are subject to transparency requirements. High-impact general-purpose AI systems which could pose systemic risks (notably those trained using a computation capability of more than 10^{25} FLOPS) must also undergo a thorough evaluation process.
>
> **Limited risk**: these systems are subject to transparency obligations aimed at informing users that they are interacting with an artificial intelligence system and allowing them to exercise their choices. This category includes, for example, AI applications that make it possible to generate or manipulate images, sound or videos (like deepfakes). In this category, free and open-source models whose parameters are publicly available are not regulated, with some exceptions.
>
> **Minimal risk**: this includes for example AI systems used for video games or spam filters. Most AI applications are expected to be in this category. They are not regulated, and Member States are prevented from further regulating them via maximum harmonisation. Existing national laws related to the design or use of such systems are disapplied. However, a voluntary code of conduct is suggested.