

From critical technical practice to reflexive data science

Simon David Hirsbrunner 

Freie Universität Berlin, Germany

Michael Tebbe

Freie Universität Berlin, Germany

Claudia Müller-Birn 

Freie Universität Berlin, Germany

Convergence: The International
Journal of Research into
New Media Technologies
2024, Vol. 30(1) 190–215
© The Author(s) 2022



Article reuse guidelines:

sagepub.com/journals-permissions
DOI: 10.1177/13548565221132243
journals.sagepub.com/home/con



Abstract

In this article, we reconsider elements of Agre's critical technical practice approach (Agre, 1997) for critical technical practice approach for reflexive artificial intelligence (AI) research and explore ways and expansions to make it productive for an operationalization in contemporary data science. Drawing on Jörg Niewöhner's co-laboration approach, we show how frictions within interdisciplinary work can be made productive for reflection. We then show how software development environments can be repurposed to infrastructure reflexivities and to make co-laborative engagement with AI-related technology possible and productive. We document our own co-laborative engagement with machine learning and highlight three exemplary critical technical practices that emerged out of the co-laboration: negotiating comparabilities, shifting contextual attention and challenging similarity and difference. We finally wrap up the conceptual and empirical elements and propose *Reflexive Data Science (RDS)* as a methodology for co-laborative engagement and infrastructured reflexivities in contemporary AI-related research. We come back to Agre's ways of operationalizing reflexivity and introduce the building blocks of RDS: (1) organizing encounters of social contestation, (2) infrastructuring a network of anchoring devices enabling reflection, (3) negotiating timely matters of concern and (4) designing for reflection. With our research, we aim at contributing to the methodological underpinnings of epistemological and social reflection in contemporary AI research.

Corresponding author:

Simon David Hirsbrunner, Research Group Human-Centered Computing (HCC), Institute of Computer Science Freie Universität Berlin. Königin-Luise-Straße 24-26, 14195 Berlin.

Email: simon.hirsbrunner@fu-berlin.de

Keywords

AI ethics, artificial intelligence, critical algorithm studies, critical data studies, data science, digital media research, human-centered computing, human-centered design, human-computer interaction, interdisciplinarity, machine learning, practice research, reflection, science & technology studies

Introduction

In 1997, Philip Agre regretted that ‘AI has never had much of a reflexive critical practice, any more than any other technical field’ (1997: 132). Agre experienced this lack of self-awareness and -critique firsthand, while he was doing his PhD and worked further at the prestigious AI Laboratory of the Massachusetts Institute of Technology (MIT) in Cambridge. He documented his discomfort with the intellectual practices in AI research in several of his later works, including the article *Toward a Critical Technical Practice: Lessons Learned Trying to Reform AI* (1997) and his book *Computation and Human Experience* (1997a). Much of his criticism revolved around the ways the AI community dealt with its own technical language and metaphors. Agre argued that, whenever problems arise, AI researchers tend to interpret these problems within its existing discourse and by using the available metaphors of this discourse. The attribution of problems to available metaphors will then motivate specific technical solutions. In contrast, the metaphors themselves will not come into question and ‘will hover in the background, acting as arbiters of plausibility without taking any of the blame’ (Agre 1997a, 46). According to Agre, the result is that ‘fundamental commitments will go unquestioned and fundamental difficulties will go unaddressed’ (Agre 1997a: 46).

Agre argued, against this background, that critical self-awareness is crucial to resurface the work of metaphors in technical work and help to diagnose, articulate and evaluate previously implicit commitments (Agre 1997a: 46). He proposed a distinct technique to interrogate the role of metaphors by replacing one set of metaphors against another. In Agre’s work, this technique is particularly exemplified by problematizing mentalist metaphors inspired by the cognitive and psychological sciences, which were then dominant in the field of symbolic AI.¹ Agre suggested, instead, to explore the productiveness of interactionist metaphors such as ‘conversation’, ‘involvement’, ‘participation’, ‘feedback’, ‘cooperation’ and ‘improvisation’, that shift the attention from machinery and cognition to dynamics and activity (Agre 1997a: 53). Agre also made clear, however, that replacing one metaphor system against another should not be seen as an end in itself, but as a means of hermeneutics (Agre 1997: 154). Reconfiguring sets of metaphors was, for him, a method to generate friction and gather new perspectives within multiple reflexive cycles of a research process. This assessment vividly illustrates the delicate position Agre had taken between computer science on the one hand and social sciences and humanities on the other hand. While being a computer scientist by training, Agre was heavily influenced by ideas of philosophers and social scientists such as Martin Heidegger, Michel Foucault, Lucy Suchman and Harold Garfinkel to name just a few. In his work, he also drew on authors who had already provided a comprehensive critique of dominant AI discourses. His most notable resources, in this regard, were Hubert Dreyfus (Dreyfus, 1972, 1982; Dreyfus and Dreyfus, 1988), as well as Terry Winograd and Fernando Flores (Winograd and Flores, 1986). This intellectual influence was so important that Agre later referred to himself as a computer scientist who turned into ‘a social scientist concerned with the social and political aspects of networking and computing’ (Agre, 1997: 132). Despite this shift in his academic positioning, Agre also retained the view that critique of AI research may not solely take the form of discursive critique. He argued, in contrast, that the interrogation of metaphors should be operationalized within and through technical work. Such alternating sessions of computer programming

with reflections on the discursive embedding of technical elements and processes was what Agre then further conceptualized as *Critical Technical Practice (CTP)*, ‘in which rigorous reflection upon technical ideas and practices becomes an integral part of day-to-day technical work itself’ (Agre, 1997a: 3). In his book (1997a), he underpinned this argument by comprehensive experimental studies carried out with/on the computer program Pengi, which Agre had written together with a colleague at MIT (Agre and Chapman, 1987).

Agre’s CTP approach had been received in several areas of computer science, most notably in the fields of HCI and critical design (Baumer, 2015; Dourish et al., 2004; Sengers et al., 2005). The field of symbolic AI research, in contrast, had generally shown little interest in equipping technical work with a critical perspective informed by the humanities and social sciences. This situation has changed dramatically in the last few years when a new generation of powerful AI technologies, namely data-driven machine learning (ML) systems, became mainstreamed in various scientific, economic and social spheres. The capabilities and massive socio-technical diffusion of these technologies brought up new concerns about harmful consequences of such algorithmic techniques to citizens, communities and society as a whole (O’Neil, 2016). This situation arguably facilitated a reflexive turn in AI-related research in recent years. A striving community of AI researchers and data scientists emerged that made issues such as fairness, accountability, transparency and ethics (FATE) in ML an established field of study discussed at dedicated conferences² and slowly integrated into computer and data science curricula at universities (Saltz et al., 2019). Some authors have also voiced criticisms towards FATE’s operationalization of AI critique and asked whether the FATE community is diverse enough to identify and address important shortcomings of ML systems and their impact on spheres of the social.³ It has been argued, in this context, that critical AI research requires more holistic approaches than those currently discussed in the FATE community. It has been asserted, for instance, that many initiatives suffer from ‘inmates running the asylum’ (Miller et al., 2017), referring to computer scientists’ own interpretation of problems in AI without integrating perspectives from other scientific disciplines or social spheres.

Against this background, it has been stated (Bates et al., 2020) that interdisciplinary settings are well suited to take into consideration more diverse perspectives of AI critique. It has, therefore, become increasingly common that computer scientists share research spaces with social scientists and vice versa to open the horizon of critical engagement and explore new perspectives on AI-related epistemological and socio-technical problems linked to AI. Such initiatives have particularly been probed and documented by researchers from the field of Science and Technology Studies (STS). Most of these researchers gathered their data as ethnographers embedded in AI research and development laboratories in academic and non-academic settings. Jaton (2017, 2021), for example, describes *ground-truthing practices* in ML drawing from his embedding in an image-processing laboratory. Henriksen and Bechmann (2020) investigated the ways predictive algorithms were made *doable* in healthcare by building on a close observation of work in a Scandinavian AI-developing company. Both Grosman and Reigeluth’s account of algorithmic normativities (2019) and Neyland’s characterization of the everyday life of an algorithm (2019) are, in turn, informed by the researchers’ embedding in scientific surveillance technology development projects. These researchers all discuss problematic elements, mechanisms and assumptions within current ML practice that would not have been brought to the fore by computer and data scientists themselves. They expand, accordingly, the space of critique in AI research and point to problems that should be tackled by ML practitioners. Jaton (2021) even goes beyond that and proposes genuine instruments (‘techno-moral graphs’) to report and compare the quantity of moral operations present in a ML project. It seems disputable, however, whether these lines of thought effectively reach AI research discourse and design practice. On the one hand, ethnographers often remain observers of the ML

design practice happening in situ, only articulating their critique much later in form of a publication and within (the bubble of) their own scientific community. The design process will, by then, long have been completed and the ethnographers may have lost contact to the ML practitioners.

On the other hand, the refined written accounts tend to remain silent about the frictions, struggles and constant re-positionings that occurred during the inter- and transdisciplinary encounters. Such resolution of conceptual inconsistencies of views is an elementary technique in scientific publication writing, as STS literature has vividly demonstrated (Latour, 1987). Resolving inconsistencies and organizing alignment is not always possible, however, particularly within interdisciplinary practice. The only way to do so in short time span of most research projects is to elect a leading discipline and to downgrade the other participating disciplines or research fields to service providers (Barry et al., 2008: 29) or observing bystanders (e.g. in case of much ethnographic research). This limits the potential of interdisciplinary research to innovate and provoke substantial change in the participating disciplines and their scientific practices. In the context of AI research, such seamless and frictionless interdisciplinary engagement risks coming across as well-intentioned but toothless reflections on social and ethical issues, without possibilities to inform and transform concrete technical practices and infrastructures.

This brings the discussion back to Philip Agre's CTP, which heavily built on the idea that struggles between technical practice and reflexive critique should not be resolved, but made productive for scientific analysis, transformation and innovation. As 25 years have passed since Agre's conceptualizations of CTP, it is worth assessing, however, to what extent the approach in its original version is still apt for guiding critical engagement with contemporary AI. The technical practices and discursive architectures of symbolic AI addressed by Agre bear little resemblance to the socio-technical settings of today's ML systems and data science practices.

In this article, we reconsider elements of Agre's CTP approach and explore ways and expansions to make it productive for an operationalization in contemporary data science. First, we propose how to reorganize Agre's CTP as a social practice. Drawing on Jörg Niewöhner's co-laboration approach (Niewöhner, 2015), we show how frictions within interdisciplinary work can be made productive for reflection. Second, we suggest that software development environments can be tailored to make co-laborative engagement with ML technology possible and productive. We specifically discuss the Jupyter Notebook (Kluyver et al., 2016) as a means for enabling social contestation and reflection in contemporary ML systems design. Third, we further suggest ways to infrastructure reflexivities in contemporary data science practice by making socio-technical constellations and their transformations visible and accessible for discursive scrutinization. This includes the recording of online team meetings, saving different stages of the technology, producing memos of identified interdisciplinary misalignments in the co-laborating team and abstracting situational mappings (Clarke, 2003; Marres, 2020) of shifts in the co-laboration and reflect on paths taken and not taken. Fourth, we document three exemplary critical technical practices that emerged out of the co-laboration: negotiating comparabilities, shifting contextual attention and challenging similarity and difference. Fifth, we propose *Reflexive Data Science (RDS)* as a methodology for co-laborative engagement and infrastructure reflexivities in contemporary AI-related research. To do so, we come back to Agre's ways of operationalizing reflexivity and introduce the building blocks of RDS: (1) organizing encounters of social contestation, (2) infrastructuring a network of anchoring devices enabling reflection, (3) negotiating timely matters of concern and (4) designing for reflection.

With our study and article, we aim at contributing to the theoretical underpinnings of methods for epistemological and social reflection in contemporary AI research. The contribution is, accordingly, an interdisciplinary one that may be productive for researchers in fields such as Science and

Technology Studies, Critical Data and Algorithm Studies, Human-Computer Interaction and Data Science.

Organizing CTP as a social practice

An essential feature of CTP is the constant struggle between technical routines and their problematization through instances of reflection. Agre's own way of operationalizing this struggle is a split identity of the individual researcher with "one foot planted in the craft work of design and the other foot planted in the reflexive work of critique" (Agre, 1997: 155). This setting requires a strong ability to switch between these identities, but also a certain skill to keep the identities apart from each other and prevent complete reconciliation and fusion. If not, CTP may be stripped of its most productive elements, which are frictions and conflicts between the two identities. It seems questionable, however, to what extent this artificial separation of identities in a researchers' mind may be sustained throughout a research process. While this arrangement may have worked for Philip Agre, it seems questionable whether CTP in its original design can easily be translated to the research realities and abilities of other researchers. Agre's capacity in *doing* CTP is rooted in Agre's deep knowledge of the AI research field's foundations and in his ability to question and overturn his own assumptions, coupled with comprehensive technical skills acquired at the MIT AI Laboratory. This artful valuation and entanglement of skills constitutes, on the one hand, a vivid example to showcase what CTP does and could possibly do. On the other hand, the strong reliance on Agre's capacities may also represent a potential limitation of the CTP approach. One may ask to what extent it is possible to achieve Agre's advanced degree of reflexivity without embodying this unique combination of personal capabilities and traits – or, to put it more bluntly, whether it is possible to conduct CTP without being Philip Agre.

An alternative way to organize frictions between technical and critical capabilities would be to organize them through a setting of interdisciplinary engagement. Such a reconfiguration of CTP would adhere to Agre's idea to make frictions between technical and critical practices epistemologically productive, but it would organize these struggles as a social activity (researchers participating in an interdisciplinary research process), rather than a mental activity (Agre's 'split identity' in a researcher's mind). It should not be taken for granted, however, that interdisciplinary engagement itself will lead to advanced reflection of theoretical positionings and disciplinary commitments. Such capacities for reflection must, in contrast, be carefully woven into the mode of interdisciplinary cooperation. According to Barry and colleagues (2008), interdisciplinary research can include different modes of conduct: (1) an integrative-synthesis mode, where several disciplines increasingly merge in view of eventually forming a new field and discipline (such as biochemistry and astrophysics). (2) In the subordination-service mode, by contrast, disciplines are organized according to a hierarchical division of labor entailing one lead discipline that defines the research questions and research design, while additional disciplines provide complementing services in this context. Arrangements within the field of digital media studies have often worked in this manner, in which scientific programmers (computer scientists) provide technical tools for the project-leading media scholars. (3) Finally, the mode of collaboration can be agonistic–antagonistic, where 'interdisciplinarity springs from a self-conscious dialogue with, criticism of or opposition to the intellectual, ethical or political limits of established disciplines or the status of academic research in general' (Barry et al., 2008: 29). Drawing on this idea of agonistic–antagonistic engagement, anthropologist Jörg Niewöhner proposed *co-laboration* as a practice that sets in motion an experimental process inducing reflexivity in all participants. For him, co-laboration has to be understood as a non-teleological, experimental practice aiming at producing a different kind of critical

reflexivity than other interdisciplinary constellations (Niewöhner, 2015: 220). Niewöhner highlights that the ultimate objective of co-laboration as an epistemic style is rather disciplinary than interdisciplinary: ‘It is aimed at producing reflexivity that challenges the dominant thought styles within a discipline to move that discipline along’ (Niewöhner, 2015: 235). Co-laboration can, thus, enable productive interdisciplinary work, while safeguarding principal disciplinary identities, priorities and practices of its co-laborators.

We believe that the agonistic–antagonistic mode of interdisciplinary engagement and its operationalization through co-laboration constitutes a promising way to organize CTP as a social process. In the following, we explain how we organized an interdisciplinary engagement in a research project at the Human-Centered Computing (HCC) research group at Freie Universität Berlin. We have been fortunate to work in institutional environments that enable experimental work on AI and comprehensive interdisciplinary discourse. All co-laborating researchers were part of the same interdisciplinary research group working in the fields of HCI and computer-supported cooperative work. The declared mission of the group is to embrace a critical practice in the design of socially responsible technologies that enable a collaboration between humans and computers. The group currently focuses, among other subjects, on explainable AI and the design of technologies for reflection. Both members of the core team were also funded by research programs that aim explicitly at interdisciplinary cooperation between the computer sciences and the social sciences. As Jörg Niewöhner has rightly pointed out, being sympathetic to the approach of co-laboration and finding money to set up a co-laboratory are two different things (Niewöhner, 2015: 237). In our case, we must acknowledge that the boundary conditions of our research have been particularly favorable to co-laborative activities enabling an agonistic–antagonistic dialog between technical and critical positionings. We struggled, however, at first to make the reflexive potential of the interdisciplinary constellation productive in practice. Overtime, however, we increasingly invented ways to structure the oscillation between technical and critical work and to make reflexivities accountable for scientific purposes.

Creating objects of social contestation and reflection for co-laborative CTP

We encountered several problems while attempting to operationalize CTP as a co-laborative practice in our HCC research group. First of all, co-laborative practice requires an understanding and clarification of one’s own and the others’ positionings (Niewöhner, 2016: 3). We started with a relatively crude understanding of the scientific positioning of the other in our co-laborating group of researchers, basically as ‘computer scientist’ versus ‘humanities scholar’. We then discovered and discussed alternative and probably more realistic positionings such as ‘data scientist’, ‘ML expert’, ‘HCI scholar’ on the one hand and ‘media scholar’, ‘social scientist’, ‘STS scholar’ and ‘interpretivist researcher’ on the other hand. The mutual discovery of more granular positionings helped us to understand the others’ fields of expertise and theoretical commitments. Second, we initially struggled with finding common reference points to coordinate both our technical and critical work. Co-laboration is fragile and marked by constant uncertainty and frequent renegotiation of the conditions of mutual engagements. As a cooperation with limited consensus, it requires certain common commonalities to keep co-laborators together. As Susan Star and James Griesemer (1989) write, such commonalities can be characterized as *boundary objects*, which may take various forms, including repositories, ideal types, coincident boundaries and standardized forms. HCI and computer-supported cooperation work scholars have argued that digital cooperative work typically involves more fluid objects that do not only enable cooperation beyond boundaries of social worlds

but also the negotiation of these boundaries (Bertelsen and Bødker, 2002; Lee, 2007). In the context of learning, we have also argued elsewhere (Hirsbrunner, 2021) that digital objects such as dynamic data visualizations can serve as anchoring devices enabling the discursive negotiation of complex matters of concern. Our own practice, then, showed that such objects can also be made productive for learning and reflection in the context of AI research. A number of techniques can be used to make the ML software development process as transparent as possible and produce artifacts as references of co-laborative reflection. In the following, we introduce Jupyter Notebooks as a tool for co-laborative reflection. Jupyter Notebooks (Kluyver et al., 2016) are software environments that can be accessed, edited and executed via a visual interface in a web browser (see Figure 1). The software projects accessed through the notebooks can either be run on a local machine or, as in our case, on a cloud server. Jupyter Notebooks organize source code in so-called ‘cells’, which are executable and produce outputs such as text, images (e.g. diagrams) and interactive visualizations. It is also possible to provide accompanying explanations to each cell in text form. Working with Jupyter Notebooks creates many advantages for software developers. It enables effective prototyping, versioning control and collaborative programming in settings of a distributed workforce and cloud environments.

In our co-collaboration, the cells of the Jupyter Notebook acted as reference points for discursive negotiations of the software development and data analysis. Referring to individual cells, we were able to discuss our (often diverging) views on the meaning and role of software components with reference to these pieces of code, visualization or accompanying description. The reference to the cells as fluid but concrete and discrete artifacts stabilized the often fragile communication and coordination processes.



```

+ Code + Text
[ ]
import numpy as np
import os
import time
import gdown
import tensorflow as tf
import tensorflow_hub as hub
import tensorflow_text

from tqdm import tqdm
import re

from google.colab import drive

from numba import jit, njit, vectorize, cuda
from sklearn.metrics import pairwise_distances

from sklearn.preprocessing import normalize
from sklearn.decomposition import PCA, TruncatedSVD
import matplotlib.pyplot as plt

from sklearn_extra.cluster import KMedoids
import seaborn as sns
#from sklearn.cluster import AgglomerativeClustering, DBSCAN, KMeans, OPTICS
import umap as umap

import pickle

```

Figure 1. View of a Jupyter Notebook running our ML pipeline. Source: our own visualization.

Infrastructuring co-laborative CTP

While the Jupyter Notebook helped us to organize in situ reflection, making the frictions, struggles and re-positionings in the project team productive in the longer term required the invention of a whole apparatus for documentation and analysis. We combined multiple techniques to trace modifications and adjustments in the Notebooks and make it available for ex-post reflection: on the one hand, different stages of code, data outputs and their visual presentations (diagrammatic visualizations) were saved by both automatic and manual means. On the other hand, we also recorded our digital project team meetings via video conferencing software and wrote text memos with particular attentiveness to situations of breakdown and repair. The heterogenous data produced by means of these techniques were analyzed in reflexive sessions by the project team. The sessions typically lasted two hours and were held both online and, whenever possible, in person. We watched excerpts of the recordings of our project meetings together and discussed major diverging points of view that occurred. We then abstracted the identified (re-)positionings into situational mappings inspired by Adele Clarke's *Situational Analysis* (Clarke, 2003; Clarke et al., 2015) and Noortje Marres' socio-technical expansion through *Situational Analytics (SA)* (Marres, 2020). In the end, we produced (auto-)ethnographic vignettes of major controversies identified through the analysis of the situational mappings. We decided to translate these auto-ethnographic vignettes of our own activities (see next section) into the third-person as a strategy of estrangement, defamiliarization and abstraction from our own perspectives (Hirschauer, 2013), enabling more effective ex-post reflection.⁴ We want to argue that the entanglement of these techniques can be characterized as an attempt to *infrastructure reflexivities* into contemporary ML practice and research.

As (Star and Ruhleder, 1996) have shown, infrastructure should be seen as a relational property and may mean different things to different people at different points in time. Others have built on this idea and showed how infrastructuring (as the practice of building infrastructures) happens on a daily basis in the field of information and communication technology development (Karasti, 2014; Star and Bowker, 2006). If one adheres to the idea of infrastructure as a relational property, infrastructures and their components may also be repurposed and subjected to new uses, as we have done in the context of our AI-related research project. We have repurposed Jupyter Notebook from an instrument for transparency, coordination and collaboration into a tool for social contestation and co-laborative reflection. In order to do so, we had to combine our work with the notebooks with other digital practices and technologies, such as video conferencing (reflective sessions) and collaborative online mapping (situational analytics). While requiring adjustments on a daily basis, this dynamic infrastructure increased our capacities for critical self-awareness in the context of our co-laborative engagement. Put differently, it enabled us to infrastructure social reflexivities in our data science practice.

Empirical case study

This section documents key moments and constellations of our AI-related research project and highlights three exemplary reflexive practices that emerged out of our co-laborative process: negotiating comparabilities, shifting contextual attention and challenging (dis-)similarities. The practices strongly relate to each other and should not be seen as discrete entities, but rather as different perspectives and weightings of a broader practice in reflexive AI-related research.

We experimented with AI-related technology within an interdisciplinary collaboration at an interdisciplinary research group. The core team consisted of a social scientist (hereafter SOSCI) and a computer scientist (hereafter COSCI). A student was also involved more closely through her

bachelor thesis, which included the implementation of software modules and a comparison of ML modules. The collaboration and software design were further informed, debated and evaluated in discussions with the leader (hereafter PROSCI) and colleagues at the research group working in the field of Human-Computer Interaction and human-centered ML. The collaboration started with a casual discussion at the research group seminar and was loosely guided by several entangled research interests and objectives. The first entry point and objective for the collaboration was to evaluate the aptness of approaches in natural language processing to support interpretivist analysis of online debates on social media platforms. SOSCI had shared his research about the discursive negotiation and appropriation of visual climate risk scenarios on YouTube (Hirsbrunner, 2021a). Building on a qualitative analysis of post-video discussions on YouTube, his study showed how users play with rhetoric tactics, such as irony and sarcasm, to cope with the perceived uncertainty regarding the credibility and plausibility of the online information presented. After the discussions in the research group, SOSCI asked COSCI whether one could imagine helping qualitative researchers in their coding process by using ML. SOSCI knew that COSCI had been a collaborator in a research project probing the use of ML within a collaborative ideation process in software development, where he collected first experiences with the technologies used in this collaboration by applying them to crowd-generated innovative product ideas (Müller-Birn and Tebbe, 2020). COSCI, then, was in the early stages of his PhD project in computer science and on the lookout for potentially interesting use cases in which he could apply his technical knowledge of ML to real-world problems. Supported by PROSCI and the institutional setting of the relevant research group, the two researchers started to work together, which led to a year-long and ongoing mutual engagement. In the beginning, COSCI and SOSCI started with a fairly concrete research question and problem to be solved: how could ML approaches support interpretivist scholars in making sense of online discussions around climate change debated on social media platforms? The researchers decided to closely adhere to the existing research practices of SOSCI and work with the same data that had already been collected in the existing qualitative study. While SOSCI had chosen to limit his qualitative coding of comments to 600 items, the idea of the collaborative ML experiment was to provide ways to explore, order, classify and analyze the whole dataset extracted from YouTube, entailing a total of 26,000 comments under the video *How Earth Would Look If All The Ice Melted* (Science Insider, 2015) uploaded by Science Insider on 18 September 2015. The comments were collected with the YouTube Data Tools (Rieder, 2015), a visual interface and tool tapping YouTube's Application Programming Interface. The original dataset included a variety of variables such as the text of the comment, author name, number of likes, number of replies and publishing data. The researchers decided to focus particularly on the text of the comments and to develop ways to classify and group comments of similar structure and meaning. The comments represented in the dataset included short expressions such as 'wow' or 'no', but also longer statements debating the likeliness of the depicted sea-level rise scenarios at hand.

SOSCI and COSCI experimented with multiple ML techniques (sentiment analysis, topic modeling (TM), text embeddings). The first approach considered thoroughly was TM by means of Latent Dirichlet allocation (LDA) (Blei et al., 2003). The latter is an established method that aims at extracting the latent topics present in a corpus of documents. This allows one to describe the corpus in terms of its topics (themes) and their distribution within the documents. Topic modeling was chosen due to its characteristics that support qualities such as transparency, explainability and interpretability. The parameters leading to modeling outputs, for example, are well understood in principle and the outputs can be interpreted relatively easily by human beings. The classifications generated through the TM technique, however, were rather disappointing. The results did not provide any more information than a word frequency list, which is a common feature in many

existing tools for (qualitative) data analysis (e.g., MAXQDA, Atlas.ti). A scrutiny of the results showed, however, that TM was simply not fit for the data at hand. While the technique has frequently been used in the comparative analysis of long texts, such as scientific publications (Griffiths and Steyvers, 2004) or interview transcripts (Baumer et al., 2017), it was simply not apt for the characterization of shorter texts, such as the YouTube comments in question.

Experiments with pretrained neural language models, which were conducted in parallel to TM, turned out to be more productive. The model chosen was the Universal Sentence Encoder (USE) (Cer et al., 2018), a language model pertaining to the family of transformers (Vaswani et al., 2017). USE was then operationalized with other methods that aim at calculating groupings ('clusters') of semantically similar text. The same dataset of 26,000 YouTube comments (see above) was processed through this method. USE produces a mapping of datapoints (i.e., the comments) in a high-dimensional space. Each comment is represented by a distinctive, long list of numbers (i.e., vectors), which encode information about its semantic content. The promise, then, is that similar lists of numbers (i.e. vectors with similar direction) represent similar semantic meanings and that these (dis-)similarities can support text interpretation. Since the mapping in a high-dimensional space is not apt for human cognition and sense-making, however, additional steps are needed to enable an interpretability of results. On the one hand, the clustering algorithm K-Medoids (Park and Jun, 2009) was applied to group ('cluster') the datapoints based on their similarity score in the embedding space. Without such clustering, the analyst would be confronted with an endless list of comments ordered only by their similarity score. Clustering, by contrast, enables the identification of patterns in the text corpus. Furthermore, a dimensionality reduction is performed to enable the representation of the output in a diagrammatic form. The Uniform Manifold Approximation and Projection (UMAP) for dimension reduction (McInnes et al., 2020) was used here, considering its compatibility with transformer-generated text embeddings (Coenen et al., 2019). The challenge of this step is to reduce the dimensionality of the vectors representing samples to two (or

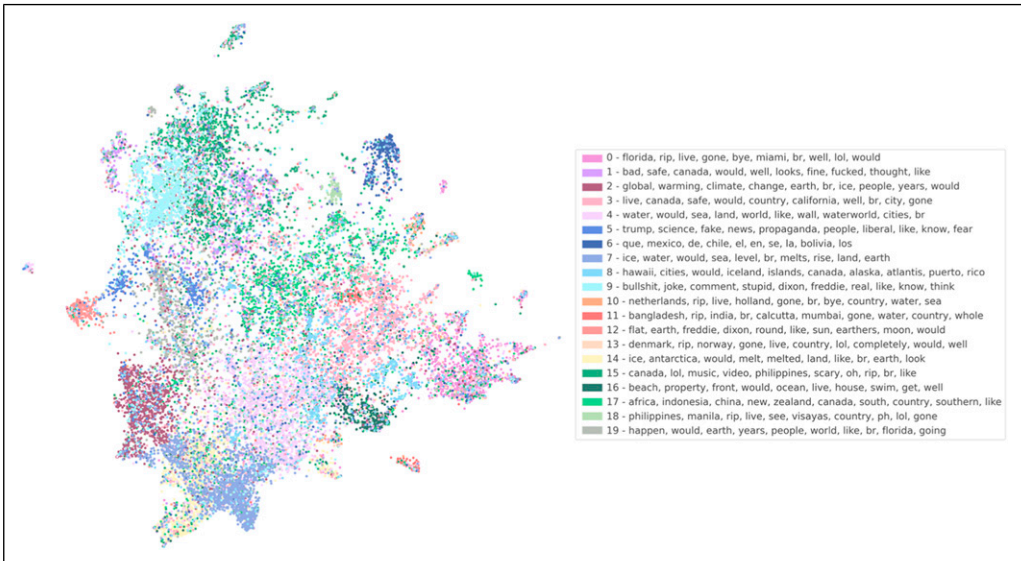


Figure 2. Automatically labeled clusters of comments associated with the YouTube video 'How Earth Would Look If All The Ice Melted'. Source: our own visualization.

three) dimensions while preserving the distances between the datapoints as accurately as possible. The dimensionality reduction made it possible to plot the results as a list in a table similar to a two-dimensional canvas, thereby enabling data exploration and analysis by a human analyst.

The ML pipeline established enabled SOSCI to engage in multiple practices of data annotation and analysis of YouTube comments. Both tabular and diagrammatic views of the data were provided to conduct these tasks. Samples of the data in one cluster, for example, could be shown as a list, starting with the comments the model deemed most similar. The clusters could also be compared by looking at a scatter plot that maps datapoints on a two-dimensional canvas (see Figure 2). Clusters of similar comments are represented through different colors. Moreover, the most frequent words of different clusters are designated, which allows a better understanding of the clusters' data and grouping characteristics.

In addition, one can also manually assign labels to the clusters. The label 'Propaganda', for example, was used for a cluster of comments that included terms like 'Trump,' 'science,' 'fake,' 'news,' 'propaganda,' 'people,' 'liberal' and 'fear.'

The pipeline was then experimentally tested and used together with students in a seminar at university. Thirty-five master students enrolled in computer science, data science, computational sciences, business informatics and interaction design joined the course focusing on critical social media analysis using mixed methods. The seminar started with the introduction of Philip Agre's CTP approach in order to highlight the instructors' understanding of 'critical analysis'. The students and instructors then used the ML pipeline established to address multiple questions, such as how to evaluate the representativeness of social media data based on thematic relevance, using qualitative methods for social media data analysis (grounded theory, discourse analysis) and identifying the blind spots of ML algorithms used in online hate speech detection.

To make this operationalization accountable to the reader, the researchers provide as a situational mapping of our unfolding co-laboration (Figure 3). The journey started on the left with a dataset of YouTube comments and diverse expectations with regard to AI-related technology. The co-

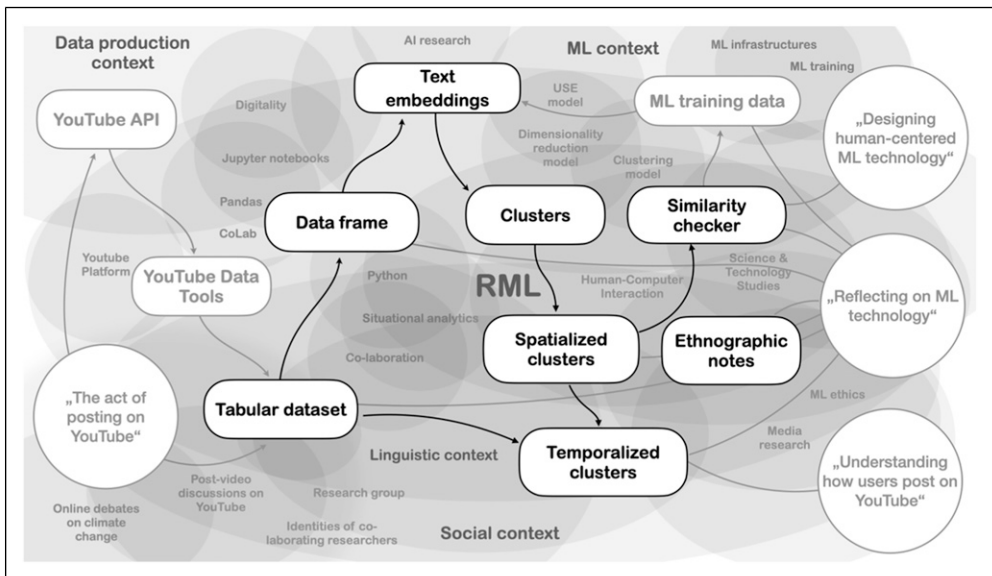


Figure 3. Situational map of our co-laboration. Source: our own visualization.

laboration intensified from left to right during various reflexive sessions organized around anchoring artifacts, which are illustrated by the plates connected through arrows. The debates brought different layers of context to the attention of the co-laborators. These layers of context are visualized as fuzzy clusters in the background. The mapping is not intended to be complete or objectively accurate, but represents another situated view on co-laborative CTP and ways to assemble it.

This first description of our activities consciously refrained from problematizing elements of the software and interdisciplinary engagement and presents both as accomplished facts. The next sections introduce alternative perspectives triggered by the infrastructured elements of co-laborative reflexivity.

Negotiating comparabilities

Nothing triggered more discussions and controversial views within the co-laboration than the concept of ‘comparison’. COSCI and SOSCI frequently argued about the ways the ML techniques enable or should enable comparisons between different datasets or datapoints in datasets. This was also a recurrent theme in the seminar organized at university. Students initially made many claims about comparisons they could achieve through the use of the ML pipeline. One student group, for example, tried to compare discursive patterns of climate activists with climate skeptics. A second group sought to identify automated content generation in YouTube comments section, thereby thriving for a differentiation between human users versus bots. A third group aimed at comparing discursive styles in online debates around climate change between different time periods. Working with the data throughout the co-laboration and seminar then brought up many issues associated with claims of data comparability. Comparison necessarily operationalizes means of categorization and reduction (Bowker and Star, 2000). It abstracts similarities and differences out of fuzzy realities and makes the latter fit for scientific analysis. This is especially true for ML techniques, which allow for the establishment of such similarities and differentiations at a large scale. Rather than criticizing the act of categorization per se, it seems productive to analyze how specific forms of comparison construct categories and what consequences follow (Deville et al., 2016: 27). The co-laborators undertook this evaluation of comparison by ML, thereby focusing especially on different disciplinary understandings of the category work and this shaped their technical and analytical practice. This can be illustrated by the problem of ‘finding the appropriate number of clusters’.

Clusters are distinct groups in data that are similar to each other whereas clusters in different groups are as dissimilar as possible. Cluster analysis is a corresponding field in computer science that aims at grouping data with computational methods based on features available to computational processing (Kaufman and Rousseeuw, 2005). In the current project, the features used for clustering the data are defined by the USE language model. In the context of the co-laboration, COSCI and SOSCI could not agree on an optimal number of clusters. SOSCI argued that he would only be able to handle ten clusters due to cognitive and time constraints. Taking as a reference his normal qualitative coding practices, he believed that taking more than 10 categories as a starting point would not be manageable. He would be obliged to go through all the clusters to build an understanding of the way the model grouped data and then develop meaningful labels in the context of his research questions. He did not trust the ML technique enough to spend several days only on the labeling of clusters. COSCI, in contrast, proposed approximately a hundred clusters. He interpreted the question of ‘how many clusters should there be?’ as a question of parameter optimization and believed that ten clusters would lead to suboptimal and, therefore, disappointing results. K-medoids, the method used to generate clusters, needs the amount of clusters it should find as an input

parameter. This cannot be automatically determined, but has to be adjusted repeatedly by the ML expert. COSCI used a heuristic method to do this – the *elbow method* (Thorndike, 1953), a widely used heuristic to determine the optimal setting of the ‘cluster amount’ parameter. A line plot is generated that shows a mathematical measure of the quality of the clustering (compactness of clusters and separation between clusters) as a function of the ‘cluster amount’. In an idealized example, this plot shows a curve with a kink (an elbow). According to the method, the optimal number of clusters can be found at this kink. The reasoning behind this is that the parameter ‘cluster amount’ should be high enough to separate all the clusters, but not so high as to split up clusters.

The reflection of these social perplexities in the team then led to reconfigurations of the practices of both SOSCI and COSCI. SOSCI noticed that he would not be able to use the clusters as a ‘scientific proof’ for the existence of certain discursive elements and show their distribution and significance in the corpus. The way the model built clusters was too far away from the categories SOSCI could use for his final argumentation (e.g., identifying discursive practices on social media platforms). Instead, the clustering enabled an alternative view of the data and, thereby, discovered new perspectives. Put otherwise, the co-laborators abandoned the view of clusters as representations of social or discursive categories and employed a more pragmatist understanding of the clustering. Rather than asking how the clustering can represent the dataset optimally, the co-laborators asked what clusters *do* within the scope of the interaction and co-laboration. This also enabled COSCI to think about the design of the ML pipeline in a different way. Optimizing parameters, such as the ‘cluster amount’ by measuring metrics was no longer the focus. This included allowing SOSCI to change the ‘cluster amount’ manually and analyze the clusters with different values. Secondly, further features were implemented that allowed for a rather agnostic evaluation of what specific clusters mean, represent or groups together. One such feature was calculating the top words in each cluster. The feature generated a list of the most relevant words in each of the clusters according to a simple formula. This helped the SOSCI to develop a feeling for clusters and undertake comparisons between groups and datapoints, without necessarily deciding what the similarities and differences represent in the whole corpus.

Shifting contextual attention

A second term that triggered misunderstandings and unraveled frictions in the co-laboration was ‘context’. While SOSCI was impressed by the way the language model could discriminate between different thematic areas, formalistic attributes and even emotional tones within YouTube comments, he was still skeptical whether and how the ML technique could be made productive in digital media research. His verbalization of the problem was that ‘the model lacks context of the data’. COSCI, then, had a very clear idea about how the model contextualized the data at hand. A reflection on this perceived problem in a reflexive session then revealed that the researchers had very different understandings of a ‘context’ based on their diverging perspectives on the data at hand.

Context, for COSCI, referred to the information that a specific model takes into account when it determines the position of a data point compared to other datapoints. The scientific publication charting the USE model describes it as follows:

The transformer-based sentence encoding model constructs sentence embeddings using the encoding sub-graph of the transformer architecture (...). This subgraph uses attention to compute context aware representations of words in a sentence that take into account both the ordering and identity of all the other words. The context aware word representations are converted to a fixed length sentence encoding vector by computing the element-wise sum of the representations at each word position. takes as input a

lowercased PTB tokenized string and outputs a 512-dimensional vector as the sentence embedding. (Cer et al., 2018: 1)

USE transforms each word into a vector that represents its semantic content in relation to (in the context of) a learned model of the distribution of the training data. The direction of the vector is sensitive to the linguistic context, that is, to the words surrounding the one currently analyzed. By relating the words that occur together in the same sentence, the model extracts something about the meaning of words and, by extension, sentences formed by combining words. The size of the linguistic context, that is, the number of preceding and succeeding words each word is related to, crucially influences the ability of the model to determine the meaning of a sentence. This can be illustrated exemplarily by the role of the word ‘it’ at the end of the following sentence: ‘Climate change is an essential threat to humanity, we have to deal with it’. The ‘it’ could refer to either ‘climate change’ or ‘humanity’. Determining this word’s meaning necessitates a knowledge of the complete sentence. The key feature of transformer-based models, such as USE, is that the operation of determining word associations can be performed for each word in parallel. This is achieved through the use of so-called ‘multi-head self-attention’. Conceptually, each self-attention head models a certain pattern of words that should be paid attention to when determining the direction of a word’s vector. Activation patterns of self-attention heads can often be identified with certain linguistic characteristics (Rogers et al., 2020). Multiple heads are combined in parallel so that each model has a different aspect of the relationships between words in a sentence. In USE, multi-head self-attention modules are combined with other types of layers and then stacked sequentially, which also allows the model to capture more abstract relationships.

For SOSCI, by contrast, context was understood as the socio-technical conditions governing data production. This understanding was strongly informed by the scientific literature in science and technology studies that challenges the status of data as ‘something given’.⁵ Geoffrey Bowker and Lisa Gitelman have famously argued that data are never ‘raw data’, but are ‘cooked’ (Bowker, 2005; Gitelman, 2013). Other authors argue in the same direction and propose alternative terms such as ‘sublata’ (achievements) (Latour, 1999) or ‘capta’ (taken) (Kitchin, 2014). The consequence of this conceptual shift for scientific analysis is that data should never be handled as a neutral or objective category, but as imprinted with the conditions of their production. These conditions determine how datasets matter or ‘act’ in the world. SOSCI had particular concerns in mind when he talked about the relevance of data context in the co-laboration. On the one hand, (re-)contextualization referred to a desideratum to be achieved by the use of the ML technique – gaining new insights about the ways people act and interact on social media platforms. Which elements determine the way people comment on YouTube? Do people refer more to the video content or to the comments of other users? To what extent are comments pre-structured through the affordances of YouTube’s post-video discussion section? SOSCI’s hope was that the language model would be able to unravel explicit markers for discursive practices (e.g. identity-building words, slogans or symbols) or distributions of formalistic features (e.g., length, language). These insights could then, hopefully, be used for higher level analytical approaches, such as controversy mapping of online debates (Marres and Moats, 2015). On the other hand, SOSCI referred to context as the very condition for the meaningfulness of information represented by the ML technique. The ML technique in question could only be trusted if its way of handling YouTube comments coincided, to some extent, with the analyst’s idea of structures in online debates. At some point in the co-laboration, however, the data appeared to be further away than ever from every possibility of interpretation in the context of social research. Put differently, the structure of the data appeared to be completely detached from the

presumed original production context. It was through another reflexive session that the co-laborators identified elements that fueled this contextual detachment and lack of interpretability.

Most notably, the algorithms in the pipeline all give preference to a spatial mapping of data over other possible representational forms. From a technical perspective, spatial mappings are considered advantageous because they turn the issues at stake into a (linear) algebra problem. Computers excel at performing operations in this domain because these operations can often be performed in parallel. This spatial mapping, however, appeared to be unsuited for the representation of the social activity at hand, namely, temporally unfolding human interactions and debates. Based on this finding, the co-laborators could then invent techniques that would approximate the data and the perceived social category in question (i.e., online interaction and debate).⁶ They introduced a new view visualizing the unfolding of clusters in a temporal perspective.

The temporal view offered a new perspective to SOSCI on the data, while maintaining the crucial contextual information necessary to engage in analytical practice. The temporal perspective (Figure 4), for example, enabled the tracing of comments and debates addressing issues such as online misinformation (cluster 5), conspiracy narratives (cluster 12) and the diminishing value of beach property due to the rise in sea levels (cluster 16). One could equally trace variations of formal structures, such as debates in the Spanish language (cluster 6), brief reactions (cluster 15) and toxic speech (cluster 9).

SOSCI and COSCI also reflected, however, on higher level issues, which were not linked to data analysis or model understanding, but to the very practices of co-laboration. It became clear that collaborative perplexities in different stages of project development unraveled different kinds of contextual constellations to be taken into account. Reflexive sessions first turned around questions of institutional support necessary for co-laboration, then to infrastructures enabling interdisciplinary

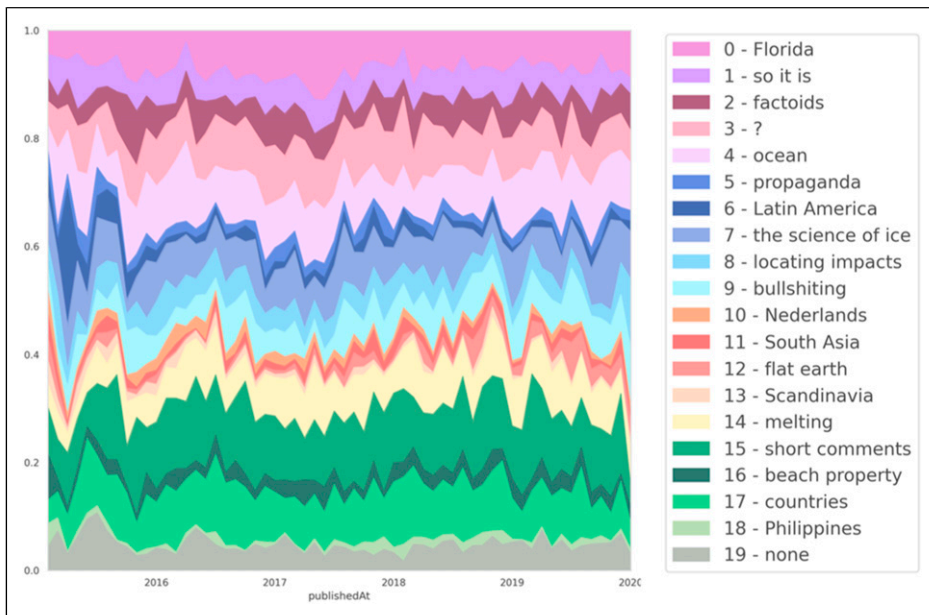


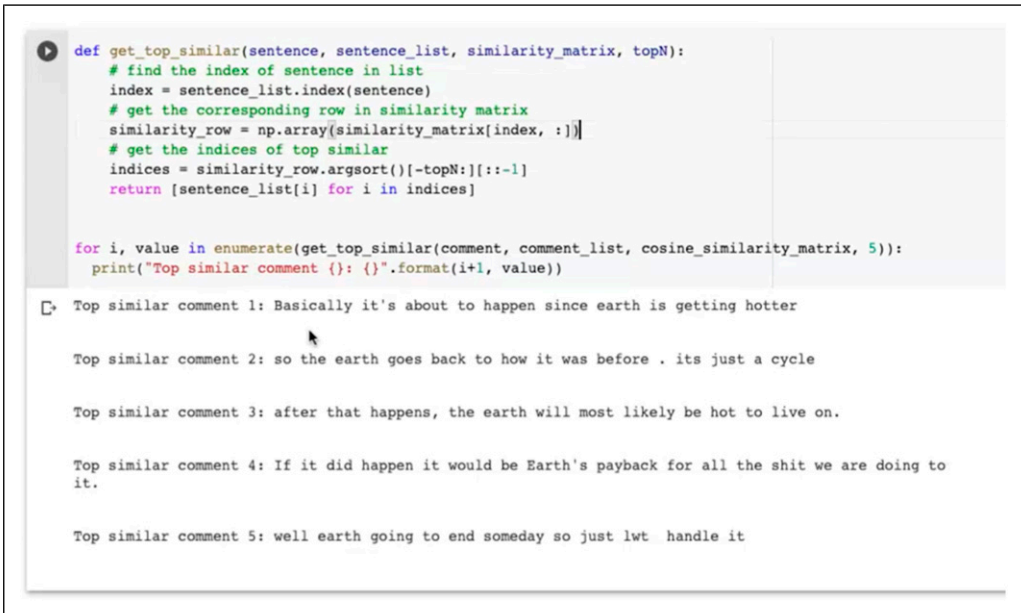
Figure 4. Temporal view of clusters, with manual cluster labels. Each colored region represents the changing amount of comments belonging to the corresponding cluster in proportion to the total amount of comments over time. Source: our own visualization.

reflection, later to logics of modeling and visualization techniques, to data analysis and so on. The co-laborators referred to this critical practice as ‘shifting contextual attention’, thereby drawing on the terminology of attention ML models described before. While the co-laboration unfolds, co-laborators can decide on new layers of context to be paid attention to. These layers of context are not ‘discovered’ or ‘reconstructed’ from an original reality or truth, but they are actively built through the moments of interdisciplinary reflection – or, as Paul Dourish has argued, context and activity have to be understood as mutually constitutive (Dourish, 2004: 29).

Challenging (dis-)similarities

While the co-laborating researchers tried to stick to the agonistic–antagonistic mode of co-laboration, it is apparent that working closely together during a considerable time changes the character of mutual engagement. Co-laborators necessarily attune to each other to some degree, which alters the prospect of using irritation for reflection. For example, early reflexive sessions profited from the different expectations of project members regarding the capabilities of ML techniques. Technology is filled with promises and its alleged potentials can be made productive for critique (Ames, 2019; Anand et al., 2018), as well as for co-laborative reflection in particular. Reflexive sessions at the beginning of the co-laboration were also able to take advantage of misaligned conceptualizations of the identities and positionings of the other co-laborators. Interdisciplinary engagement necessarily involves practices of ‘othering’ (Jensen, 2011). Agonistic–antagonistic modes of interdisciplinarity tend to pay attention to these practices of othering and to the shifting identities and positionings of co-laborating researchers.

The longer SOSCI and COSCI worked together and thereby necessarily became experts of the technology at hand, however, the more the focus of reflection moved to particular operating modes of the chosen ML technique (language models, text embeddings). The mode of reflection shifted from the analysis of breakdown and repair to more targeted activities of technical inquiry (Baumer, 2015). One object of reflection in these later stages of the co-laboration was triggered by the relationship between context and activity discussed further above. It was frequently an object of debate in the co-laboration whether a certain way the model determined (dis-)similarities was produced in the present or whether it was imprinted into the technique beforehand. This differentiation was important as it would determine whether a certain mechanism could be changed and improved or whether it has to be taken as given. In the specific case, this problem related to the characteristics of USE as a pretrained language model. The new context that emerged in the co-laboration was, accordingly, the way the USE model had been trained. These training conditions would necessarily create a given ‘context’ out of hand for the researchers and not to be altered easily through current ‘activity’. A number of aspects were identified as relevant to the configuration of the training context: training data, methods and purpose. Similar to other language models, USE was designed to solve language-related tasks, such as question answering, classification or semantic search. It was trained to perform well in such tasks by both unsupervised and supervised learning. On the one hand, the model had been trained on publicly accessible, online text ensembles, such as Wikipedia or web news aggregators (unsupervised learning). In this process, the model ‘learned’ which words and sentences are often used in close proximity to each other. When a text, for instance, mentions ‘climate change’, it might also mention ‘environment’, ‘disaster’ or ‘politics’. On the other hand, USE was trained with datasets that had been manually annotated by humans (supervised learning). Such annotation processes are commonly carried out as task fulfillments by both crowdworkers (e.g., Amazon Mechanical Turk) and expert annotators.



```

def get_top_similar(sentence, sentence_list, similarity_matrix, topN):
    # find the index of sentence in list
    index = sentence_list.index(sentence)
    # get the corresponding row in similarity matrix
    similarity_row = np.array(similarity_matrix[index, :])
    # get the indices of top similar
    indices = similarity_row.argsort()[-topN:][::-1]
    return [sentence_list[i] for i in indices]

for i, value in enumerate(get_top_similar(comment, comment_list, cosine_similarity_matrix, 5)):
    print("Top similar comment {}: {}".format(i+1, value))

```

Top similar comment 1: Basically it's about to happen since earth is getting hotter
 Top similar comment 2: so the earth goes back to how it was before . its just a cycle
 Top similar comment 3: after that happens, the earth will most likely be hot to live on.
 Top similar comment 4: If it did happen it would be Earth's payback for all the shit we are doing to it.
 Top similar comment 5: well earth going to end someday so just lwt handle it

Figure 5. Evaluating how the USE model determines similarities between comments. Source: our own visualization (Notebook screenshot).

It makes sense to analyze these datasets more closely in order to understand the context the model considers when processing data. The SNLI data corpus (Bowman et al., 2015), for example, was used to train the USE model. The dataset was constructed in a crowdsourcing effort in which a total of 2500 crowdworkers contributed on Mechanical Turk. These crowdworkers were given sentences and asked to supply three other sentences, one for each category (entailment, contradiction, independence). These annotators would be asked, for example, to identify semantic entailments or contradictions: ‘Gravity exists on earth’ entails ‘When a ball is dropped it falls to the ground’, whereas ‘Global temperatures are rising’ contradicts ‘Global temperatures are decreasing’. Other sentences do not have such logical relationships and are, therefore, considered independent. The sentences that were given to workers were sourced from image descriptions of Flickr images (Young et al., 2014), however, the images were not shown. After sentence pairs had been collected, they were validated by another set of crowdworkers, specifically, thirty workers that were highly trusted according to their performance in past tasks. It is not clear, however, what these tasks were and why they enabled these workers to be good judges of semantic entailment or contradiction. The detailed consideration of this dataset hints at potential limitations of the model. For instance, the dataset was sourced from Flickr image descriptions and thus presumably only contains descriptions of things that can be depicted in images, while abstract concepts (e.g., feelings, values) are not covered. Contrary to what the dataset’s name suggests, the model has only been trained to infer logical relationships between concepts and its ability to seamlessly generalize to the abstract *in the same way* humans would is not guaranteed.

After a reflexive session, the co-laborating researchers decided to make this new problematization of the training context productive within the design of the software. The idea was to design a *similarity checker* as a module of the ML pipeline. The similarity checker would allow researchers to develop an understanding about what pieces of texts the language model deemed similar and

negotiate these similarities within the co-laboration. This would then allow deeper investigations into the reasons for particular claims of similarity and difference. As a matter of fact, the similarity checking feature was already an object of debate in an early stage of the co-laboration. It was an element in the common practice in software development to playfully engage with the functions of software or, in this case, an ML model. This practice is usually carried out by giving inputs to the system that test its functionality. Edge cases are often given as inputs that are expected to result in interesting output. At the time of this early project meeting, however, the functionality of evaluating the calculated similarities by the ML technique was not considered to be an interesting feature for SOSCI, but only a tool for computer scientists to evaluate whether and how the model works. Based on the five similar comments shown to him (see [Figure 5](#)), he considered the model to work improperly.

SOSCI considered the calculated similarities of comments as meaningless based on his thematic expertise regarding online climate change discourses. From this perspective, comment 1 in Figure X has to be considered as fundamentally different if not opposite to comment 2. While comment 1 accepts anthropogenic climate change as a reality, comment 2 describes ‘a cycle’ and the situation ‘going back to normal’, thereby evoking key rhetorical figures of climate change deniers. The two comments, therefore, have to be considered as diametrical opposites in the discursive space presented in the social science literature (e.g., [Nisbet, 2009](#)). Following this situation of perplexity linked to apparently meaningless results of the ML technique, the co-laborators banned the similarity checking functionality from the software and discussions, assigning it solely to technical model evaluation and, thus, to the responsibility sphere of the computer scientists involved.

The functionality’s role and value of the functionality changed, however, after the reflection on ‘context’ and ML training data. Its position changed from the technical sphere (governed by the computer scientists) to that of interdisciplinary negotiation and reflection (all co-laborators). To enable this new role, the similarity evaluating functionality was transformed into an independent software module (i.e. a cell in the Jupyter Notebook) and tailored to trigger co-laborative engagement and reflection. This can be illustrated by one exemplary activity: co-laborators would get inspiration from the real YouTube comments and invent new sentences to ‘trick’ the model and use their domain expertise and common knowledge to identify blind spaces in the model space.

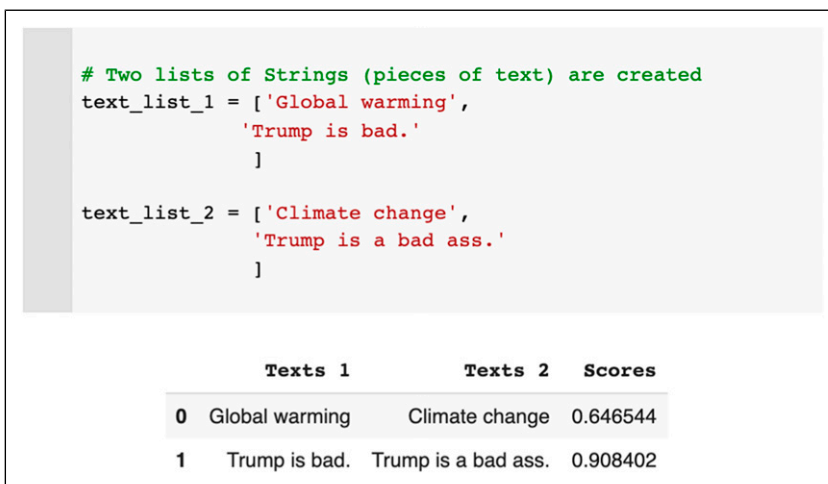


Figure 6. Similarity checker (input and output in Jupyter Notebook). Source: our own visualization.

For example (see [Figure 6](#)), the USE model attributes a similarity score of 0.64% to the two terms ‘global warming’ and ‘climate change’, which appears very low given the strong relationship between the two. By contrast, USE attributes the very high similarity score of 0.9 to the sentences ‘Trump is bad’ and ‘Trump is a bad ass’, despite the obvious and consequential difference between them (negative connotation versus positive connotation of former US President Donald Trump). This failure of the model to identify this obvious difference in terms of the mundane understanding of language led to a discussion of the real-world consequences of such shortcomings. Building on these reflections, the co-laborators also addressed these questions in a session of their seminar at university. The case of fighting hate speech and disinformation online was used to evaluate how automatic detection mechanisms based on ML techniques would identify harmful content and trigger sanctions against social media users. The seminar involved experiments with multiple ML techniques and a discussion of scientific literature addressing the failures of hate speech detection. This literature shows, for example, that many automated techniques fail to understand nuances such as irony and sarcasm and that they can be tricked by integrating simple, positively connotated words such as ‘love’ ([Gröndahl et al., 2018](#)). In sum, the example of the similarity checker shows that collaborative reflection in the context of AI research cannot only be triggered by an analysis of breakdown and shifting positionings in cooperative relationships, but can also be imprinted into the materiality and practice of ML.

Conclusion: reconsidering reflexivity in AI research

In this section, we come back to Agre’s understanding of reflexivity in CTP and compare it to our own reflexive research practices in contemporary ML and DS. For Agre,

“Reflexive research cultivates a critical self-awareness, including itself among its objects of study and developing useful concepts for reflecting on the research as it is happening.” ([Agre 1997a: 27](#))

It is challenging to position Agre’s understanding of reflexive research among other conceptualizations of reflexivity in the scientific literature. As a computer scientist by training and interdisciplinary scholar by heart, he took a great liberty to mix different traditions of the social sciences and humanities. One can, however, identify several tactics for reflexivities in Agre’s work and then put it in relation with similar practices described in the scientific literature.

Agre’s reflexivity

Agre’s reflexive practices include (1) tracing the historical constitution of AI research, (2) unraveling its self-conception as a research field, (3) conducting a critical discourse analysis of the field’s terminology, (4) describing his work and shifting positions in AI research and (5) making these techniques productive within experimental AI modeling and programming. None of these techniques alone achieves the level of reflexivity imagined by Agre, but only their close entanglement, adjustment and reconciliation, which is ultimately represented by the CTP approach. In his critical evaluation of the historical constitution of AI as a field of study, Agre was clearly influenced by historians of science and technology and their reading of reflexivity as self-reference to knowledge production in one’s own scientific field ([Ashmore, 1989: 32](#)). He references key figures of the field such as Thomas Kuhn and David Bloor, but also elaborates on his personal friendship and intellectual exchange with Paul Edwards. Agre’s discourse analysis of symbolic AI terminology has, as well, strong conceptual overlaps with Malcolm Ashmore’s work ([1989](#)), proposing a radical

reflexivity that engages in deconstructive instances of constructivist arguments. While not explicitly mentioning Ashmore, Agre has adopted his terminology ('reflexive thesis') and it is, therefore, quite certain that he was aware of this work. Agre's understanding of reflexivity also resonates with scholarship highlighting the situatedness of knowledge (Haraway, 1988), and scientific knowledge in particular. Reflexivity, in this context, is understood as a capacity to take this situatedness into account and make it productive for scientific knowledge production. It acknowledges that the meaning of knowledge is always dependent upon the person who created it (Rose, 1997), which naturally includes the position and perspective of the researcher. For Philip Agre, this kind of standpoint reflexivity (Lynch, 2000) included a documented assessment and evaluation of both his own theoretical positionings and his technical practice. An important feature of Agre's reflexivity in the context of AI research is its conceptualization as material-semiotic practice. Reflexive research should, according to Agre, not be limited to mental instances of critique (Agre, 1997: 154), but be embedded in and make constructive recommendations to technical practice.

"As I worked my way toward a critical technical practice, this was the part that I found hardest: maintaining constructive engagement with researchers whose substantive commitments I found wildly mistaken. It is tempting to start explaining the problems with these commitments in an alien disciplinary voice, invoking phenomenology or dialectics as an exogenous authority, but it is essentially destructive. The constructive path is much harder to follow but more rewarding. Its essence is to evaluate a research project not by its correspondence to one's own substantive beliefs but by the rigor and insight with which it struggles against the patterns of difficulty that are inherent in its design." (Agre, 1997: 154f)

This constructive attitude towards technological design is the essence of CTP and differentiates Agre's approach to reflexive research from most other understandings of critical research and their reflexivity.⁷ This productive stance also explains the influence of CTP on design studies and Human-Computer Interaction, incorporated in approaches such as *Reflective HCI* (Dourish et al., 2004), *Reflexive Design* (Sengers et al., 2005) and *Reflexive Informatics* (Baumer, 2015) who all link back to Agre and his CTP.

Reflexive data science

In our paper, we aimed at highlighting additional devices of reflexivity that may enrich the original CTP approach and make it fit for an operationalization in contemporary AI-related research. We propose the term *Reflexive Data Science (RDS)* for this kind of research practice. The now-common term 'data science' was chosen to acknowledge the tailoring of our approach to today's socio-technical settings of data-driven ML systems, which are fundamentally different from Agre's original context of symbolic AI. Drawing on our empirical study, we suggest that RDS operationalizes the following strategies and suggest some elements for further research:

(1) *Organizing encounters of social contestation.* We propose to organize conducive constellations for social and (inter-)disciplinary contestation in AI-related research inspired by agonistic-antagonistic modes of interdisciplinarity (Barry et al., 2008) and co-laboration (Niewöhner, 2015). Such social contestation could either be stabilized in an institutional manner (e.g., as mode of conduct of a research group), or organized on a more situative basis between individually co-laborating researchers. As we have shown in our study, social contestation must not necessarily happen along disciplinary boundaries. It can, equally, include other layers of belongings and commitments that produce social frictions. We think, however, that a natural expansion of the interdisciplinary scope

of RDS would be to seek a participation of non-academic actors in the co-laboration. This would possibly require a shift of the attentional foci of reflexive sessions from epistemological considerations to other concerns with AI (e.g., the scrutinization of ethical questions linked to particular ML techniques and application contexts).

(2) *Infrastructuring a network of anchoring devices enabling reflection.* RDS employs a comprehensive network of (digital or analog) tools that enable discursive negotiations, the analysis of socio-technical (re-)positionings, and other ways of reflection. This may include anchoring devices (Hirsbrunner, 2021) for discursive negotiation of programming code such as Jupyter Notebooks, software versioning systems comparing different states of the technology, but also video-meeting software, (auto-)ethnographic notes and visual mapping tools. A promising direction for further research in an STS context would be to investigate whether we could actually speak of reflexivities in AI infrastructure or even of *reflexive AI infrastructures*. This would require a theoretical reconsideration of different understandings of reflexivity and infrastructure, and it would create a productive linkage to today's prevalent discussion around *explainable AI* (Gunning et al., 2019).

(3) *Negotiating timely matters of concern.* Diverging from Agre's original CTP, we are skeptic whether a critical scrutinization and replacement of dominant metaphors in today's data science practice will be particularly effective and meaningful. Compared to the relatively closed and stable conceptual world of symbolic AI research, today's data science practice handles metaphors more dynamically. Many metaphors are, as e.g. 'comparison', 'context' and 'similarity' discussed in our study, iteratively generated as boundary concepts enabling cooperation and coordination of heterogeneous actors. We believe, therefore, that discursive critique of concepts and metaphors in contemporary AI-related research should also take a more tentative character prompted by the use of language in a situated context. RDS will require a careful shifting of contextual attention during the research process depending on situated constellations of co-laborative contributors, the state of technology, investigative insights and failures.

(4) *Designing for reflection.* While infrastructuring reflexivities (2) is about the thoughtful entanglement of distributed techniques enabling reflection, RDS may also think about explicit functionalities of software triggering reflection of co-laborating scientists or other stakeholders. The AI-related software developed in the co-laboration is now accessible as *Reflexive ML Toolbox* in a repository on the Github platform (Tebbe et al., 2021). The availability for re-use brings up the question to what extent reflexivity (i.e., a capacity for reflection) is or could be explicitly imprinted into AI-related software design. This may include functionalities tailored for reflection such as the exemplary similarity checker discussed above. As a possible direction of future research in the HCI-context, it is promising to think about the design of elaborated strategies for reflection guidance, including features of technical interfaces, but also reflexive data structures, explanations and practices.

Acknowledgements

We would like to thank our colleagues at the HCC research group for their valuable input, Isabel Schmuck for her implementation work, the editors and anonymous reviewers for the productive feedback and Philipp Saunders for his impeccable proofreading of the manuscript.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Geo. X – the research network for geosciences in Berlin and Potsdam and by the Cluster of Excellence “Matters of Activity. Image Space Material” funded by the German Research Foundation (EXC 2025).

ORCID iDs

Simon David Hirsbrunner  <https://orcid.org/0000-0001-5529-4171>

Claudia Müller-Birn  <https://orcid.org/0000-0002-5143-1770>

Notes

1. Symbolic and connectionist (or subsymbolic) AI represent two opposed paradigms of AI research. Symbolic approaches develop explicit relationships between concepts based on formal logic. Connectionist techniques, on the other hand, are based on statistics and learn from large amounts of data. Today's most common AI approaches such as neural networks, ensemble models, regression models, decision trees and support vector machines belong to the connectionist tradition. For a comparison of theoretical underpinnings of symbolic and connectionist AI, see [Dreyfus and Dreyfus \(1988\)](#).
2. See for example the yearly FAccT conference: <https://facctconference.org/> (Accessed 15 Feb. 2022)
3. See www.odproject.org/2019/07/15/critiquing-and-rethinking-fairness-accountability-and-transparency/ (Accessed 4 Sept. 2021)
4. Considering the auto-ethnographic perspective of the paper, an anonymization is neither possible nor intended.
5. ‘Datum’ and ‘data’ etymologically refer to ‘something given’ in Latin.
6. As a matter of fact, recontextualizing reality through symbolic systems and technical mediation is something we do all the time. We experience a situation and document its unfolding in a text, say in social research within an ethnographic memo. The structure of the signs at hand (text) has no resemblance to the shape or feeling of the situation we experienced. Nevertheless, nobody would challenge the general idea that we can reflect on the original experience by looking at a written story describing it.
7. A notable exception is Niewöhner's conceptualization of reflexive research, which also highlights the aspect of reflexivity as material-semiotic practice ([Niewöhner, 2018](#)).

References

- Agre P (1997) Toward a critical technical practice: lessons learned in trying to reform AI. In: Bowker G, Gasser L, Star L, et al. (eds) *Social Science, Technical Systems and Cooperative Work: Beyond the Great Divide*. Erlbaum.
- Agre P (1997a) *Computation and Human Experience*. Cambridge: Cambridge University Press. DOI: [10.1017/CBO9780511571169](https://doi.org/10.1017/CBO9780511571169).
- Agre P and Chapman D (1987) *Pengi: An Implementation of a Theory of Activity*. In: Proceedings of the Sixth National Conference on Artificial Intelligence, Seattle, 1987, pp. 196–201.
- Ashmore M (1989) *The Reflexive Thesis: Writing Sociology of Scientific Knowledge*. University of Chicago Press.

- Ames MG (2019) *The Charisma Machine: The Life, Death, and Legacy of One Laptop Per Child. Infrastructures*. Cambridge, MA, USA: MIT Press.
- Anand N, Gupta A and Appel H (eds) (2018) *The Promise of Infrastructure*. Durham: Duke University Press.
- Barry A, Born G and Weszkalnys G (2008) Logics of interdisciplinarity. *Economy and Society* 37(1): 20–49. DOI: [10.1080/03085140701760841](https://doi.org/10.1080/03085140701760841).
- Bates J, Cameron D, Checco A, et al. (2020) Integrating FATE/critical data studies into data science curricula : where are we going and how do we get there? In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, January 2020, pp. 425–435.
- Baumer EPS (2015) Reflective informatics: conceptual dimensions for designing technologies of reflection. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul Republic of Korea, 18 April 2015, pp. 585–594. ACM. DOI: [10.1145/2702123.2702234](https://doi.org/10.1145/2702123.2702234).
- Baumer EPS, Mimno D, Guha S, et al. (2017) Comparing grounded theory and topic modeling: extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology* 68(6): 1397–1410. DOI: [10.1002/asi.23786](https://doi.org/10.1002/asi.23786).
- Bertelsen O and Bødker S (2002) Interaction through clusters of artefacts. In: 11th European Conference on Cognitive Ergonomics (ECCE–11), Catania, Italy: 8.
- Blei DM, Ng AY and Jordan MI (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research* 3(Jan): 993–1022.
- Bowker GC (2005) *Memory Practices in the Sciences. Inside technology*. Cambridge, Mass: MIT Press.
- Bowker GC and Star SL (2000) *Sorting Things Out: Classification and Its Consequences. Revised. ed. edition*. Cambridge, Mass. / London, England: MIT Press.
- Bowman SR, Angeli G, Potts C, et al. (2015) A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, September, pp. 632–642.
- Cer D, Yang Y, Kong S, et al. (2018) *Universal Sentence Encoder. arXiv:1803.11175 [cs]*. Available at: <http://arxiv.org/abs/1803.11175> (accessed 23 November 2020).
- Clarke AE (2003) Situational analyses: grounded theory mapping after the postmodern turn. *Symbolic Interaction* 26(4): 553–576. DOI: [10.1525/si.2003.26.4.553](https://doi.org/10.1525/si.2003.26.4.553).
- Clarke AE, Friese C and Washburn R (eds) (2015) *Situational Analysis in Practice: Mapping Research with Grounded Theory*. London: Routledge.
- Coenen A, Reif E, Yuan A, et al. (2019) *Visualizing and Measuring the Geometry of BERT. arXiv:1906.02715 [cs, stat]*. Available at: <http://arxiv.org/abs/1906.02715> (accessed 23 June 2021).
- Deville J, Guggenheim M and Hrdličková Z (2016) Introduction: the practices and infrastructures of comparison. In: Deville J, Guggenheim M, and Hrdličková Z (eds) *Practising Comparison: Logics, Relations, Collaborations*. Mattering Press, 17–41.
- Dourish P (2004) What we talk about when we talk about context. *Personal and Ubiquitous Computing* 8(1): 19–30. DOI: [10.1007/s00779-003-0253-8](https://doi.org/10.1007/s00779-003-0253-8).
- Dourish P, Finlay J, Sengers P, et al. (2004) Reflective HCI: towards a critical technical practice. In: Extended abstract of the 2004 Conference on Human Factors in Computing Systems, Vienna, Austria, 24–29 April 2004. DOI: [10.1145/985921.986203](https://doi.org/10.1145/985921.986203).
- Dreyfus HL and Dreyfus S (1988) Making a mind versus modeling the brain: AI back at a branchpoint. *Daedalus* 117(1): 15–43.
- Dreyfus HL (ed) (1982) *Husserl, Intentionality, and Cognitive Science*. Cambridge, MA: MIT Press.
- Dreyfus HL (1972) *What Computers Can't Do: A Critique of Artificial Reason*. 1st edition. New York: Harper & Row.
- Gitelman L (ed) (2013) *'Raw Data' Is an Oxymoron*. Cambridge, MA: MIT Press.

- Griffiths TL and Steyvers M (2004) Finding scientific topics. *Proceedings of the National Academy of Sciences* 101(suppl 1): 5228–5235. DOI: [10.1073/pnas.0307752101](https://doi.org/10.1073/pnas.0307752101).
- Gröndahl T, Pajola L, Juuti M, et al. (2018) All you need is ‘love’: evading hate speech detection. In: Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security, Toronto Canada, 15 January 2018, pp. 2–12. ACM. DOI: [10.1145/3270101.3270103](https://doi.org/10.1145/3270101.3270103).
- Gunning D, Stefik M, Choi J, et al. (2019) XAI—Explainable artificial intelligence. *Science Robotics* 4 (37). DOI: [10.1126/scirobotics.aay7120](https://doi.org/10.1126/scirobotics.aay7120).
- Haraway D (1988) Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14(3): 575. doi: [10.2307/3178066](https://doi.org/10.2307/3178066).
- Henriksen A and Bechmann A (2020) Building truths in AI: making predictive algorithms doable in healthcare. *Information Communication and Society* 23 (6): 802–816. DOI: [10.1080/1369118X.2020.1751866](https://doi.org/10.1080/1369118X.2020.1751866).
- Hirsbrunner SD (2021) *A New Science for Future. Climate Impact Modeling and the Quest for Digital Openness. Locating Media/Situierte Medien. Bielefeld: Transcript*. <https://www.transcript-verlag.de/978-3-8376-5265-9/a-new-science-for-future/>
- Hirsbrunner SD (2021a) Negotiating the data deluge on YouTube: practices of knowledge appropriation and articulated ambiguity around visual scenarios of sea-level rise futures. *Frontiers in Communication* 6. DOI: [10.3389/fcomm.2021.613167](https://doi.org/10.3389/fcomm.2021.613167).
- Jaton F (2021) Assessing biases, relaxing moralism: on ground-truthing practices in machine learning design and application. *Big Data and Society* 1(8): 1–15. DOI: [10.1177/20539517211013569](https://doi.org/10.1177/20539517211013569).
- Hirschauer S (2013) Verstehen des Fremden, Exotisierung des Eigenen. *Ethnologie und Soziologie als zwei Seiten einer Medaille. Ethnologie im 21*: 229–248.
- Jaton F (2017) We get the algorithms of our ground truths: designing referential databases in digital image processing. *Social Studies of Science* 47(6): 811–840. DOI: [10.1177/0306312717730428](https://doi.org/10.1177/0306312717730428).
- Jensen SQ (2011) Othering, identity formation and agency. *Qualitative Studies* 2(2): 63–78. DOI: [10.7146/qs.v2i2.5510](https://doi.org/10.7146/qs.v2i2.5510).
- Karasti H (2014) Infrastructuring in participatory design. In: Proceedings of the 13th Participatory Design Conference on Research Papers - PDC '14, Windhoek, Namibia, 2014, pp. 141–150, ACM Press. DOI: [10.1145/2661435.2661450](https://doi.org/10.1145/2661435.2661450).
- Kaufman L and Rousseeuw PJ (2005) Finding groups in data: an introduction to cluster analysis. *Wiley Series in Probability and Mathematical Statistics*. Hoboken, NJ: Wiley.
- Kitchin R (2014) *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Thousand Oaks, California: Sage.
- Kluyver T, Ragan-Kelley B, Pérez F, et al. (2016) Jupyter Notebooks – a publishing format for reproducible computational workflows. *Stand Alone*: 87–90. DOI: [10.3233/978-1-61499-649-1-87](https://doi.org/10.3233/978-1-61499-649-1-87).
- Latour B (1987) *Science in Action: How to Follow Scientists and Engineers through Society*. Cambridge, Mass.: Harvard University Press. Cambridge, Mass.: Harvard University Press.
- Latour B (1999) *Pandora's Hope: Essays on the Reality of Science Studies*. 1st edition. Cambridge, MA: Harvard University Press.
- Lee CP (2007) Boundary negotiating artifacts: unbinding the routine of boundary objects and embracing chaos in collaborative work. *Computer Supported Cooperative Work (CSCW)* 16(3): 307–339. DOI: [10.1007/s10606-007-9044-5](https://doi.org/10.1007/s10606-007-9044-5).
- Lynch M (2000) Against reflexivity as an academic virtue and source of privileged knowledge. *Theory, Culture and Society* 17(3): 26–54. DOI: [10.1177/02632760022051202](https://doi.org/10.1177/02632760022051202).
- McInnes L, Healy J, and Melville J (2020) *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. *arXiv:1802.03426 [cs, stat]*. Available at: <http://arxiv.org/abs/1802.03426> (accessed 1 March 2021).

- Marres N (2020) For a situational analytics: an interpretative methodology for the study of situations in computational settings. *Big Data and Society* 7(2). DOI: [10.1177/2053951720949571](https://doi.org/10.1177/2053951720949571).
- Marres N and Moats D (2015) Mapping controversies with social media: the case for symmetry. *Social Media + Society* 1(2): 2056305115604176. DOI: [10.1177/2056305115604176](https://doi.org/10.1177/2056305115604176).
- Miller T, Howe P and Sonenberg L (2017) *Explainable AI: Beware of Inmates Running the Asylum or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences*. *arXiv:1712.00547 [cs]*. Available at: <http://arxiv.org/abs/1712.00547> (accessed 6 September 2021).
- Müller-Birm C, Mackeprang M and Tebbe M (2020) Ideas to Market - Entwicklung eines toolgestützten Vorgehens zur interdisziplinären und branchenübergreifenden Generierung von Verwertungsoptionen - Teilvorhaben 2: Nutzerzentrierte Entwicklung einer Tool-Suite für den kollaborativen Aufbau und Nutzung von Terminologien. Final Project Report. <https://www.tib.eu/de/suchen/id/TIBKAT:1748522787/IDEAS-TO-MARKET-Entwicklung-eines-toolgest/C3/BCTzten?cHash=6c187eef9dda9e0baccdd61981686407> (accessed 13 October 2022).
- Niewöhner J (2016) Co-laborative anthropology: crafting reflexivities experimentally. In: Jouhki J and Steel T (eds) *Etnologinen tulkinta ja analyysi. Kohti avoimempaa tutkimusprosessia*. Helsinki: Ethnos, 81–124.
- Niewöhner J (2015) Epigenetics: localizing biology through co-laboration. *New Genetics and Society* 34(2): 219–242. DOI: [10.1080/14636778.2015.1036154](https://doi.org/10.1080/14636778.2015.1036154).
- Niewöhner J (2018) Assembling Comparators – Assembling Reflexivities. *Science as Culture* 27(4): 563–568. DOI: [10.1080/09505431.2018.1519533](https://doi.org/10.1080/09505431.2018.1519533).
- Nisbet MC (2009) Communicating climate change: why frames matter for public engagement. *Environment: Science and Policy for Sustainable Development* 51(2): 12–23. DOI: [10.3200/ENVT.51.2.12-23](https://doi.org/10.3200/ENVT.51.2.12-23).
- O’Neil C (2016) *Weapons of Math Destruction*. How Big Data Increases Inequality and Threatens Democracy. Broadway Books.
- Park H-S and Jun C-H (2009) A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications* 36(2): 3336–3341. DOI: [10.1016/j.eswa.2008.01.039](https://doi.org/10.1016/j.eswa.2008.01.039).
- Rieder B (2015) *YouTube Data Tools (Version 1.23)*. [Software] Available at: <https://tools.digitalmethods.net/netvizz/youtube/>
- Rogers A, Kovaleva O and Rumshisky A (2020) A primer in BERTology: What we know about how BERT works. *arXiv:2002.12327 [cs]*. Available at: <http://arxiv.org/abs/2002.12327> (accessed 23 September 2021).
- Rose G (1997) Situating knowledges: positionality, reflexivities and other tactics. *Progress in human geography* 21(3): 305–320.
- Saltz J, Skirpan M, Fiesler C, et al. (2019) Integrating ethics within machine learning courses. *ACM Transactions on Computing Education* 19 (4): 1–26. DOI: [10.1145/3341164](https://doi.org/10.1145/3341164).
- Science Insider (2015) How earth would look if all the ice melted. Available at: https://www.youtube.com/watch?v=VbiRNT_gWUQ&t=21s (accessed 26 August 2021).
- Sengers P, Boehner K, David S, et al. (2005) Reflective design. In: Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility, New York, NY, USA, 20 August 2005, pp. 49–58. CC ‘05. Association for Computing Machinery. DOI: [10.1145/1094562.1094569](https://doi.org/10.1145/1094562.1094569).
- Star SL and Bowker GC (2006) How to infrastructure. *Handbook of new media: Social shaping and social consequences of ICTs*: 230–245.
- Star SL and Griesemer JR (1989) Institutional ecology, translations’ and boundary objects: Amateurs and professionals in Berkeley’s Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science* 19.3: 387–420.
- Star SL and Ruhleder K (1996) Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information Systems Research* 7(1): 111–134.

- Tebbe M, Hirsbrunner SD and Müller-Birn C (2021) *Reflexive Machine Learning Toolbox (Version 0.1)* [computer program]. <https://github.com/FUB-HCC/Reflexive-Machine-Learning-Toolbox>
- Thorndike RL (1953) Who belongs in the family? *Psychometrika* 18(4): 267–276. DOI: [10.1007/BF02289263](https://doi.org/10.1007/BF02289263).
- Vaswani A, Shazeer N, Parmar N, et al. (2017) *Attention is All You Need*. *arXiv:1706.03762 [cs]*. Available at: <http://arxiv.org/abs/1706.03762> (accessed 23 June 2021).
- Winograd T and Flores FF (1986) *Understanding Computers and Cognition: A New Foundation for Design*. Intellect Books.
- Young P, Lai A, Hodosh M, et al. (2014) From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2: 67–78. DOI: [10.1162/tacl_a_00166](https://doi.org/10.1162/tacl_a_00166).

Author biographies

Simon David Hirsbrunner, is Geo.X fellow and postdoc at the Human-Centered Computing (HCC) research group of the Institute of Computer Science at Freie Universität Berlin. He conducts research in the field of Science & Technology Studies on matters such as digital communication, collaboration and interdisciplinarity in science, as well as AI and data ethics.

Michael Tebbe is a PhD candidate and researcher in interactive machine learning at the Cluster of Excellence ‘Matters of Activity’ and the Human-Centered Computing (HCC) research group of the Institute of Computer Science at Freie Universität Berlin.

Claudia Müller-Birn is a professor at the Institute of Computer Science at Freie Universität Berlin, head of the Human-Centered Computing (HCC) research group and principal investigator of the Cluster of Excellence ‘Matters of Activity’.