Humanistic AI: Towards a new field of interdisciplinary expertise and research

Mats Fridlund[1,2], David Alfter[1,3], Daniel Brodén[1,2], Ashely Green[1,4], Aram Karimi[1,3], and Cecilia Lindhé[1,2]

[1] *Gothenburg Research Infrastructure in Digital Humanities (GRIDH), University of Gothenburg, Renströmsgatan 6, Gothenburg, 412 55, Sweden*
[2] *Department of Literature, History of Ideas and Religion, University of Gothenburg, Renströmsgatan 6, Gothenburg, 412 55, Sweden*
[3] *Department of Philosophy, Linguistics, and Theory of Science, University of Gothenburg, Renströmsgatan 6, Gothenburg, 412 55, Sweden*
[4] *Department of Historical Studies, University of Gothenburg, Renströmsgatan 6, Gothenburg, 412 55, Sweden*

Abstract

The Gothenburg Research Infrastructure in Digital Humanities (GRIDH) have participated in projects within various humanities fields that utilise as well as develop research tools and infrastructural resources that incorporate applications of 'artificial intelligence' (AI). These applications can include natural language processing, machine learning, computer vision, large language models, image recognition algorithms, classification, clustering, and deep learning. This paper advances the term 'humanistic AI' to describe an emergent form of interdisciplinary practice that uses and develops AI-based research applications to answer humanities research questions together with its entangled humanistic reflection. We coin this term to make implicit and visible the epistemological and material particularities of its practice and the new forms of knowledge its affordances make possible. The paper presents GRIDH projects within 'humanistic AI' together with its developed AI resources and applications.

Keywords

Research infrastructure, interdisciplinarity, critical digital humanities, artificial intelligence

1. Introduction

The recent surge in interest in the academic impact of ChatGPT and other applications of 'artificial intelligence' or 'AI', mainly overlook how humanities researchers have long been using and developing AI. Most prominently within humanities disciplines such as corpus linguistics and language technology, but also in digital humanities and traditional disciplines such as archaeology, comparative literature, and history. In the latter cases, this use is often less prominent in that it tends to be embedded in language technology tools and algorithms, such as topic modelling and word embeddings. With the expanding use of digital methods together with a rising critique within the humanities against the unreflective (dare we say, naive) use of biased and potentially dangerous AI applications, we propose a conceptualisation of 'Humanistic AI', to allow such uses to be discussed in a more structured and nuanced manner.

This paper is a first tentative attempt to develop Humanistic AI as a concept describing an emerging field of interdisciplinary research and expertise. We explain what we mean by Humanistic AI and lay out its main practical and conceptual dimensions, followed by describing the involvement of the Gothenburg Research Infrastructure in Digital Humanities (GRIDH, formerly Centre for Digital Humanities) in projects utilising, developing or interrogating AI. We conclude by discussing the concept's usefulness in wider communications.

Huminfra Conference 2024, Gothenburg, 10-11 January 2024.

mats.fridlund@gu.se (M. Fridlund); david.alfter@gu.se (D. Alfter); daniel.broden@gu.se (D.Brodén); ashely.green@gu.se (A. Green); aram.karimi@gu.se (A. Karimi); cecila.lindhe@gu.se (C. Lindhé)

2. Conceptualising Humanistic AI

What do we mean by 'Humanistic AI' and what are its different elements and dimensions? In short, the term refers to activities within humanities research and cultural heritage, that use, develop or study AI tools and applications. Below we briefly describe what we mean by 'humanistic' and 'AI', followed by a discussion of the emergence of the term and its different meanings within various AI fields and how we position ourselves in using the concept. Finally, we describe the three main elements of Humanistic AI as we understand the concept and its practice within the humanities.

2.1. Meanings of 'Humanistic' and 'AI'

Within AI, 'humanistic' can be used to designate 'humane' or 'human-like' functionalities and behaviours as well as to describe aspects related to humanities disciplines or knowledge domains. For us, the term 'humanistic' in 'Humanistic AI', refers to the latter sense, to designate an activity or research-directed project within cultural heritage and the humanities. Notably, among historians, there is a consensus that 'the humanities' consists of a complex of socially and historically constructed academic disciplines and practices perceived as distinct and yet under continuous renegotiation [1]. For instance, in Sweden many humanistic disciplines move across different university faculties, and before the 1960s the faculty of humanities represented both the humanities and the social sciences [2].

Secondly, the meaning of 'AI' stands for artificial intelligence whose meaning is somewhat more problematic due to its increasingly widening meanings within its different academic and public contexts. To clarify and critique the various uses and abuses of the AI term a number of alternative and differentiating terms have been introduced such as 'augmented intelligence', 'intelligence augmentation', 'automated approaches', 'autonomous systems' and 'intelligent systems. The "classic" AI textbook describes it as a field "concerned with not just understanding but also building intelligent entities—machines that can compute how to act effectively and safely in a wide variety of novel situations", encompassing "logic, probability, and continuous mathematics; perception, reasoning, learning, and action; fairness, trust, social good, and safety; and applications that range from microelectronic devices to robotic planetary explorers to online services with billions of users." [3]. When thinking about AI, many people without advanced technical expertise would imagine autonomous robots such as seen in TV shows, fearsome "creatures" that surpass human intelligence. While this certainly can be referred to as AI, our approach is more grounded in using it to refer to a *field that develops and studies intelligent machines, as well as to those algorithms and machines themselves*. In particular this refer to machine and software applications from the AI subfields of Expert Systems, Machine Learning (ML), Natural Language Processing (NLP), Speech Recognition, Computer Vision, Robotics, and Genetic Algorithms, which in themselves includes specific applications such as clustering, deep learning, image segmentation, text classification and topic modelling.

Despite these many alternatives to AI, we have, aware of the diversity and range of meanings discussed above, chosen to retain the use of the AI term as it encompasses many significant applications used within the humanistic domain. Furthermore, this is also due to the emerging and increasing contemporary use of AI connected to various purportedly 'humanistic' purposes which is something we see ourselves as well positioned to engage with.

2.2. Emergence of Humanistic AI

The last two decades have seen the term 'Humanistic AI' being used in different ways within the AI domain that partly overlaps our use. For instance, in 2003 the term 'humanistic AI' was used to describe one main trajectory within the design of intelligent machines, trying to emulate human cognitive capabilities, rather than mimicking the anatomical functioning of the human brain [4]. More recently, the term 'Human-Centered AI' (HAI) has also been used for similar AI activities and processes. Not rarely, such efforts are shaped by a rationale implying that HAI is not just efficient but also more fair, compatible, and 'humane', in augmenting rather than 'replacing' human decision-making. Furthermore, there are a range of AI development activities similarly drawing on

HSS perspectives, described in terms such as 'Responsible AI', 'Ethical AI', 'Fair AI', 'Trustworthy AI' and, at times, 'Humanistic AI' [5].

One example is the Media Lab at Swedish KTH Royal Institute of Technology that see the humanities provide "a critical perspective" as well as "a source of innovation in AI" and engages in interdisciplinary research combining "advanced engineering with philosophy, art, aesthetics and other disciplines from the humanities" to "develop a strong humanistic stance with respect to AI to avoid a situation of technological intelligence overrunning humanism" [6]. Another example is University of Bologna's Humanistic AI unit that similarly describes Humanistic AI as a "novel branch" reframing the study of "the embodied human mind and social and cultural contexts, as well as their reciprocal relations", applying AI techniques to humanities that "range from the classification, exploration, management, and preservation of cultural heritage, archives, or demo-ethno-anthropological materials" [7]. Our conceptualisation aligns with these efforts insofar as it concerns the application of AI to humanities rather than the design of HAI as well as developing digital resources in an interdisciplinary context augmented by the reflective and critical faculties of humanities scholars. It should also be noted that a similar but opposite trend is recently surging in popularity within AI, namely 'human-in-the-loop architectures', i.e. machine learning architectures where human knowledge is provided before or during the training phase in order to overcome the limits of modern AI [8].

2.3. Elements of Humanistic AI

AI is involved in humanistic research through three main areas of practice: humanistic researchers using various existing tools incorporating AI applications; the use of AI tools and knowledge to develop custom-made resources for humanities researchers; and humanists' interrogating AI through analysis and reflection on the impact of the AI tools' embedded positions ('biases') and affordances. All of this, either by individual humanistic researchers or as interdisciplinary collaborations. On one level, Humanistic AI can be understood in relation to other concepts, such as, Critical Digital Humanities [9] and Critical Code Studies [10], in that it concerns critical, interdisciplinary oriented reflection on sophisticated digital methods, software, tools, etc., as well as the socio-cultural production of knowledge in "digitalised" society. On another level, our notion of Humanistic AI is more of a tentative framework constructed around the somewhat contested concepts of 'DH' and 'AI' rather than an agreement on approaches and objects. Thus, we delineate a wide range of activities united by a heterogeneous aim to explore AI within the domains of the humanities.

2.3.1. Using AI

Applications of AI used in DH projects involve a range of diverse techniques and methods which includes vector representation for text, contextual search, data annotation, clustering, image classification, and recognition. Specific examples of applications implemented by GRIDH include advanced word embeddings (such as Word2Vec or FastText) to create vector representations of textual content allowing for semantic similarity analysis, topic modelling, and contextual understanding; word embeddings in combination with domain-specific ontologies to enhance the semantic understanding; topic modelling techniques, such as Dynamic Topic Modeling (DTM) or Term Frequency - Inverse Document Frequency (TF-IDF), to capture evolving themes and topics in historical text; semantic search to clarify meaning of queries and documents and to improve search recall precision; and image colour clustering based on similarity of embeddings and calculation of 'nearest neighbours'.

2.3.2. Developing AI

DH research often involves complex issues not easily solvable by simply applying existing AI applications designed for general purposes and tasks. Thus, in contrast to simply *using* AI, DH projects may also set out to *develop* AI applications for their specific purposes. This can be done in different ways, such as training classifiers, fine-tuning existing or training new transformer models from scratch based on specific text or image corpora. Such GRIDH applications include computer vision and deep learning techniques for automatic image annotation, object detection and segmentation for image

labelling. However, developing more general AI applications requires a deeper understanding of the underlying principles (and implications), and large amounts of training data, a constraint often hard to satisfy in the humanities (e.g. the documents to be analysed are in extinct languages, or the artefacts under scrutiny no longer exists).

2.3.3. Interrogating AI

The last element entails applying humanistic research-based reflection and critique to interrogate the implications of the AI tools and methodologies used. The core of humanities scholarship concerns applying hermeneutics, (source) criticism and reflection concerning methods, tools and data used in producing humanistic knowledge, including potential societal impacts. Also, some humanities disciplines specialise in reflecting on the development, use and impact of digital technologies, such as digital humanities, information science, media studies, practical philosophy, and science and technology studies (STS). Such AI reflexiveness at times comes as explicit interdisciplinary studies including humanities scholars, exemplified by an archeological study stating that "the outcome of any AI approach and its interpretation will differ," enabling "us to reflect on the potential limitations of our digital technology to avoid taking its results as answers to our research questions about human beliefs, ideologies, and creativity." [11] However, such interdisciplinary interrogation also comes tacitly in project conversations with humanist scholars probing the interpretative limits and affordances of the data generated by AI tools. This often entails making obtuse AI algorithms fathomable, as humanist researchers, to quote a digital humanist, "can never afford to treat algorithms as black boxes that generate mysterious authority" and to use them, "we have to crack them open and find out how they work." [12]. Or at least try, as working out their inner workings can require elaborate analysis of models and outputs, at times involving separate projects analysing model performance and training that incorporates human-in-the-loop components or active learning.

3. Humanistic AI projects at GRIDH

The Humanistic AI projects at GRIDH focus on one of the three elements above or combine different forms of them and separates into two categories: text-based and multimodal AI projects. They somewhat overlap as text-based projects often include analyses or manipulation of images in the form of digital image files of text documents that are OCRed or not.

3.1. Text-based AI projects

3.1.1. The *Nordisk Familjebok* tool

A research infrastructure project initiated by GRIDH and developed together with Data as Impact Lab at the University of Borås, that created a open digital resource (*nordiskfamiljebok.dh.gu.se*) of the two first editions (published 1876–99 and 1904–26 respectively) of the encyclopedia *Nordisk Familjebok (NF)*, a standard reference work for studying 19th and 20th century Swedish society. The scholarly use of NF was augmented by implementation of advanced 'likeness' search functionalities using a Word2vec-model based on the KB-BERT large language model of the KBLab of the National Library of Sweden, making it an AI use as well as AI development project.

3.1.2. The New Order of Criticism project

The project's comprehensive approach aims to provide rich insights into how readers perceive and engage with books in the Swedish language, facilitating the development of user-centric applications like personalised book recommendations. To accomplish this large, pre-trained Swedish language models like BERT and ELMo are leveraged to conduct sentiment analysis and classification of newspaper book reviews. The project employs fine-tuning techniques to tailor these models to specific tasks, ensuring high performance. It goes beyond simple sentiment polarity analysis by implementing

aspect-based sentiment analysis, entity recognition, and emotion detection. The research in the project also includes advanced visualisation methods for presenting findings and addressing ethical considerations in AI.

3.2. Multimodal AI projects

In several GRIDH projects AI is used to study, analyse and manipulate non-text data, most often digital images and at times digital audio and video, and sometimes also associated with geospatial data. Thus, when we talk about multimodal projects, we mean it as a description of our non-text focused AI projects using one or several other data modes than text. Often these projects also involve analysis of text data. In some cases, these projects are 'genuinely' multimodal, where data of multiple types are combined for analysis to add additional context and at times also include analysis of multiple data types in parallel.

3.2.1. Projects using image clustering

GRIDH are using ML algorithms and developing interactive visualisations for image clustering of several types of image content. In the literary 'lab' developed for Litteraturbanken (LB), GRIDH use machine learning algorithms to cluster images of illustrations, initials, graphics ornaments, and sheet music extracted from the LB's repository of 19th century works using object detection. The aim is to enhance the visualisation of literary reuse and similarity, as well as provide future researchers with easy data access. The *Ivar Aroseniusarkivet* (Ivar Arosenius Archive) project uses methods and concepts developed by Douglas Duhaime and the Yale DHLab [13], as well as the Nasjonalmuseet in Norway [14] to visualise the archive's images of the artwork. The TSNE projections of RGB images and a gallery of each image's nearest neighbours are displayed in an interactive frontend. GRIDH aims to use additional clustering algorithms and improved interactive visualisation to increase the archive's accessibility to the public and researchers.

3.2.2. Projects using Augmented Reality

In the project 'Rock Art in Three Dimensions', an application was developed using two different Augmented Reality (AR) technologies; markerless image detection trained on natural features where a device tracks its position through image recognition of natural features, and plane tracking which recognises horizontal and vertical surfaces using the technique Visual Inertial Odometry (VIO). Markerless image detection made possible to detect and add contextual information to rock carvings without any physical additions to the rock art site. Thus, the interpretations could be served in a digital form while the physical environment was kept untouched, thereby aligning itself with the values of conservation [15].

3.2.3. Projects using automatic speech recognition

One frequently mentioned AI application is automatic speech recognition (ASR) which serves as the foundation of the project 'Terrorism in Swedish politics (SweTerror)' that studies parliamentary speech on terrorism 1968–2018. The speech technology trains and adapts deep neural networks to better cope with speaker biases, especially gender, and train, compare and inspect models trained on speech to detect historical changes in parliamentary speech. Also, the project's text analysis relies on state-of-the-art LT methods using AI, such as word pictures, topic modelling and word vectors. [16].

4. Conclusions

In this paper, we have tentatively suggested Humanistic AI as an apt concept for discussing an emergent field in the intersection of the application of AI tools and the interests that fall within the domain of critical digital humanities and adjacent fields of the humanities. By addressing the core elements of what could be considered Humanistic AI and concretising it by presenting some projects involving

GRIDH, we have sought to demystify the notion of applying AI within the humanities. In a way, we have tried to show that humanists are already active within AI, sometimes without realising the depth, degree, or character of their involvement. In writing this paper, we have debated the ambiguous and partly contested concept of AI, and in this come to terms with the extent to which the notion of Humanistic AI can be useful in describing the work at GRIDH concerning the development of resources and infrastructure with elements of AI. This usefulness becomes apparent, not least when communicating what we "do" to external partners and fellow humanist researchers as well as other academic and non-academic stakeholders, interested in the applications of AI to the humanities.

Acknowledgements

We are very grateful for the support from and collaboration with our research and development partners and staff at CDH and GRIDH who made the projects discussed above possible, in particular Johan Eklund, Christian Horn, Johan Ling, Arild Matsson, Gustaf Nelhans, Rich Potter, and Victor Wåhlstrand Skärström.

References

[1] R. Bon. *A new history of the Humanities*, Oxford University Press, Oxford, 2013.
[2] A. Ekström & H. Östh Gustafsson (Eds.). *The humanities and the modern politics of knowledge*, Amsterdam University Press, Amsterdam, 2022.
[3] Russell, Stuart J., and Peter Norvig. *Artificial intelligence: a modern approach*. Harlow, 2021. 4th ed, 7, 19.
[4] Krishnakumar, Kalmanje. "Intelligent Systems for Aerospace Engineering: An Overview." *Von Karman Institute Lecture Series on Intelligent Systems for Aeronautics* (2002).
[5] Saheb, Tahereh, Sudha Jamthe, and Tayebeh Saheb. "Developing a conceptual framework for identifying the ethical repercussions of artificial intelligence: A mixed method analysis." *Journal of AI, Robotics & Workplace Automation* 1.4 (2022): 371-398.
[6] https://www.kth.se/hct/mid/research/media-lab/about-1.929121
[7] https://centri.unibo.it/alma-ai/en/scientific-units/humanistic-ai
[8] Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). "A survey of human-in-the-loop for machine learning". *Future Generation Computer Systems, 135*, 364-381.
[9] D. Berry and A. Fagerjord. Digital humanities: Knowledge and critique in a digital age, Polity, London, 2017.
[10] Mark C. Marino, *Critical Code studies*, MIT press, Cambridge, Massachusetts, 2020.
[11] Horn, C., Ivarsson, O., Lindhé, C., Potter, R., Green, A., & Ling, J. (2022). "Artificial intelligence, 3D documentation, and rock art: Approaching and reflecting on the automation of identification and classification of rock art images." *Journal of Archaeological Method and Theory*, 29, 188–213.
[12] T. Underwood. "Theorising research practices that we forgot to theorize twenty years ago", *Representations*, 127:1 (2014): 64–72.
[13] https://douglasduhaime.com/posts/identifying-similar-images-with-tensorflow.html.
[14] https://www.nasjonalmuseet.no/en/about-the-national-museum/collection-management---behind-the-scenes/digital-collection-management/project-principal-components/
[15] Westin, J., Råmark, A. & Horn, C. (2023). "Augmenting the Stone: Rock Art and Augmented Reality in a Nordic Climate". Conservation and Management of Archaeological Sites.
[16] J. Edlund, D. Brodén, M. Fridlund, C. Lindhé, L-J. Olsson, M. Ängsal and P. Öhberg, "A multimodal digital humanities study of terrorism in Swedish politics: An interdisciplinary mixed methods project on the configuration of terrorism in parliamentary debates, legislation, and policy networks 1968–2018", in: K Arai (ed.): *Intelligent Systems and Applications: Proceedings of the Intelligent Systems Conference (IntelliSys) 2021*, 2, Springer, Cham, 2022, pp. 435–449.