Humane AI at Utrecht University – a position paper

Tejaswini Deoskar, Pim Huijnen, Dominik Klein, Evelyn Wan *(sectorplannen kernteam Humane AI)*

Draft 1.1, May 2024

Introduction

There is a widespread consensus that artificial intelligence (AI) will play a significant role in the future of the Humanities in the Netherlands, which is why AI received a place in the most recent Sectorplannen SSH. AI has been part of the UU Humanities Faculty's self-understanding for a long time, witnessed, among other things, by the fact that it coordinates one of the oldest BA programs in AI, and plays a significant role in the corresponding MSc program (both turning 35 this year). The Sectorplannen are providing opportunities to explore and deepen the connection between AI and the Humanities in new and innovative ways in collaboration with, but also beyond those programs. Central to this connection is to "make the development of [AI] models in service to humanity as much as possible"[1] – in short: to create '*Humane* AI'.

As the core team (*kernteam*) Humane AI, we have been tasked with fleshing out the opportunities of the Sectorplannen for AI education in the Humanities. This position paper is the first result of that work. It starts with laying out a working definition of AI (Section 1). We then articulate aspects of the "humane" elements in AI, particularly with a view of the strengths of the UU humanities faculty (Section 2). We follow that with an ambitious view of the knowledge and skills that humanities students ideally leave their programs with (Section 3). The last two sections lay out general considerations regarding AI-related curriculum innovations in our faculty (Section 4), followed by specific proposals in Section 5 to operationalize those.

This position paper has been prepared after consultation meetings and input from José van Dijck, Martha Frederiks, Jan Broersen, Antal van den Bosch, and Federica Russo.[2] The policy officer of this kernteam is Jennie Weemhof and its coach Hugo Quené. It is important to note that this position paper is a 'living' document, and will evolve, as our work with the core team continues, with inputs on specific educational formats from educational and program directors, as well as with the fast-evolving changes in the field of AI itself.

Apart from these consultations our core team works in conjunction with a number of related initiatives at the Utrecht University Humanities faculty, most specifically the committee implementing digital learning goals – *digitale leerlijnen* – into all BA programs of the humanities, and the working group guiding the faculty into dealing with Generative AI. We detail relations to other initiatives at the end of this document.

1. What do we talk about when we talk about AI?

Before fleshing out what *Humane* AI entails, let us first briefly consider what we mean by artificial intelligence (AI) in general. The term stems from the 1950s, while the concept in popular imagination can be traced much further back. AI as a field of study underwent a fundamental shift in the 2010s — from what is called 'symbolic' or even 'good old-fashioned' AI (GOFAI) to a predominance of data-driven neural network-based AI — which makes defining AI itself an act of historical contextualization.

[1] SSH Raad, *Sectorplannen 2022/2023: Samen vooruit. Investeren in de basis voor wetenschappelijk onderzoek en onderwijs, versterken van de maatschappelijke veerkracht.* (Toelichting, bestedingsplannen en bijlagen, 2023), 143.

[2] More consultation meetings, particularly with experts outside the faculty, are planned for the upcoming months.

Definitions

Although scholars sometimes mean different things when they talk about AI, the literature provides some common denominators that may serve as defining traits. Take the recent definitions of AI by Montomeyer as "[t]he computational design of intelligent capacities based on information-processing techniques concerning large databases"[3], by Suleyman as "the science of teaching machines to learn humanlike capabilities"[4], or the NWO definition, "the science and engineering of making machines intelligent and collaborative. AI solutions enable machines to assist in tasks that require intelligence, such as reasoning, learning, finding information, understanding text, speech and images, listening and speaking in dialogue systems, and optimising complex systems" [5].

While they provide some important cues for what we might consider typical characteristics of present-day AI, these definitions suffer from a lack of clarity as they attempt to define (artificial) intelligence by referring to 'intelligent capacities' or 'humanlike capabilities'[6] without specifying these further. Fortunately, there are definitions of AI that free us from this Gordian knot. The two formulations below provide a clear working definition of 'intelligence', and also serve to also illustrate how fundamentally AI has changed in just a few years:

> Russel and Norvig[7] : *"We define AI as the study of agents that receive percepts from the environment and perform actions."*

> Boucher[8]: *"Systems that display intelligent behaviour by analysing their environment and taking action – with some degree of autonomy – to achieve specific goals."*

The two phases of AI

Crucially, Boucher in 2020 places emphasis on "systems" with a degree of "autonomy", a fundamental shift brought about by the introduction of deep learning by Geoffrey Hinton in 2012 and the subsequent advancement of artificial neural networks (ANN's), a radical departure from the 'symbolic' AI of before. On the other hand, this new approach could thrive because it went hand in hand with an immense increase in the availability of all kinds of—either born-digital or digitized—data: "The tremendous growth of data-driven AI is, itself, data-driven".[9]

In contrast to 'Symbolic' AI, which could be rule-based or data-driven, Neural AI crucially depends on large amounts of data that it learns from. This dependence on large quantities of data is what the

[3] Montemayor, Carlos. *The Prospect of a Humanitarian Artificial Intelligence: Agency and Value Alignment*. London New York Oxford New Delhi Syney: Bloomsbury Academic, 2023, xiii.

[4] Suleyman, Mustafa. *The Coming Wave*. First edition. New York: Crown, 2023.

[5] NWO AI Research Agenda for the Netherlands: https://www.nwo.nl/sites/nwo/files/documents/AIREA-NL%20AI%20Research%20Agenda%20for%20the%20Netherlands.pdf

See James Bridle's claim that "the most significant [quotidian] definition of intelligence is *what humans do*." (Bridle 2022). Bridle brings up the Turing test as a case in point, and argues for a fundamentally different approach towards intelligence.

[7] Russell, Stuart J., and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Third edition, Global edition. Prentice Hall Series in Artificial Intelligence. Boston et.al.: Pearson, 2016, p. viii.

[8] Philip Boucher, *Artificial intelligence: How does it work, why does it matter, and what can we do about it?* (European Parliamentary Research Service; Luxembourg, 2020), p. vi.

[9] Boucher, *Artificial intelligence*, p. 4.

latter part of Montemayor's definition ("information-processing techniques concerning large databases") refers to. Algorithms that are trained on labeled or unlabeled data are so central to current AI development that the field of machine learning (ML), once peripheral to AI, is now virtually equated with AI at large. However, historically, AI consists of a variety of different approaches, that include Machine Learning, but also Knowledge Representation, Reasoning, or Agent Based Simulations to name but a few. In this document, we build on a broad conceptualization of AI that contains Machine Learning as a central element but is by no means restricted to it. And indeed, various experts in the field see the exclusive focus of AI on (neural) Machine learning as a transient phenomenon, especially given that neural systems are notoriously difficult to make transparent or explainable.

For all their differences, from the above definitions a number of characterizations can be extracted that together seem to describe what it means to demonstrate 'intelligence':

1. An awareness of – that is, capacities to perceive and analyse – the environment
2. The ability to make decisions and take action on the basis of step 1.
3. Going through steps 1 and 2 to achieve specific goals
4. The property of having some degree of autonomy in doing so

This is, in its most abstract form, what the literature means when AI is concerned. Clearly, defining AI in this way leaves ample room for interpretation. The kernteam's outlook on AI is to purposefully use this ambiguity for reflection. We do not depart from a narrow understanding of AI with clear purposes and aims in mind. Rather, dealing with AI at our university, and specifically in the Faculty of Humanities means questioning the positionality, ambiguity and scope of the key notions involved. 'Autonomy', for example, is a highly ambiguous term that can take on a range of different meanings depending on perspective, as well as on purpose. It is not only connected to questions of technology and methodology, but also to ethical and legal considerations and, as we have seen, to cultural and historical perspectives. To a somewhat lesser extent the same goes for 'awareness'. Although this term is already elaborated in the definition above, both 'to perceive' and 'to analyze' can, obviously, mean a range of things in themselves. Montemayor lists a number of common types of perception in this regard – "inferential, perceptual, rational, linguistic and other cognitive capacities"[10] – which are all subjects of study and reflection at our university.

Whose AI?

Finally, our team's current take on AI involves the critical reflection of the term 'AI' itself. The above definitions testify of at least three things that AI can refer to: a scientific field ("the science", "the study"), the development of computational techniques ("computational design") *or* the result of that work ("systems" or, for that matter, tools, methods, products, engineering solutions, etc.). It is important to keep in mind these different dimensions, for they impact how we regard AI when it comes to factors like, again, autonomy. Most importantly step 3 ("to achieve specific goals") begs the question of *whose* goals an AI system pursues: the user's or the programmer's.

On a deeper level, our faculty's connection to AI aims to question its embeddedness – at least as a rhetorical device – in Big Tech. Dealing with AI means accounting for the fact that our current understanding of it—if not the term AI itself—is not a purely academic one. On the contrary: "The definition of intelligence which is framed, endorsed and ultimately constructed in machines is a profit-seeking, extractive one", if we follow Bridle.[11] Contextualizing AI in this way is an important part of

[10] Montemayor, *Humanitarian Artificial Intelligence*, p. xiii.
[11] Bridle, *Ways of being*, p. 17.

AI education at the UU. This is what makes it all the more crucial for a humanities faculty to define and propagate the counternarrative, of ‘Humane AI, to the technological narrative.

2. The ‘Humane’ in Humane AI

Humane AI can most simply be defined as 'human-centered' AI, i.e. AI that puts humans and their needs in the center of its focus. Humane AI requires us to take stock of the effects of AI on society and how we would build AI that is in accordance with human/ humanistic values. In the humanities, we particularly recognise the multiple subject positions of our students and staff:

- As **AI and data scientists**, who conduct scientific research on AI, or use data and digital tools for disciplinary research
- As **designers**, who do AI, i.e. who build models, code, and co-create with technology
- As **content creators,** who make and publish digital written, graphical, video or audio content for audiences
- As **users and data subjects**, who employ AI to pursue personal ends, and should be cognizant of issues like privacy and surveillance
- As **critical evaluators** of technology, who should become digitally literate, well-versed in technical know-how as well as cultural-philosophical perspectives to critically engage with the technology

Using the above subject positions, Humane AI can be seen as[12]:

- **An opportunity for innovation in research and teaching:**
  Recent developments continue to create new modes of inquiry that change our way of research. Literacy and innovation in these areas would prove fruitful for our staff and students. For instance, network analysis of Twitter (now X) data assists in understanding the emergence of social movements and their central actors. Large Language Models help in detecting patterns within bodies of text, that relate to how scientific disciplines are formed or how citation networks emerge. They can also be used for exploratory data analysis and will be increasingly adopted as assistive research tools and methods. Agent-based models allow experimentation or visualization of effects of repeated communication in large groups. ChatGPT may be used for brainstorming and thought experiments, e.g., reasoning in a historic context about how life has been at a given place and time.
- **An opportunity for creating new directions in research:**
  AI and its rapid development increasingly shapes the world around us. This raises new research questions for the humanities, concerning both AI technologies as well as their effects on society. On the technological side, advances in Large Language Models are driven by and raise new research questions within linguistics. Concerning the societal effects of AI, disciplines such as religious studies may be interested in how digital technologies influence religious beliefs and practices while philosophy and media studies pose questions on the current era of post-truth politics. Moreover, there is a need to assess AI models in their reliability and predictive accuracy, but also how and whether they may be shaped by non-epistemic values or whether they reproduce and aggravate existing biases. Such questions are typically studied within philosophy of science.

[12] The task of building Humane AI curriculum is well-connected to the focus areas Applied Data Science, Governing the Digital Society and Human-Centred AI.

- **An opportunity for raising new methodological questions[13]:**
  Computer-assisted/ generated knowledge and methods require us to rethink the philosophy and method behind knowledge production in the humanities. This connects to discipline-specific practices and how computer-assisted/ generated knowledge and data-driven methods change the onto-epistemological landscape. For example, how do humanities scholars make use of data science and modelling in studying philosophy? How do digital techniques like distant reading disrupt longstanding practices of hermeneutics and close reading in literary studies? And more broadly, how does technology (re-)mediate our experience of sociality and general belief in the world (e.g. in the study of polarisation, echo chambers, conspiracy theories, etc.)?
- **An issue of agency and creativity**[14]:
  Where do we situate human agency and creativity in our engagement with AI, and how does AI afford new forms of agency and creativity? What does it mean to live in a hybrid environment where we live with digital artefacts? From co-composing stories with ChatGPT to creative coding, and programming social robots, the humanities afford much space to experiment with technology. As content creators, we are also creative makers, writers, and creators on the various platforms we use. We expect students to enter the job market with the necessary digital skills, knowledge, as well as a mindset that would help them thrive in this hybrid world.
- **An issue with historical continuity**:
  Contextualisation of AI in its history of development is important. Not only should this history be situated as a history of technology (e.g. cybernetics, space travel projects, human computing, neural networks, AI) and a history of knowledge, but it should also be recognized as social and cultural history. As argued by Matteo Pasquinelli in *The Eye of the Master: A Social History of Artificial Intelligence* (2023), "the inner code of AI is constituted not by the imitation of biological intelligence but by the intelligence of labour and social relations." Studying AI is then not just about computation and technology itself, but also the system of social organisation that it requires, inaugurates, or supports. For instance, this can mean situating AI in a longer historical trajectory around labour organisation and systems of scientific management, where AI continues and intensifies post-Fordist factory logic. This brings the question of how humane AI has to include studying not only the algorithms, but also their process and means of production.
- **An issue of (democratic) governance[15]:**
  Governance of AI and technology is a key concern today. As argued by Shoshana Zuboff (2019), surveillance capitalism points to the widespread collection and commodification of personal data by Big Tech corporations. Companies gather data not only to predict behaviour but also to influence habits, from purchasing commodities to voting, and most technologies are built on market principles rather than on public values. In response to this, and recognising our position as data subjects and democratic citizens, how might we navigate this phenomenon and drive the development of technology towards the open futures and open societies that hold up democratic values such as transparency, accountability, and fairness? Is it possible to make technology 'humane' by design? The EU, for instance, has made regulation as part of its key tactics. This thread requires learning more about evolving AI regulation, supervisory boards, and benchmarks.

[13] There has been, for instance, discussions for a philosophy of science for the humanities course for the entire faculty. See below for a concrete proposal in relation to AI.
[14] This connects, for instance, to the SIG AI in Cultural Inquiry and Art.
[15] This connects to the Governing the Digital Society focus area, amongst others.

- **An issue of discrimination and bias**[16]:
  Algorithmic bias and data bias are ongoing issues that have to be tackled. Already addressed in books like Cathy O'Neil's *Weapons of Math Destruction* (2016), Virginia Eubanks' *Automating Inequality* (2018), Catherine D'Ignazio and Lauren F. Klein's *Data Feminism* (2020), Wendy Chun's *Discriminating Data* (2021), there is a steady stream of literature tackling the nature of big data and AI algorithms as inherently biased and problematic. In some sense, discrimination is part and parcel of the algorithmic function of AI -- whose purpose often is to distinguish and categorise data in order to discriminate (i.e., differentiate) between different entities or properties inherent in data sets. Thus, a relevant task is to discuss and distinguish situations where discrimination is justifiable from those where it is not. It is also about creating awareness of the fact that AI technologies shape and influence our understanding of the world if they prioritize and re-reproduce certain data over others. These datafied categorisations form new classifications of data subjects that may be discriminatory in nature (e.g. being involved in crimes). Bias can be created through algorithms, but also occur as a result of missing data, incomplete data, and (labelled) data that carries human prejudice (Crawford & Paglan 2001). Relevant questions, here, are how such biases can be identified and mitigated on various levels, both conceptually and technologically, see Hedden (2021).

3. What should (Humanities) students learn about AI?

Although the field of AI is a fast-moving target, fundamental skills and knowledge can be identified that students must possess, in order to be equipped as scholars, as future participants of the job market, and as members of society. The overarching learning goal that this knowledge and these skills help achieve is **empowering our students to ask the right questions** when it comes to AI. To do so, teaching AI at the Humanities faculty should focus on teaching student sufficient skills to use, understand and reflect on AI technology.

What do we mean by 'asking the right questions'? AI systems are often positioned as "black-boxes", an impression that is actively kept alive by many industry players –an example is the talk about Artificial General Intelligence, in the context of Large Language Models, as an autonomous being. In a similar vein, the question of where AI is heading is often treated as pre-determined. Such narratives encourage students and academics alike to hand over agency to outside players. We need to bring the agency in AI development back to us – and we think that the humanities do and should play a vital role in this. In the proposals that follow, we therefore keep as our guiding principle the urgent need to teach students to ask questions of the kind "Where do *we* want AI to head?" "How can *we* steer that direction?", without limiting our scope to questions in the vein "Where is AI heading?" "How do we use it? or "How do we cope with it in the Humanities?"

Data driven AI systems are often perceived as value neutral, basing their output on seemingly objective data alone. It is important to challenge this view, and to alert students of issues such as bias or discrimination through AI, and how and where human judgments and agency enter the AI building process. An important learning goal is that all students become aware of the relevant debates and, thus, acquire a critical lens for assessing AI systems and their effects. Raising the right kind of questions, in other words, is important to maintain intellectual agency—to not only become critical

[16] This connects to the KI and computational linguistics curriculum, and research conducted by the Utrecht Data School, amongst others.

users of AI, but to remind ourselves that we are also designers, data subjects, critical evaluators, et cetera (see above). To achieve this, our students should acquire both knowledge and skills. One does not go without the other: reflecting on AI only goes so far without knowing how AI works, while working on (or with) AI only makes sense in as much as one knows why, how or to what end. Our inventory of what we consider necessary learning goals pertaining to AI for Humanities students is, therefore, based on this (mutually informing) distinction between 'knowing' and 'doing'. Both branches are further specified below.

We consider all listed aspects crucial to achieve the goals of educating a new type of professional and scholar, who not only has the ability to formulate the right questions from (inter)disciplinary perspectives to AI, but also sufficient knowledge and skills to answer successfully in the light of demands of that same discipline. Such a professional scholar is well-versed in using specific systems, including having adequate understanding of their powers, inner workings, and limitations. Finding the correct balance between 'knowing' and 'doing,' as well as deciding the width and depth of teaching in the four sub-branches will differ from domain to domain.

In the following, we list both knowledge and skills we deem relevant for students.

3.1 AI knowledge

3.1.1 Reflective knowledge: AI as a field of knowledge

The goal of this sub-branch is to introduce students to AI as something that can be reflected upon. This can be done by studying AI in its political, social or cultural context, in the present day or by taking into account its historical development. A part of this may also be creating an awareness of the interconnection between literary and media narratives (science fiction) and their roles in shaping the histories and futures of technology.

3.1.2 Technical knowledge: AI as an approach

This sub-branch aims to introduce AI as an approach and method to use and reflect upon. Its aims are to enable students to explain, on a level that is highly domain-specific, how machine learning in general and large language models in particular function. It also involves teaching how to deal with data and, depending on the field, how the logic of computation works. Ideally, students learn to relate the methodologies that working with AI require to approaches from their respective fields, so that they are able to make informed research-related choices. Note that this section focusses on *knowledge* related to AI, while skills, including the ability to use AI systems are listed in section 3.2.

A) Large Language Models

Among current AI developments, Large Language Models including ChatGPT stand out as perhaps having the most wide-ranging impact on all disciplines in the Humanities and will arguably continue to do so in the near future. This includes their use in education, for composing text, in writing assignments, for information gathering, and for coding and other assignments. Consequentially, it is important for all humanities students to have at least some level of understanding of these models.

B) Machine Learning

Much of current AI developments are based on machine learning, and many models in use in society are machine-learnt models. It is important for all students to understand *at a high level* what is meant by machine learnt models, their distinction from rule-based models, the advantages and disadvantages of each, the role of data and statistics in machine-learning.

C) The use of Data

Data is a fundamental ingredient of modern AI models. In addition, the digitization of society points to the importance of all students being knowledgeable about fundamental issues related to the use of data. Examples are biases in data, data curation, representational issues in data, issues regarding privacy of data, as well as issues like access to data, and at what cost, etc.

D) Algorithmic thinking

In algorithmic thinking, the goal is not to develop the skills of coding per se, but rather to develop skills of analytical thinking via some form of "coding" and understand the logic of computer programs. This can be "creating coding" where the output is a piece of music or visual or coding for printable objects. Algorithmic thinking could also be introduced in the form of the interactive coding environment that Jupyter notebooks provide, for instance, for "soft" tasks, like querying data or other. The notebooks are an easy entry point to invite algorithmic thinking, or computational thinking, by the students.

3.2 AI skills

3.2.1 Theoretical skills

This sub-branch is aimed at fundamental skills in the age of AI.

A) Skills related to logic, argumentation, and academic writing

We argue that skills related to logical and analytical thinking, argumentation, including presenting evidence for arguments in structured and transparent ways—skills that are intrinsic to humanities education—are ever more important in times of AI. We can no longer discount the utility of ChatGPT-like tools in the writing process, both in education and the post-education careers of students. Combining the use of these tools with named skills presents new opportunities for the learning process. They enable students to assess the quality of the output of tools like ChatGPT in terms logic and argumentative structure, while using them more effectively within the boundaries of sound scholarship, for instance as a sparring partner, a time-saving tool, or to empower second-language writers to write more fluently.

B) Coding and Programming skills

Programming or coding skills are necessary to interact with AI technologies at a certain level, but also to be able to critically evaluate how they work. It is important to recognise that many humanities students currently do not have a natural inclination or interest in coding. However, recent advancements in AI, ironically, have made acquiring coding skills much more feasible. Moreover, coding is already taught in several programs in the faculty. Finally, coding is a useful, transferable, and empowering skill for humanities students on its own.

Contingent on the field of study, coding skills can aim at an in-depth skillset, but also at introducing students to coding in the first place. This can be done in the form of tutorials with existing code in the above-mentioned notebooks, by setting up interdisciplinary student projects (for example with AI or computer science students), and/or by empowering students to use the abundance of online resources that provide guidance and assistance (Stackoverflow.com, theprogramminghistorian.org, etcetera). For the more coding-heavy fields, the teaching meant here may pertain to cutting-edge AI techniques.

3.2.2 Applied skills

This sub-branch is aimed at skills education not for the sake of itself, but to teach how to functionally apply AI skills in, for example, research projects.

A) Co-creation and collaborative writing

Teaching aimed at strengthening students’ skills in using AI technology for creative and professional work, such as digital musical compositions in musicology, using generative AI to support (screen)writing in literary/ film studies, working with computer scientists to give personality to and program social robots in theatre studies, learning how to professionally carry out translations using machine-supported techniques in linguistics, etc.

B) Disciplinary connections to AI

This final section encompasses the large amount of disciplinary connections to AI, where the objective is to explore and answer discipline specific questions and relations to AI. There can be a large variety of topics covered, including inter-disciplinary and trans-disciplinary ones. Such courses may provide a particular discipline specific perspective, and be embedded in specific relevant programs.

It must be noted that the Humanities faculty contains several groups that have a direct role and expertise in developing AI techniques, i.e. “doing AI” (for example, large groups in TLC, and in F&R). These groups contribute to teaching in the bachelor KI and master AI programs in natural language processing, machine learning, ethics, philosophy of science, and logic-based reasoning models in AI. Such existing expertise in the faculty thus includes technical knowledge of the inner workings of AI models, and methods of assessment of AI within the disciplines of Linguistics and NLP and philosophical reasoning (metrics and methods). Such technical expertise can be harnessed in development of the general skill-set of “thinking about” AI, asking the right questions from the other disciplinary perspectives, or skills training.

4. Towards teaching innovations

AI is not a technological matter alone but a societal, philosophical and technological problem all at once. It is, therefore an inherently inter-disciplinary one in which humanities disciplines and skills are equal and necessary partners. Moreover, it is important to recognise that humanities students and scholars may no longer be able to work effectively in their own disciplinary fields without appropriate knowledge of AI tools and techniques. Thus, proposed course innovations have a strong interplay with interdisciplinarity. On the one hand, these considerations call for broad and interdisciplinary course proposals, for instance by designing multi-disciplinary courses and minors or by opening existing courses to further disciplines. Moreover, some of the more technical courses offer themselves for a broad, domain neutral approach.

It is also important to note that many students may choose an AI course for disciplinary motivation, rather than for an interest in AI or interdisciplinarity per se. Thus, it may help students’ appreciation of digital topics to provide technical courses not in isolation, but to interweave them with a disciplinary perspective.

Concrete proposals are presented in the next section. These are guided by a number of objectives and constraints.

Objectives:

- Maintain and strengthen expertise within the faculty. While co-operations with neighboring faculties as well as other UU institutions such as the library can be beneficial, it is imperative to ensure that a significant part of teaching AI knowledge and skills for humanities students is provided by the Humanities faculty itself.
- All stakeholders appeared, in general, welcoming to teaching innovation in relation to AI, as there is a shared feeling that current AI-based developments will have significant impacts on many disciplines and collaborations, and on the role of humanities, as well as on students' perspectives on the labor market. We see a broad consensus that we need to react to this in our teaching.
- Align our plans with the introduction and consolidation of Digital Literacy as a learning objective in the Humanities education programs (2023-2026).
- Keep in mind the need to square the current speed of AI development with the lengthy process of developing and deploying new courses, together with the fact that, once established, a course will stand for several years before its first major overhaul. Thus, teaching formats (courses, minors, etc.) need to be described at a level that is sufficiently abstract to remain relevant for 5-8 years, while still being concrete and hands on enough to be interesting. In the best case, a general course description focusses on abstract developments and templates that can then be filled in with up-to-date material by the respective lecturer(s).
- Strengthening education in AI should not increase the already high work pressure of the teaching staff.

Constraints:

- Teaching novel AI based techniques can require technical expertise that differs significantly from a classic humanities skillset and that may not be present in every disciplinary or thematic group (and also not available in the respective labor market). Thus, one aspect of teaching innovation is to find ways for making AI based skills available to the relevant programs, for instance by offering staff training or connecting potential instructors across programs and possibly departments or Schools. When successful, this can result in strengthening the education profile in smaller and vulnerable programs, and more efficient use of teaching and disciplinary expertise.
- Teaching innovation can be complex and time-consuming, especially when trying to place courses into existing programs that are already fully planned, so that new courses would need to replace existing ones. Teaching offers within running programs often strike a delicate balance between needs and desires of the various stakeholders, but also of a program's intended learning outcomes. Thus, innovation within existing programs can require careful and delicate balancing, especially when existing courses are to be replaced.

Given these considerations, the most feasible locations for innovation seems to be either (i) *within* existing courses, (ii) in the electives or (iii) in minors. These mainly pertain to the BA, although

In line with the different objectives sketched out in section 3, proposed innovations come in various shapes. Firstly, knowledge, skills and attitude deemed relevant for every humanities student may best be taught in broad courses located at School/department level or even faculty level. Special topics related to individual fields, on the other hand, better fit in a minor or in individual courses – either within a single department or shared between departments/*afdelingen*. Lastly, skills and technical education calls for small instructional units that could either go as modules into existing courses or into electives of individual programs.

Broad Courses Within our work, we identified a core understanding of AI that we deem relevant to many or all humanities BA students. These core contents call for broad courses, ideally situated in the first year. Designing such a course requires balancing two factors. On the one hand, designing a course broadly, at the School/department level or even beyond, might not only be efficient, but also facilitate finding the relevant instructions. On the other hand, situating such courses at a smaller level, targeted only at individual programs or small subsets thereof, might tie in AI instruction better with disciplinary contents and perspectives.

New Minors and elective courses. A second level of proposals are new minors. One example here is a four-course minor on (un) fairness and discrimination through algorithms. Such a minor should have a technical component and should also include a broad number of perspectives in- and outside GW, including philosophy, law or gender studies. Such minors may be composed of existing courses, new courses, or a mixture thereof.

Innovations within current individual courses. First, humane-AI related teaching innovations can and will take place within individual existing courses. Laudably, a number of existing courses already contain individual lectures or blocks relating to current AI innovation and their effects. Here, we seek to foster and deepen/broaden this development where appropriate. Innovation within existing courses is propelled through individual conversations with course and program coordinators. Though re-designing elements of existing courses may go hand in hand with extending their target group, for instance towards a later cross-listing in various programs.

AI techniques for Humanities A last level of innovation is to make short instructions in concrete AI techniques available to humanities students. The idea here is that a number of fields have indicated that it would be profitable if a subset of their students became versed in concrete techniques, e.g. related to social network analysis or data scraping. As some disciplines cannot offer these themselves, either for a lack of space in the curriculum or a lack of teaching staff, this should ideally be done through a faculty-wide offer of specialized courses. We are still deliberating about the best setting for offering these, though both the library, as well as the projected expansion of REBO's skills academy seem natural loci.

AI oriented inter-disciplinary research theses Apart from teaching-based innovation as articulated above, inter-disciplinary bachelor and master theses, jointly supervised by staff with differing expertise offer unique opportunities for combining humanities disciplinary perspectives with AI.

5. Concrete Proposals

Taking into account the considerations laid out in the previous two sections, we here present a number of initial ideas towards course innovation. These proceed in the order introduced in section 4, i.e. from broad courses to minors, in-course innovation and new technical modules. Throughout the description, we link to the learning goals introduced in section 3 by means of pictograms.

Broad Courses, or Course Modules:

Description: This proposal refers to basic knowledge (see 3.1.2) that students must possess regarding AI. The training could be organised in different ways, either as a complete course, or in terms of short modules, possibly online, that are taught by relevant staff, and are put together in one course, or distributed amongst several courses. The latter has the advantage that short modules could be integrated within existing disciplinary courses, helping to understand and practice AI skills within a disciplinary perspective. The latter format also invites for a "flipped classroom" approach, where students watch modules online on their own, but discussion sessions are organised in seminars/werkcolleges in relevant courses.

The course/course modules proposed here are high level and discipline neutral, and may satisfy portions of the Digital Leerlijn requirements, especially concerning Machine Learning / Large Language Models.

A. Large Language Models

- 1-2 hour module on how an LLM works
- 1-2 hour module on use of LLMs (including responsible use)

B. Machine Learning for the Humanities

- 2 or 4 hour module: distinguish rule-based AI and machine learnt AI; advantages and disadvantages; some concepts related to the fundamental data-driven and statistical nature of machine learning models, some insights into data selection and labelling, as well as the process of AI building and deployment.
- This knowledge set can also be more in-depth by having a half-course, modelled along the lines of "Methods and Statistics" courses that are shared between programs.

C: Humane Data in AI

- 2 or 4 hour module on Bias and Representation in ML based AI models (LLMs or others..)

New Minors and elective courses.

Example I: Minor: AI, Data and Fairness

Description: This minor sheds light on societal implications of AI technologies, specifically in relation to fairness and discrimination. It aims to study this problem from a multi-angular perspective, combining a thorough formal foundation with philosophical perspective and legal perspectives on discrimination and AI. Moreover, the minor also introduces practical approaches towards reducing or mitigating bias and discrimination through AI.

Courses are ideally a mixture of formats that are developed for this minor and other existing courses. Instructors should mainly be from GW, though occasional outside expertise (REBO) may be needed.

Proposed course structure:

C1: Formal basics (understanding of ML, possibly also other tools, understanding of data labelling and its effects, explainable AI, Formal measures of fairness, remedies of unfairness) Could coincide with one of the courses developed in the previous section on broad courses/modules

Teachers: ideally GW internal. Possibly from LLI (Logic Language and Information) group in TLC?

C2: Philosophy of AI decision making

(What is bias / what is discrimination, when and why are they wrong. Privacy. data autonomy. Epistemic rights towards a decision system. Explanation, Predictive accuracy, AI in public policy….)

Teachers: From FenR, likely combination between theoretical philosophy and ethics.

C3: [tentative:] Origins and effects of AI discrimination (Continuity with other forms of discrimination, digital colonialsm. Effects of discrimination)

Teachers: MCW?

C4: Legal perspectives and building fair AI (how to show discrimination of a system. Admissible uses of data….) Also: Influencing the AI building process

(two halves of the course, likely involving lectures from Rebo)

Example II: Keuzeruimte course: Post-humanities in the Age of AI

This course focuses on the philosophical implications of AI, specifically focusing on the ontology, epistemology, methodology, and ethics that surround the theme of AI in the humanities. Taking on a posthuman perspective, the course positions AI as something that extends but also challenges human exceptionalism/ anthropocentrism within humanistic thinking, with attention to perspectives from feminist science and critical race theory. Some potential themes for the course:

- What is intelligence?: Exploring the relationship between various interpretations of intelligence in the philosophical discourse around what artificial intelligence means, e.g. in relation to non-human intelligences (plants/ animals etc.).
- What makes us human?: Exploring related concepts on cyborg, posthumanism, transhumanism, and study how AI challenges our biological bodies and extends our perception.
- AI and Creativity: Analysing the relationship between artificial intelligence and creativity in fields such as literature, art, and music etc.
- Are we living in a simulation?: Exploring the nature of being and reality and the ontological implications of incorporating AI (or more specifically, extended reality technologies like augmented reality, virtual reality) in our daily lives.
- Ethics and AI: Investigating the ethical considerations and moral dilemmas arising from the themes above.

Example III: Keuzeruimte course: Narratives of AI: From Text to Computation

This course focuses on the role of language and discourse in shaping AI, that combines insights from literary studies, media studies, history, and computer science. There is an intrinsic link between text, coding, language and computation. This link can be observed on the level of computer programming and design (e.g. code, hypertext, natural language processing, etc.), on the level of human-computer-interfaces (e.g. search engines; voice prompts; text-to-image/ video generators), and on the level of imagination and design (e.g. AI imaginaries and narratives through science fiction). The course surveys and builds a historical narrative around the role of text and textuality in computation and AI development. Potential authors: Lev Manovich; Katherine N. Hayles; Wendy Chun; Mark Hansen; Stephen Cave, Kanta Dihal and Sarah Dillon.

Innovations within current individual courses (or merge / split structures between courses)

Example 1: Cross-listing of courses between RMA (Linguistics) program and the MSc AI program

🌐

Strictly speaking this is not a proposal, as it is already in the process of being implemented in 2024-25. It is mentioned here as an example of a *low-hanging fruit* – a proposal that had low overhead and can make a small sub-group of courses in a Humanities program make inter-disciplinary connections to AI, and in the process increase the enrollment of students to the courses.

Description: The RMA (Linguistics) program is in the process of being overhauled in the past year for independent reasons. As part of this overhaul, several elective courses in the program were changed. Two of these courses could be cross listed as secondary electives in the AI Master program (Beta faculty):

Example 1a. Topics in AI and Language

This course is part of the elective package of the RMA Linguistics. The course name was changed, and the course designed to become a course that will cover specific topics in Linguistics that related to AI. The course is research based and topics can change per year – the specific topics can largely be determined by the expertise of the inter-disciplinary staff-member teaching the course, which provides flexibility in staffing. If limited to the RMA Linguistics, about 10-12 students are expected to enroll; however, by cross-listing with AI, the expected number of students could be doubled.

Constraints: This proposal needs staff members/teachers with a strong inter-disciplinary connection to AI research and development. It was easily feasible in part because several linguistics staff already teach in the AI program. However, it could serve as a model for courses in MA Philosophy programs in F&R, where a similar cross-listing could be proposed and implemented.

Technical adjustments: AI program courses are 7.5 ECTS by default, and RMA courses 5 ECTS. A separate and extra 2.5 ECTS module will be provided to AI students who enroll (else those students will fall short of their program ECTS requirements by 2.5 if they enroll in the course as a secondary elective).

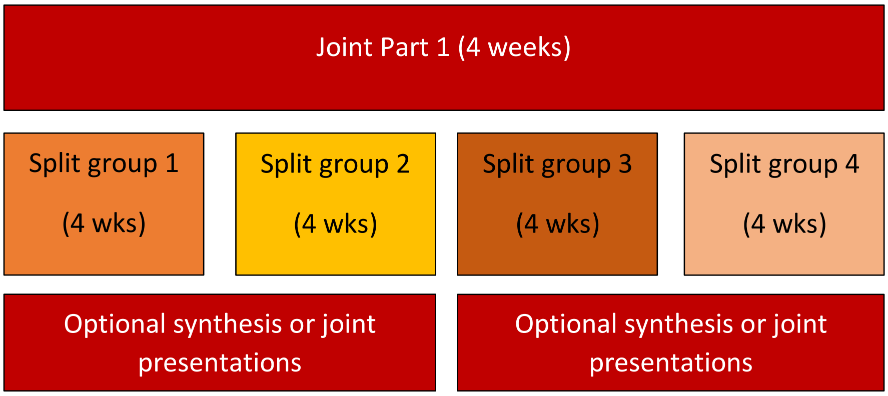
Example 1b. Experimentation in Psychology, Linguistics, and AI

This course is one of the primary electives in the AI MSc program, run by TLC. It covers experimental methods in AI, e.g. with human subjects in the domain of language or other cognitive areas, along with statistics. This course was in turn cross-listed as a secondary elective for the RMA Linguistics, and also serves as the *statistical methods* course for RMA students interested in a thesis in an experimental area (for whom it is obligatory). By doing so, all Linguistics RMA students obtain an additional secondary elective in a new area, and also a statistical methods course, without the costs of a separate course. Linguistics RMA students can avail of an “early exit” option for 5 ECTS, but are free to take the course for 7.5 ECTs.

Example 2: Innovation of "Tools and Methods" courses in BA programs (possibly with a merge/split structure)

Background: In a meeting of TLC kern-team members and TLC BA/MA co-ordinators (13/02/2024), there was expressed the concern that some "Digital Methods/Digital Tools" courses in BA programs are subject to the dichotomy that on the one hand, a (large) subset of students find these topics either difficult or irrelevant to their disciplinary interests, while on that other, some students are interested in these topics and do not get sufficient depth from the courses. Moreover, it was observed that (i) there needs to be more focus on fundamental concepts rather than training in using software tools (ii) the specific tools or software used could be more specialized (e.g. per program, or based on a track in a program).

Description: A selected group of courses that fall under the category of digital tools or methods course could be modified. The structure of joint courses (as proposed by Talen en Culturen KernTeam) can be applied.



Joint Part 1 (4 weeks)

Split group 1 (4 wks)

Split group 2 (4 wks)

Split group 3 (4 wks)

Split group 4 (4 wks)

Optional synthesis or joint presentations

Optional synthesis or joint presentations

Example 2a:

A joint Methods course already exists between two programs in TLC

- Taalwetenschaap and Taal & Communicatie (TW2V19002, *Methods and Statistics 1*)

Potential new program groupings for other joint Methods courses following this model could be:

- Methods course for the "languages" programs (Spanish, Italian…)
- Inter-departmental Methods courses could be identified as well. (Onderwijsdirecteur's input would be valuable here to evaluate feasibility, and to identify potential groups.)

Example 2b:

A similar structure could apply to MA courses. Examples of existing joint courses are *Distant Reading: Digital Tools and Textual Analysis* (TLRMV16202) (Comparative Literature Studies, and Dutch Literature and Culture)

AI techniques for Humanities:

The field and topic of Humane AI provides an excellent playing ground to experiment with new teaching formats. First of all, teaching Humane AI is interdisciplinary almost by definition. Whether AI (as machine learning, LLM's, or any other shape or form) is introduced in teaching as an analytical tool, as a sphere of innovation, or as a topic of reflection, it will always require the connection between technical and methodological skills and Humanities domain knowledge. The easiest way to make this connection is to have students from different disciplinary backgrounds work together.

Example 1: Hackathon lab

A challenge-based educational format in which Humanities students are paired with Computer Science students to work on a particular task within a given amount of time. The task can either be a concrete question and/or a specific dataset (such as the Panama Papers) that the students have to jointly work on. The task should be formulated so that the application of some form of AI is vital, as is the specific skill set that the Humanities student brings. This is work that is typically done for multiple hours at a time within the time span of a week, after which every group presents their work and a jury may decide on who has presented the most elegant or efficient solution to the task at hand. It can also be combined with tutorials, guest lectures, etc. Moreover, this teaching format would ideally be structured as a CEL course. The work of the student pairs would gain real urgency if the question and/or dataset are coming from an external partner (and their solutions might have a real benefit).

Example II: Supervised machine learning lab

An educational format that revolves around the introduction of supervised ML. The epistemological and methodological approaches of many Humanity programs are not readily translatable to the principle of supervised ML. This is one of the reasons why few Humanities students 'officially' learn about these principles. At the same time, they may benefit from knowing about supervised ML tasks all the same—for their future careers, if not for their own research papers/thesis. Particularly when large amounts of data are involved, supervised ML may add to their existing methodological toolbox. An entire course on supervised ML might be too niche for most Humanities students, but a methodological lab session of 4-5 weeks might suffice to train them in the techniques and attitudes needed to apply ML. It could revolve around a specific dataset that the students (from different backgrounds) jointly label, train, and use, so as to learn the entire supervised ML pipeline. The lab could start with a specific question that the students jointly work on. This will lower the great deal of work involved in the labeling/training phase, and give the students the opportunity to not only learn the principles but also see the results of their work (sentiments in a social media dataset, objects in a dataset of photographs or paintings, handwriting recognition in a historical dataset of handwritten texts). Depending on the dataset, it could also be added with new labels in each iteration of the lab, to create our very own enriched Humanities dataset that enables all kinds of predictive research. This lab teaches students how to apply supervised ML, but also provides many opportunities for methodological and epistemological reflection when it, for example, comes to questions of labeling correctly or asking binary questions (yes-no, good-bad, positive-negative). Again, this lab format can be presented as a CEL module if the dataset is provided by an external partner. In that scenario, they would offer the real-life data and the urgency, we the expertise, the work, and the critical reflection.

Example III Raspberry lab

The Raspberry Pi is an extremely affordable minicomputer that is developed especially for experimental and educational settings. The latest version (RP5) has the power to do AI tasks like speech or image recognition. We could offer a lab session in which students from different backgrounds work together on a specific task in which they have to build the RP (add peripherals etc.) and program it, apply it to a real-world setting, and analyse the data. This could be the analysis of noise volume in the university library, the number of persons visiting a specific building, or any of the tasks under 4b. There are dedicated websites full of RP projects, which makes the challenge of programming the RP highly feasible for non tech-savvy students. Besides getting introduced into the hands-on work, this particular lab creates ample possibilities for reflection. This involves the same kinds of questions as module 4b stimulates, but physically placing hard-to-see mini-computers over the campus and collecting visual or audio information also raises many ethical and legal issues that we aim to train our students in. As a consequence, this module also provides opportunities to talk these through beforehand, invite experts, and jointly find workable solutions for data-collection

Inter-disciplinary RMA/BA research theses

Joint supervision of bachelor or master theses, by specific interested staff members, and AI experts (either outside the faculty, or within the faculty), is a good method to foster innovation that spans both research and education. Such joint supervisor needs some structural flexibility, since the number of students interested in such joint or interdisciplinary projects are hard to predict, and many staff may refuse such projects if their teaching time is already full. (For instance, *dcu*s for such supervision can come out of the research budget of individual staff members in the year of supervision but counted in the next year – such measure may encourage more interesting interdisciplinary interactions).

6. Related initiatives

Relations between the kernteam and current efforts to implement ***digitale leerlijnen*** move in various directions. Given that the digital learning goals required a longer administrative coordination process, they do not yet reflect various recent developments such as the prominence of LLM over the past 1-2 years. On the one hand, courses proposed in this document may help various programs meet their digital learning goals as defined in their specific OERs, and on the other hand, some of the desiderata for all humanities students identified in this document can be seen as an addition to the leerlijnen defined and may also be included in future versions thereof.

A further relevant entry is the **Working group on Generative AI.** This working group is tasked with creating information about and defining policies for the use of Generative AI in teaching and examinations – both by instructors and students. This working group has put forward introductory material on Generative AI and its use and pitfalls (cf. here) and has also formulated sample regulations and policies that can be put into OERs as well as course manuals. As of 2024/25, BA programs in GW are required to have some sessions on Generative AI and its use in their first year. Here, the workgroup coordinates and assists with the creation of this teaching material.

A main difference to the kernteam AI is that the workgroup on Generative AI has a much narrower focus, adressing generative AI only, and is more directed to think about a recently available tool and its use for education by students and teachers. Relevant debates here are how this tool could change teaching and examinations, how it could be used productively and whether and where its use calls for regulation. The kernteam, on the other hand, is less concerned with tools used *for* teaching/learning and more with i) learning about AI and ii) using AI as tool for research tasks (rather than as a tool within education). Of course, overlaps between the kernteam's task and the working group on generative AI exist, especially concerning the new teaching modules. Here we are in ample communication to coordinate these efforts.

Relatedly, there is the **Taskforce AI and data science.** While having its main focus on research activities, the taskforce also has a subcommittee dedicated to teaching. The latter is currently working towards an overview of AI and data science related teaching in- and outside GW. Especially knowledge and contacts outside our faculty may come in handy for the kernteam in pursueing inter-faculty proposals.

Another relevant entity is the **Center for Digital Humanities (CDH) (**http://cdh.uu.nl/) intending to promote DH methods in the humanities. While the kernteam is exclusively tasked with teaching innovation, one strong focus of CDH is on staff development. In light of our proposed courses of AI literacy a number of staff members have uttered a wish to have related educational offers for them. Such courses may be provided by CDH, which might also help teachers to develop their skills in teaching AI related content.

List of Sources

Boucher, Philip. 2020. *Artificial intelligence: How does it work, why does it matter, and what can we do about it?* European Parliamentary Research Service; Luxembourg.

Bridle, James. 2022. *Ways of Being: Beyond Human Intelligence.* London: Allen Lane.

Chun, Wendy Hui Kyong. 2021. *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition.* Cambridge, Massachusetts: MIT Press.

Crawford, Kate and Trevor Paglen. 2021. Excavating AI: The politics of images in machine learning trainings sets. *AI & Society,* 1-12.

D’Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism.* Cambridge, Massachusetts: MIT Press.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* New York, NY: St. Martin’s Press.

Hedden, Brian. 2021. On statistical criteria of algorithmic fairness. *Philosophy & Public Affairs.* 49 (2), 209-231

Montemayor, Carlos. 2023. *The Prospect of a Humanitarian Artificial Intelligence: Agency and Value Alignment.* London et.al.: Bloomsbury Academic.

Pasquinelli, Matteo. 2023. *The Eye of the Master: A Social History of Artificial Intelligence.* London: Verso Books.

Russell, Stuart J., and Peter Norvig. 2016. *Artificial Intelligence: A Modern Approach.* Third edition, Boston et.al.: Pearson.

Suleyman, Mustafa. 2023. *The Coming Wave.* New York: Crown.

SSH Raad. 2023. *Sectorplannen 2022/2023: Samen vooruit. Investeren in de basis voor wetenschappelijk onderzoek en onderwijs, versterken van de maatschappelijke veerkracht.*

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* London: Profile Books.