Science & Technology Studies

4/2022

Science & Technology Studies

ISSN 2243-4690

Open access & copyright information

Science & Technology Studies

Volume 35, Issue 4, 2022

Editorial

Articles

Book reviews

Visit our web-site at

www.sciencetechnologystudies.org

Editorial

Antti Silvast

Technical University of Denmark, Denmark/ aedsi@dtu.dk

Dear Colleagues,

It is with great pleasure that I have accepted the role of the new Coordinating Editor of *Science & Technology Studies* (S&TS). The post starts from this issue and lasts for four years. I want to wish all the best to our former Coordinating Editor, Salla Sariola, and thank her for her excellent work. During her term, the journal has grown extensively and continued its dedication to being a fully Open Access journal with no Article Processing Charges. It is a great opportunity to continue working with the solid infrastructure that Salla has helped build for the journal. At the same time, I look forward to collaborating with the skilled Editorial Team whose work I will coordinate.

My name is Antti Silvast and I have acted as an Editor for this journal since 2014. In addition, I was a Guest Editor for three special issues for S&TS from 2012 to 2014. I have a long experience in editing, including my role at S&TS, and starting as a newsletter editor for a network of the European Sociological Association in 2010.

As a journal, we will continue to be committed to open publishing, normatively and in the applied sense. Together with Salla and the Editorial Team, we carried out the work in migrating the journal to its current editorial system, the Open Journal System (OJS) in 2014. OJS is one of the main open-source publishing infrastructures in the world. For S&TS, using the OJS has contributed actively to the improvement of the editing process both technically and substantively.

For the future, our core values hence include:

- To maintain rigorously the Impact Factor, coupled with a strong commitment to Open Access, no Article Processing Charges, and open publishing via the open-source software infrastructure OJS. Our Impact Factor has risen several years in a row, and to our knowledge, we are among the very few STS journals that have both Open Access and an Impact Factor. The S&TS journal is also included in the DOAJ (Directory of Open Access Journals), a community-curated list of open-access journals. Our Open Access is possible because of the financial support of the European Association for the Study of Science and Technology (EASST) and the Finnish Science Foundation via the Finnish Society for Science and Technology Studies.
- To uphold the rigor of the field of STS by an editorial commitment to advancing scholarly studies of science and technology. This means not only including studies that apply STS concepts, as is increasingly the case in many interdisciplinary fields, but to expect our papers to contribute to the advancement of STS debates and knowledge.
- To maintain our commitment to a collegiate and supportive approach. The Editorial Team sees it as the journal's duty to contribute to preserving and improving both the high quality of research in the field of STS and the inclusiveness and healthy research environment of the field. This also means that all our editors

are invested in an egalitarian 'flat hierarchy' concerning publication decisions. This means they have the space to influence the editorial process and are expected to be able to work both in the team and independently.

As 2022 is coming to a close, I want to highlight that the next year marks the journal's 35th anniversary. This makes S&TS among the oldest published journals in this field. The first issue of S&TS, then named *Science Studies*, was published in 1988. *Science as Culture* was launched in 1987, *Science, Technology and Society* in 1996, and the *Nordic Journal of Science and Technology Studies* in 2013. S&TS was preceded by *Science, Technology, & Human Values* (1967) and *Social Studies of Science* (1970).

In October, the STS community heard the very sad news that Bruno Latour passed away at the age of 75. Obituaries have been published all over the world. His contributions to the field of STS were essential and he will be sorely missed.

Kind regards
Antti Silvast

Testing Emergent Technologies in the Arctic: How Attention to Place Contributes to Visions of Autonomous Vehicles

Marianne Ryghaug

Department of Interdisciplinary Studies of Culture, Norwegian University of Science and Technology (NTNU), Norway/marianne.ryghaug@ntnu.no

Bård T. Haugland

Department of Interdisciplinary Studies of Culture, Norwegian University of Science and Technology (NTNU), Norway

Roger A. Søraa

Department of Interdisciplinary Studies of Culture, Norwegian University of Science and Technology (NTNU), Norway

Tomas M. Skjølsvold

Department of Interdisciplinary Studies of Culture, Norwegian University of Science and Technology (NTNU), Norway

Abstract

There are great expectations around the future of autonomous vehicles (AVs). Such visions often picture vehicles that work *everywhere* without human interference. In this article we use empirical data from a pilot project taking place in the Norwegian Arctic to explore the place-specificity of such technologies. The case study is used to demonstrate how new configurations of emergent technologies are shaped by the places where the trial unfolds; and how insights produced through working on and with this site contribute to changing visions of AV technologies into questioning issues of transferability and scalability. In this way, the paper contributes to discussions of how pilot projects and testing of emergent technologies in the real world relates to the re-configuring of visions and expectations. The paper highlights how emerging technologies might transform societies, infrastructures and vehicles towards more computerized configurations in ways that are not anticipated or discussed in public and therefore seldom governed.

Keywords: automated vehicles, testing; place; Arctic; scalability, connectivity, Intelligent Transport System

Introduction: Transport systems, infrastructure and the impacts of autonomous vehicles

The globalization of food markets and associated food-chains has resulted in vast demand for long-distance transport of livestock, meat and fish (Anderson et al., 2018). These transportation activities depend on large technical systems such as road infrastructures and large fleets of vehicles to transport goods from production sites to markets. The transport of salmon from fish farms on the coast of northern Norway to high-end Asian markets is a good example. Large volumes of fish are brought to shore by boat, transported on trucks through Norwegian landscapes with rough roads and challenging driving conditions, before reaching Finnish airports where the cargo is flown to Japan.

Today, policy makers and goods transport actors are working to transform transportation practices to improve environmental and climatic performance and to increase profit margins. Amongst these actors, there are strong visions and expectations for the role of emerging technologies in such processes of change (Mladenović et al., 2019). Technologies like ´connected´ or ´autonomous vehicles´ (AVs) and Intelligent Transport Systems (ITS) are examples of innovations that many believe are likely to transform the transport sector in the near future (e.g. Sovacool et al., 2019; Stilgoe, 2018; Mutter, 2019), providing seemingly universal solutions to diverse challenges associated with transportation. While actors such as the European Commission claim that we are only a few years away from a reality where autonomous vehicles are the norm (EC, 2017), and industrialists have argued that it is only a matter of months until the most important technological challenges facing full AV implementation are solved (Duarte and Ratti, 2018; Koetsier, 2020), the potential social, economic, environmental and practical implications of autonomy, automation and digitalization in the transport sector are contested (Haugland and Skjølsvold, 2020).

Innovation within this field is often conducted through demonstration projects, test beds, field trials and pilot projects. Such projects[1] are at the centre of our approach in this study, as we zoom in on one site where intelligent automotive technologies are being developed and tested under arctic conditions in the north of Norway. Through this, we seek to gain new insights about how visions of intelligent automotive transport futures are enacted, but our ambitions are also broader. Pilot projects do not only discretely test new technologies. These sites are places where 'visioneering' is transformed into materiality (Engels et al., 2017), potential sites of 'anticipatory governance' (Guston, 2014) and milieus where the ethics of invention (Jasanoff, 2016) are shaped in rapidly evolving fields. These sites constitute important geographical locations for studying emerging technologies, as well as the shaping of knowledge claims and visions about future societies.

Actors involved in the case we study, mobilize the characteristics of the place to lend credibility to the tested technologies. Compared to Silicon Valley and other sites associated with artificial intelligence and driverless vehicles, northern Norway provides an altogether different set of challenges for AVs. Hence, if successful, the test site might become somewhat of what Gieryn (2018) calls a truth-spot for AVs. Interesting questions for us are how studying pilot projects may contribute to our understanding of current innovation practices and how pilots relate to visions of technology introduction, scalability, and place? By investigating such questions, we bring to the surface otherwise marginalized debates and alternative visions of technological pathways for automation and digitalisation of large technical transport systems.

Studying emergent technologies: the role of experimentation and pilots

In recent years, a scholarly interest in a ´sociology of experimentation´ has boomed. Research in this field studies societal experimentation and testing in real-world social environments (van de Poel et al., 2017; Marres, 2019). Studying experiments beyond the laboratory have been flagged as central, as they clearly constitute places where new forms of governance, economy and subjectivity are invented (Engels et al., 2019; Van de Poel et al., 2017; Marres and Stark, 2020; Gross and Hoffmann-Riem, 2005). While experimental devel-

opment of new socio-technical configurations provides opportunities for experimenting both with new socio-political orders and technology (Marres et al., 2018; Marres, 2019), some scholars have noted that it is quite rare that pilot projects do more than test technologies under standard societal conditions (Schot and Steinmueller, 2018). However, AV performing tests in public streets have recently prompted STS-scholars to raise new questions concerning the relationship between such innovation activities and the social. Many of these real-world intelligent vehicle tests explicitly focus on social phenomena and thereby, do not comply with the "social deficit" associated with testing reminiscent of older STS accounts of testing (Marres, 2019; Pinch, 1993). Actors conducting AV testing, for instance, often highlight that improving the understanding of vehicle-pedestrian interaction is a key element of the test (Haugland and Skjølsvold, 2020; Marres, 2019).

Thus, while these tests operate with rather narrow understandings of sociality one cannot claim that they are void of social concern. Our point, however, is that we should not only see these occasions as attempts of transferring the tests from laboratories to social environments such as public streets. We should also explore the relations *between* real world sites of testing, their relations to the environments they are part of, and focus on how such a move can illuminate social change, more broadly.

Many assumptions about the future of AVs and ITS are based on what we might call technological hype produced by media actors, policy makers, consultants, and companies promoting AVs (Stilgoe, 2020). Hype, however, does not mean insignificance. STS scholars have illustrated the 'constitutive' nature of promises, e.g. within literature on technology expectations. Visions, expectations, and technological hype are not only predictions of the future, they also produce futures (Van Lente and Rip, 1998; Borup et al., 2006; Skjølsvold, 2014; Pollock and Williams, 2010; Stilgoe, 2020). This makes such predictions interesting research objects. It also points to the importance of scrutinizing who predicts what and why and the importance of studying emerging technologies at an early stage, "before they become just another fact of life" (Stilgoe, 2020: 5), both in the quest to govern technologies, and to be able to understand the transformative power of technology in order to be able to resist, stop, slow down or redirect technological trajectories (Jasanoff, 2016).

Pilot projects, trials and experiments are important sites, where the discursive elements of expectations are made concrete and material (Engels et al., 2017). They represent an approach to innovation which signals ambitions of making technologies that function when implemented in society (Skjølsvold et al., 2020; Ryghaug and Skjølsvold 2021b) and tend to have a dual set of ambitions: On the one hand, they seek distinct and localized lessons. On the other hand, there is often an outspoken ambition of scaling up and to apply what has been tested in one setting universally (Naber et al., 2016; Ryghaug et al., 2019; Engels et al., 2019). Classic STS-accounts note how technologies become shaped by their social surroundings (e.g. MacKenzie and Wajcman, 1999; Williams and Edge, 1996) or through the work of relevant social groups (Pinch and Bijker, 1984), which is echoed in accounts of how pilot projects for technologies are often shaped by a combination of local concerns (Skjølsvold and Ryghaug, 2015; Ryghaug and Skjølsvold, 2021a) and wider repertoires of interests, understandings and competence circulating through international networks (Bulkeley et al., 2014; Engels et al., 2019). In this paper, we build on such perspectives from STS, and aim to contribute to discussions of how and in what ways pilot projects, experiments and testing of emergent technologies in the real world relates to the re-configuring of visions and expectations.

Social implications of AVs and different levels of automation

In the discussion above, we mainly engage with the enactment and materialization of expectations in concrete trials. However, there are also strong visions for how AVs will affect life on the roads more broadly. This is visible in the increasing media- and scholarly interest in AVs (Stilgoe, 2020; Duarte and Ratti, 2018; Shladover, 2018; Sperling et al., 2018), in part shaped by vehicle development, but also wider transport and mobil-

ity developments, e.g. Mobility-as-a-Service, traffic management, and IT applications for transport in smart cities. Important interactions have also been established between automation and innovations in modes of ownership and fuels (Hopkins and Schwanen, 2018). Today, new cars can automate tasks that until recently have had to be performed by the driver through technologies such as automated and adaptive cruise control and lane assistance systems. Fully automated – or popularly called "driverless" or "self-driving" cars – have arguably gone from being interpreted as highly unlikely, to becoming what many think are an inevitable part of our near future, soon to be found driving down every street (Sperling et al., 2018; Stilgoe, 2017; 2020).

The hype and expectations both in terms of technology development and how AVs might change societies, have led social scientists and others to critically engage with such visions, and to reflectively probe potential societal implications of AVs. Examples include questioning whether AVs will lead to safer environments for pedestrians (Combs, 2019), increased vehicle miles travelled, (negatively) impact public transport, reduce the overall number of vehicles and parking spaces (Duarte and Ratti, 2018; Soteropoulos et al., 2019) and if AVs would demand more or less road infrastructure; contribute to increasing urban sprawl, or rather attract more residents to city centres if they are freed from congestion and pollution. Loss of social safety and privacy have also been identified as potential social implications (Blyth, 2019). AVs may potentially impact many aspects of our lives.

Through reviewing the literature on the effects of automated driving, Milakis et al., (2017, 2018) divided the implications of AVs into: (i) day-to-day usage impacts (travel costs and choices), (ii) impacts to long-term decisions (vehicle ownership, sharing, residence choice, land use and infrastructure), and (iii) overall societal impacts (energy, environment, equity and health). Others have discussed the potential implications of AVs by simplifying them into extreme future transportation systems scenarios, such as the "Heaven" and a "Hell" scenario described by Sperling et al. (2018) where the Heaven scenario focuses on effects such as improved safety, accessibility and equity among travellers and a Hell scenario characterized by further entrenchment of private vehicle ownership and negative effects such as increased vehicle use, suburban sprawl, fossil fuel usage and less use of public transit and active travel modes.

While all the above tend to be discussed as impacts, or effects of technology, they are in reality parts of the societal visions and expectations for how AVs will change the world. Hence, a move from the study of pure discourse to the study of materialization, is also a move towards studying consequences and implications in the making. Not long ago, laboratory tests were the norm for AV development (Leonardi, 2010), but a rapid rise in real-world testing to learn and to proceed to higher levels of intelligence has ensued (Stilgoe, 2017). To us, this also entails the making of sites that on the one hand tests technologies and social aspects, but which on the other hand also contributes to the production of new visions and expectations. One practical consequence of the move from laboratory to street, is that the different levels of automation have become omnipresent in discussions of AVs. These levels serve as a solidification and standardization of certain technology expectations. As one can read on the website of the US National Highway Traffic Safety Administration (NHTSA, 2020):

> Fully autonomous cars and trucks that drive us instead of us driving them will become a reality. These self-driving vehicles ultimately will integrate onto U.S. roadways by progressing through six levels of driver assistance technology advancements in the coming years. This includes everything from no automation (where a fully engaged driver is required at all times), to full autonomy (where an automated vehicle operates independently, without a human driver).

Here, the NHTSA refers to the J3016 Levels of Driving Automation standard developed by the Society of Automation Engineers (SAE). This standard divide driving automation into six distinct levels, ranging from level 0 (no automation) to level 5 (full automation).[2] At level 5, automated features allow the vehicle to "drive everywhere in all conditions" (SAE International, 2016). The SAE standard, originally developed to elucidate the challenge of automating the driving task (Stayton and

Stilgoe, 2020), has come to define level 5 automation as the singular goal of transport automation (Ganesh, 2020; Hopkins and Schwanen, 2021). This suggests that, at some unspecified point in the future, self-driving vehicles will be capable of operating within any environment without needing support from 'smart' infrastructures. Such a future is promoted by Tesla, as well as other AV proponents (Stilgoe, 2018). For a vehicle to operate without concern for its specific environment, however, it is crucial that the technology learns how to drive in different environments.

Street trials with AVs on public roads are said to provide the variety necessary for learning vehicles how to drive under every single circumstance. Testing under real-life conditions is important in order to benefit from machine learning. Such testing allows the technology to learn from unexpected situations that would be difficult to simulate (Stilgoe, 2017; Marres, 2019). If most road automation trials are about displacing innovation activity and experiments from the laboratory to the real world to do experimental innovation (Laurent and Tironi, 2015) in line with the logic of data-intensive machine learning which requires learning from as many and varied situations as possible, it should be important that these trials are not always "displaced" to very similar environments. On this basis one should expect that real-world AV trials were conducted in very different environments (arctic, tropical, etc.) with different characteristics (urban, rural, road topography and geometries) and under different conditions (weather, traffic, pedestrians etc.) in order to ensure successful operation in all possible contexts.[3]

The early history of AVs had prominent plans for integrating car innovation and road infrastructure (Stilgoe, 2018). From the 1950s until quite recently it was assumed that, in order to get self-driving cars to operate well, they would require communication with equally intelligent highways and road infrastructures (Wetmore, 2003). However, in the last couple of years, innovations experimenting with intelligent road infrastructures such as responsive traffic light systems or concepts of fleet steering and truck platooning[4] (like we focus on in this article) have not been given equal weight. Current field tests focus mainly on cars and associated automotive technologies driven by platform companies such as Google, Uber and Tesla (Stilgoe, 2018). Early trials were also typically done in remote and confined spaces, such as the Mojave Desert and Nevada Desert, although AV trials in cities and urban areas have become more common (Hopkins and Schwanen, 2018; Marres, 2019). Such urban AV trials have, however, typically been configured in specific parts of the city, such as new residential and/or commercial developments (e.g. Greenwich Peninsula in UK) and sites characterized with lower traffic flows and less complex road configurations (Hopkins and Schwanen, 2018; Haugland and Skjølsvold, 2020).

Many of the test sites that have already been studied in Europe also have been heavily prepared and facilitated to curtail interaction between intelligent vehicles and other road users (Marres, 2019; Haugland and Skjølsvold, 2020). Thus, there is clearly an ´unevenness of laboratorization´ going on (Hodson and Marvin, 2009).[5] This deserves more attention when trying to anticipate the futures that could surround self-driving cars, which futures such cars might enable, what futures those advocating such technologies might push for, and likewise what future transport scenarios become disfavoured by increased focus on AVs (Haugland, 2020).

Testing emergent innovations in different environments and under particular hash conditions is obviously important for both technical and non-technical reasons, as we also need to empirically examine different ways in which road trials of intelligent automotive technology contribute to the production of new visions and expectations and configure relations between society and innovation in new ways. Thus, in this article we have chosen to focus on a case study representing a test site for intelligent transport technologies that clearly stands out from typical urban test sites in warmer climates: a test site along a long road stretch in a remote area north of the Arctic Circle.

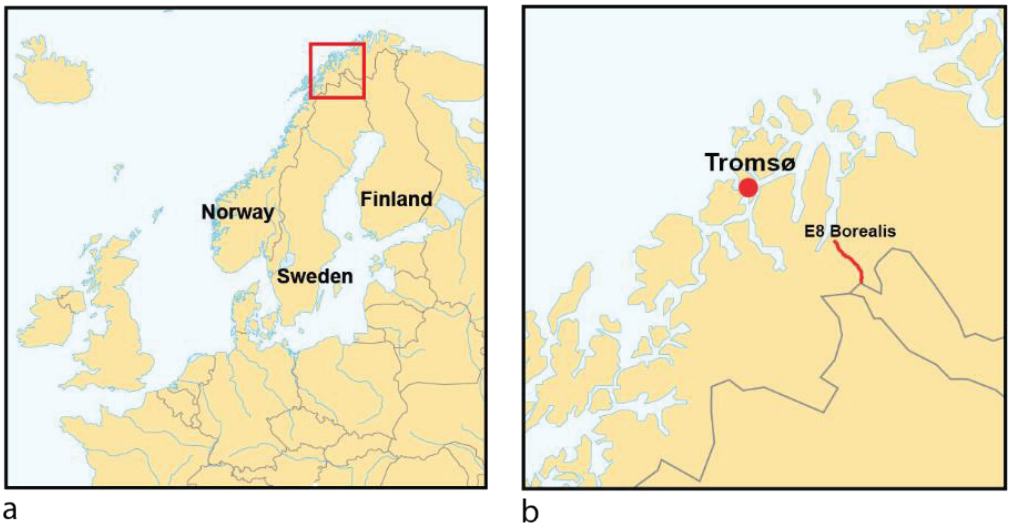
Infrastructures and other elements of the built environment in polar regions have traditionally been given little attention in the literature (Schweitzer et al., 2017). The particular "laboratory" reputation of the Arctic, as a technology-intensive locality that renders tensions between human and technology in these settings unavoid-

able (Usenyuk et al., 2016) should however be acknowledged. For instance, it has been shown that continuous modification, tuning and even redesign of technology *in situ* have been necessary for humans and machines to function in these extreme settings, often leading to new design principles (Usenyuk et al., 2016) and new insights highlighting the contextual relevance of design, innovation and policy implications of developing technologies for extreme and uncertain environments (Usenyuk et al., 2020). Others have focused on the important role of the state in building infrastructure in the polar regions (Schweitzer et al., 2017).

In the following, the Borealis test site on European route E8 will be analysed – a case that does not focus primarily on vehicle technologies but may represent a particularly hard case when it comes to testing intelligent transport technologies and road infrastructures in rural and remote settings. By zooming in on one particular field site where intelligent automotive technologies are piloted we are able to articulate more precisely not only *what* is being tested and how this relates to the place, but also how these innovation practises can question the whole narrative of "placelessness" that characterize the current vision of AVs (Hopkins and Schwanen, 2021). Thus, the case study is used to demonstrate how new configurations of emergent technologies are shaped by the places where the trial unfolds; and how insights produced through working on and with this site contribute to changing visions of AV technologies into questioning issues of transferability and scalability. The pilot test under scrutiny in this article, also contributes to illuminating some unforeseen glitches in the technology at hand. Consequently, and indirectly, it also points towards questioning the transferability of knowledge gained through other AV test sites and field trials, acting as truth spots (Gieryn, 2006) for claims about AVs.

The case study: Developing and testing intelligent transport technologies in the Arctic

The Borealis project, chosen as case study for this paper, has been described as both a research and development programme and a national test laboratory for new technology covering a 40-kilometre-long stretch of the European route E8 in Skibotndalen, 69 degrees north in the Arctic reaches of northern Norway.[6] E8, which stretches 1,410 kilometres from Tromsø, Norway to Turku, Finland and goes through Skibotn to Kilpisjärvi is one of five Norwegian road sections selected as pilots for the development and testing of ITS solutions in Norway. While the pilot project in Norway from Skibotn to Kilpisjärvi has been named Borealis, it also has a Finnish counterpart project running from Kilpisjärvi to Kolari, named Aurora. However, in this article we will mainly focus on the



Figures1a & 1b. 1a Norway's placement in Northern Europe. 1b The location of the Borealis project (© Kartverket under a CC BY 4.0 license, modified by the authors).

Norwegian part of the project run by Norwegian Public Roads Administration (NPRA).

The paper is based on observations and qualitative interviews of key actors of the Borealis project, including a site visit by three of the authors to the Borealis test site during the first on-site testing in the winter of 2019. During the site visit we also took part in several meetings between different actors operating in the project, mostly consisting of (chief) engineers and leaders responsible for conducting the testing on behalf of NPRA, as well as representatives from several technology developers engaged in different sub-projects of Borealis. In these meetings and during the on-site activities, status, progress, challenges, and ways forward were topics that were discussed. In addition to these participations and field-observations, our main empirical data material consists of eight semi-structured interviews with key participants of the Borealis pilot project. These actors consisted of senior engineers, planners, test-leaders, and test-conductors, companies involved, and local government.[7] The interviews, ranging from 25–85 minutes in length, were conducted by the authors and subsequently transcribed verbatim. In addition, written sources, and updates on the project in meetings and seminars and more informal briefs given to us by the NPRA as part of a larger project in which Borealis was picked as one of the cases, contributed with additional insight about the pilot activities after the visit. These additional sources, as well as our interview data, provide us with a rich material for thoroughly analysing the pilot project and its operation. In the next sections we will zoom in on the innovation strategies that have governed the pilot project and how place specific concerns are raised through the implementation of this project setting out to test intelligent transport technologies in the Arctic.

Innovation strategies of the Borealis pilot activity and test site

The Borealis test site was organized by the NPRA to test and develop Intelligent Transport Systems (ITS). ITS is an umbrella term that covers technology and computer systems in the transport sector. In an ITS system communication can flow from one vehicle to another, from the vehicle to the roadway or from the roadway to the vehicle. Examples of such technologies are real-time information about weather, road surface conditions and traffic accidents, automatic scanning of the vehicle's brakes, and warnings of wildlife or other obstacles on the roadway. According to the NPRA, ITS combines technology and computer systems with a dual goal: For road users and transport operators, ITS can make the drive safer, more efficient, and more environmentally friendly. For those who operate and maintain the road, ITS can make it easier to implement the right measures at the right time.

In the start of the project, the NPRA enlisted the help of the interest group ITS Norway to host a workshop and an idea-competition. Subsequently, the NPRA received 36 ideas of which they chose 16 concepts they deemed interesting, before ultimately selecting 8 projects for funding. Table 1 gives an overview of most prominent technologies that were tested within the Borealis project and related concepts discussed by test side organizers and participants during our visit to the test site and in interviews.

The NPRA organized the innovation process so that these firms could implement their technologies alongside the road chosen to be a test site. However, before doing so, they needed to know the real problems in this north region. They therefore organized dialogue meetings and interviewed different stakeholders and users of the roads such as local industry (customs, fishing industry, businesses) and those using the road (like truck and bus drivers) to identify their needs and what their problems were. These insights were fed back to technology developers in and after these meetings.

According to the NPRA, this road was selected for its socio-economic significance, especially with reference to the road's high importance in exporting fish from fish farms by the Atlantic Ocean to European and Japanese markets. Thus, its importance as a corridor for transporting fish from the Norwegian coast to an airport in Finland cannot be overstated. Time is a complicating factor in this regard. The fish should be at the Finnish airport no more than 18 hours after being loaded onto the truck. Driving the stretch takes 16 hours, giving the truckers no more than two

Table 1. The Borealis project: concepts and technologies that were tested

Technology	Description
Truck platooning	Technology for linking of two or more trucks in a convoy. Combining communications technology and advanced driving support systems allows the vehicles to maintain a pre-determined distance, reducing air drag and thus also fuel consumption.
LIDAR technology	LIDAR technology uses a pulsed laser to determine the distance from the LIDAR and to an object. In the Borealis project, LIDAR was mounted on poles, to judge the technology's merit in identifying trucks coming to a stop in slippery uphill slopes.
Parking sensors	A set of parking sensors were dug into a stretch of the road. The sensors use the magnetic field generated by the mass of a passing vehicle to identify the vehicle type. Like the LIDAR, these sensors may also identify vehicles coming to a stop.
Smart signs	Digital signs placed along the road. The text on the signs is editable, and the signs are connected to communications infrastructure. These signs are capable of displaying alerts received from other infrastructure, such as the aforementioned LIDAR, as well as information about road and weather conditions.
I2V and V2V communications	Different solutions for facilitating communication between infrastructure and vehicles (I2V) or between vehicles (V2V). These technologies include both software for processing and distributing alerts and hardware for passing these alerts from infrastructure to vehicles or between vehicles.
Distributed acoustic sensing (DAS)	Acoustic cables cast into the road. When a vehicle passes over the cables, noise is introduced to the signal passing through the cables. This signal noise might then be used to identify the kind of vehicle passing or follow the vehicle's trajectory.
Roadside cameras	Combining Bluetooth, wi-fi, and cameras, these cameras were intended to contribute to the estimation of travel-time, counting vehicles, and give the proper authorities an overview of an unforeseen situation, e.g., an accident.
Clocking-app	An app surveying and suggesting adjustments in vehicle speed to avoid waiting time and traffic jams. As Northern Norway has multiple locations with narrow tunnels and bridges, this app would allow these sites to be traversed in a problem-free manner by avoiding oncoming traffic at these narrow sites.
Travel-time estimation	A set of algorithms combining available data, for example weather data, road conditions, previously registered travel times, etc., to predict travel-time
Relevant Concepts	
Smart roads	Umbrella term for ITS technologies, for instance technologies mentioned here are technologies that detect vehicles moving upwards, and then different alerts are set if the speed is declining or it stops completely. It is also sent back to the Road Traffic Centre, possibly also directly to the smart signs.
Communications infrastructure	A general communications infrastructure was established along the road. This would enable the above technologies to communicate with each other or with, for example, the Road Traffic Centre or road maintenance providers in the case of unforeseen events.
C-ITS	Cooperative intelligent transport systems and vehicle-to-vehicle and vehicle-to-infrastructure communication which is seen as one step towards more autonomous or automatic vehicles, and cooperative awareness messaging

hours to spare for unforeseen events with potentially big economic losses in the case of delay. The two hours margin could easily be "eaten up" by for instance, a bit of trouble, or a slow loading process at Skjervøy (the largest fish landing in the area), and weather conditions. With icy roads an additional 30 minutes fitting chains on the tires, and another 15 minutes removing them, leaving only a 15 minutes margin for travelling the whole road stretch. Thus, getting stuck on the notoriously difficult to drive road stretch of the E8 road in Skibotn, would result in serious trouble. The cargo would risk decay in both taste and value, the plane would be lost, the sushi would lose its 'rigor mortis' and the Japanese would not be willing to pay a premium price for the fish. The alternative would be freezing the fish, reducing the market value by seventy-five percent. Thus, improving solutions for when to slaughter the fish, when to send the trailers and on to which route, may have significant importance for fish farmers and the local economy.

Thus, the Borealis test site was chosen both because of its important role in the local community and for its difficult test-site environment, as the E8 road had very demanding winter conditions and a large share of heavy vehicles and trucks trafficking the road. The fact that the road stretch itself was nick-named "the road from Hell" and considered a terribly difficult stretch of road to drive as the weather in the area often caused chaotic situations: trailers sliding off the road, trucks with difficulties getting up the long and quite steep hill of the Skibotn valley, the road getting blocked for hours by trucks in need of vehicle assistance, was seen as an advantage for experimenting. From the point of view of the test organizers, it was considered a particularly tricky place for demonstrating intelligent automotive technologies. As one of the NPRA project leaders noted: "If you think of self-driving vehicles in terms of the school system, then Arizona is kindergarten", the flat and confined trials of the Netherlands as elementary school, driving on European roads as secondary school and driving in Finland, as high school as you will have to deal with snow. However; "if you manage to drive the Skibotn valley down, then you are at the PhD level". Thus, by operating tests in this harsh environment the NPRA was deliberately striving to test technologies under difficult circumstances. The harsh winter conditions and demanding road infrastructure were considered as advantages and something Norwegian research communities could capitalize on. According to one of the NPRA engineers,

> the special challenges we face in relation to positioning, communication, and that they do winter testing in Norway – If you manage to attract foreign companies to do so, then we have succeeded because we get technologies that are more robust and more beneficial. Our goal is not that Norwegian industry will make the cars (…) but they have to function in Norway.

Here, we clearly see how field scientists, or in this case, the engineers strived to justify their choice of the specific place and the research site as being analytically strategic (Gieryn 2006) in that it uniquely displayed certain forms of process with great interests for technological advances.

The NPRA saw the trials as a way of showcasing the difficult circumstances that intelligent automobile technologies must be able to operate under, highlighting that the harsh winter conditions are an opportunity for Norwegian innovators, who could attract foreign companies by using the circumstances to develop more robust intelligent automotive technologies. However, this argument was not always easy to convey to other actors working on automation and ICT solutions. The argument that, for the technology to work, it had to handle all kinds of situations, was often met with pointing to the peculiarities of the place: that it is not like that everywhere. However, for the NPRA engineers the question should be the opposite: how many days of snow the European economy would be able to handle if everyone was driving their own AVs. Thus, we see that the peculiarities of the site played a double role; both as a credibility-enhancing geography (Gieryn, 2006) that was required in order to develop reliable technological solutions that could work 'anywhere', but at the same time contested by some of the IT and automation industry experts because Norway was not exactly 'anywhere' – hinting towards the fact that other sites were perceived as more naturalized 'anywheres' or 'placeless places' that could enhance the credibility of scientific claims.

According to the engineers involved in making the smart signs and communications solutions, the Borealis project was exciting because it would give answers on how far one could get regarding self-driving vehicles by using existing technology and what new technologies were needed to make it work. Following this line of reasoning, it was important for the Borealis project to be open about *the limitations* of the technology and what possibly could be tested in the trial. However, this kind of critical remarks about limitations of the technology were sometimes sanctioned by the IT and automation industry actors in the project, who were afraid that such remarks might harm them.

Thus far, this article has focused on conditions that are important for the NPRA to consider when setting up the trial and some of the rationale for creating a test site under such difficult conditions. The position taken by the NPRA seemed to strive towards more robust technology, but also creating

more socially robust knowledge (Nowotny 2003; Stilgoe 2017) about self-driving and AV technology; to manage technological expectations so that they better aligned with societal needs and urgent challenges to be solved.

Another result of this was that the NPRA deliberately worked to keep the trial relatively small and to minimize the complexity by reducing the number of technologies being tested and to manage expectations and visioneering. While some of the industrial partners from the IT and automation industry wanted to "conquer the world" by designing a comprehensive digital platform capable of handling all the data gathered and processed in the Borealis project, the NPRA worked to keep the test site focused on solving local pertinent issues. Thus, instead of buying into the bold visions around big data and machine learning associated with AVs they decided to focus on small use-cases concerning how to use existing intelligent automotive technologies to improve the building and use of tunnels. This they saw as something that could potentially add value both for NPRA and the local community as it would keep them from having to build new tunnels.

Thus, technological solutions used to showcase the project, such as platooning, were not really something that was tested out, although these concepts featured prominently in the public accounts of the projects. This, however, did not mean that platooning and self-driving vehicles did not play a role in the project. Platooning self-driving trucks and AVs seemed to play a role in allowing for particular ways of operating the innovation process. It was seen as crucial for branding the project and had created a lot of media attention and political support. This type of big and shiny visions was regarded as important tools to give NPRA engineers finances and leverage: as "building blocks needed to be able to work undisturbed". Thus, the innovation process was deliberately set up with the inherent duality: to, on the one hand upholding big shiny visions about self-driving cars and technological advances, on the other hand pushing for technological sobriety and realism within the project team in order to be able to push the development forwards realistically and stepwise.

Place-based challenges of the trial site

There were several issues concerning correct positioning of vehicles in the area. Some of these problems related directly to the positioning of the site near to, and at, the border to Finland. For the technology to work properly it was important that technology developers could use correct maps. However, at the borderline between Norway and Finland, an unanticipated problem was identified: Gaps in the Global Positioning System (GPS) maps between the national borders. As one of the engineers explained:

> One thing is to find out where you are going, but it must be connected to a map. And that link between the position and the map is made a little bit differently in each country. We have a small gap of ten centimetres where there really is nothing! On Norwegian maps there is nothing, and on Finnish maps there is nothing. But, of course, there is something there! But it is these systems that make it wrong.

Although the digital maps showed a non-existent area, this area was of course very real in the physical world. Thus, the engineers working in the project clearly found challenges that needed to be addressed for the intelligent automotive technologies to function accurately in this type of environment.

Tectonic plate movement represents another challenge to the accuracy of GPS. As explained by one of the engineers: The GPS and the associated Global Navigation Satellite System (GNSS) refer to a fixed position on the *globe*, not the Earth's *surface*. As the tectonic plates imperceptibly drift over time, this means GPS positioning will not match the maps of the actual terrain. In one instance, the NPRA found that a GPS map produced in 1989 consistently positioned all the marked road signs incorrectly, by approximately 0.5 metres. This was after some time discovered to result from continental plates drifting 46 centimetres north-east since the maps were made. This made the engineers question the prospect of using GPS for AV positioning, as this would require a positioning accuracy of at least ten centimetres for the vehicles to drive safely. As tectonic plate activity influences GPS positioning, this would require maps to be updated regularly.

The above challenges would be present across the globe. The upper reaches of the northern hemisphere, however, came with their own set of problems in relation to GPS. First, GPS relies upon a set of four satellites for accurate positioning. As most of the satellites are located to the south-west of Northern Norway, geometry dictates that accurate positioning is harder to achieve as the angles from the satellites and onto the globe flatten. To achieve accurate positioning in these areas, one would need satellites in hyperelliptic orbits.

The prospect of accuracy was challenged further by the quite frequent natural phenomenon of Northern Lights (Aurora Borealis in Latin) made driving by GPS signals difficult as well, as one of the interviewees explained:

> Everyone says the Northern Lights are terrific! Let's send all the Asian tourists with autonomous vehicles to look at the Northern Lights! But what does the Northern Lights do to satellite navigation? It is a living hell for GPS signals, it destroys them! So, then the uncertainty increases, and we start driving the tourists into the ditches (…)

The amount of Northern Lights in the area is considerable, thus this would lead to a huge challenge. But also, the latitudes would reinforce this as the challenges related to GPS positioning mentioned above. Altogether this makes it difficult to have a good positioning fixed in the high north.

We find this to be an interesting juxtaposition concerning Norway's dual identity as one of the foremost countries both on natural phenomena as the Northern Lights, and as a frontrunner in emerging AV technology testing. As noted above, interviewees deemed navigating AVs in the north as problematic due to lacking GPS accuracy. Within positioning, one usually went by "three meters, 50% of the time" as the standard rule, which was considered fine for landscape measurements. A similar standard of levels of accuracy would however be fatal for positioning AVs that would require much more accurate positioning to be considered safe. This kind of inaccuracy of using GPS also led to discussions about other types of inaccuracies that the engineers struggled with within the automation projects.

When talking about intelligent automotive technologies, one of the things that puzzled the engineers in the project was the accuracy and myriad of signals that would be going to and from, and in between vehicles and the road infrastructure. They felt very unsure about the prospect of every problem being solved by local sensors, as was often presented as the answer when raising questions about the inaccuracy of GPS positioning. They would not get any good answers from industry partners about what levels of accuracy they could expect regarding sensor technology. According to the NPRA engineers, local sensors might also fail. Without a proper distribution of signal wavelengths when using LIDAR technology for local positioning, an automated vehicle might experience 'sensor blackout' when interpreting an outgoing signal from another car as a returning signal from its own sensor. This points to both the importance and limitations of back-up systems and made the ITS developers aware of the importance of making technologies that communicate across different domains, and the importance of adapting and/or upgrading infrastructure in order for self-driving vehicles to work.

Furthermore, there were the more mundane problems relating to winter conditions, such as snow and ice, which influenced the effect of sensors tested in the Borealis project. Snow on the sensors meant you could no longer get much information from them. Therefore, building infrastructure that also could help with the positioning was important. Sensors are well-known to have serious problems during difficult weather conditions, and the project therefore focused on infrastructure (for instance in signposts) along roads and in tunnels that could help with the positioning of AVs.

In fact, massive investments and technologies were built into the roads and infrastructures in order for the technologies tested at this site to work. Examples were costly broadband cables in the ground that ensured communication and connectivity on the test site and 120 sensors installed in the asphalt to detect and send signals back to the traffic centre if vehicles stopped in the hills. Consequently, the vision of AVs being able to navigate the world's complexity using

only its sensors and processors seems far-fetched. In addition to the webs of social and technical connectivity so-called autonomous vehicles rely upon in order to work (Stilgoe, 2017), their implementation also appears to demand expensive and comprehensive infrastructure upgrades which are unlikely to be undertaken indiscriminately across the nation's road network, as illustrated well by this case.

Cross national cooperation and different innovation cultures

The agreement between Finland and Norway which was initiated on a high departmental level to develop innovations and think anew regarding freight transport in this particular pilot, also seemed to have some significance for the focus of the project. The Finnish part was carried out by commercial actors and therefore governed by more price-concerned thinking. The Norwegian side of the project was on the other hand, run by a public body, the NPRA with a different innovation approach to this kind of project. The fact that Norway still has a lot of engineers employed in the Public Roads Administration, compared to other European countries that have replaced many of their transport engineers and with procurement officers also play a role and had consequences for the division of labour between the two test sites.

The technologies tested also had national imprints that aligned with cultural standards around safety and risks. For instance, communicating directly to the driver by mobile phone was unheard of in the Norwegian context, in contrast to the Finnish. This was, according to the NPRA engineers, rooted in cultural differences: In Norway where traffic security was higher up on the agenda than in other countries, engineering strategies were accordingly risk averse. This also impacted who were regarded as important players to cooperate with. If you designed systems that communicated directly to the car (for instance telling the car to slow down because of an incident on the road ahead) consequently car manufacturers were considered more important. This was one reason that Norwegian Borealis actors considered cooperation with car or truck manufacturers more important than on the Finnish, Aurora, side of the border where one designed systems that communicated with the driver.

These examples clearly indicate how innovation activities were shaped by different engineering cultures and standards related to risks and safety. Technologies developed were shaped by their social surroundings, by local concerns and wider repertoires of interests, understandings and competences.

Conclusion

Most research on autonomous vehicles (AVs) focuses on vehicle technology or seek to anticipate the societal impacts of autonomous vehicles (Milakis et al., 2017). We have argued that to understand the direction of innovation, its potential consequences, and to reflect on the governance of these emerging technologies, we need to study the sites where they are currently developed and tested. Today, such innovation increasingly unfolds in real-world environments such as in test beds, street trials and pilot projects.

In this paper, we have used empirical data from one such pilot project situated in the Arctic to point out the place-specificity of testing and its consequences for visions of autonomous vehicles. The paper challenges some of the dominant visions about AVs and ITS, especially related to the often-overlooked networked aspects of these emerging technologies. Thus, the analysis reveals several challenges associated with digitalization of the transport sector and intelligent automotive systems which today are being ignored in most scholarly and public debates about self-driving cars and intelligent transport systems that deserves further attention.

First, we establish that the development and testing of intelligent automotive technologies is shaped by the place in ways that have serious consequences for the trustworthiness of current AV and ITS visions. The analysis of the Arctic test site demonstrates that testing of intelligent transport systems is tied to geographically specific needs and problems such as unreliable GPS signals and sensors related to inadequate maps and positioning systems, influence of Northern Lights, and difficult weather conditions. Thus, we point out several technological challenges usually

ignored when discussing the future of AVs. Visions of AVs as able to navigate the complexity of the world using only its sensors and processors is likely to be misleading. 'Placelessness' is however, an important feature of level 5 automation, with vehicles being expected to work without any human interference *everywhere* and under all conditions. It should be noted though, that the placelessness of AVs is about adaptability of technology to different circumstances, not about universalism as such. Referring to the Borealis site, an NPRA engineer argued that "if [a technology] works here, it will work anywhere". This statement suggests that the particularities of place can be taken into consideration, to the extent that a technology might work anywhere. Theoretically, placelessness is achievable. However, as suggested by our case study, it can only be approached by focussing intensely on the particularities of places. This also means that placelessness can only be imitated, and not truly achieved: in order to give the impression of placelessness across vast geographical swathes, a wide variety of place-specific factors have to be compensated for, whether through further technological development or additional, supporting infrastructure. This means that although placelessness might be possible in theory, it is both hard-won and improbable in practice, as both AV infrastructure and automated driving is highly place specific.

Second, this points to the fact that achieving full automation would require substantial work and sizable infrastructure investments, to such an extent that an indiscriminate implementation across the globe is entirely unlikely – one need only consider the substantial number (and complexity) of environmental factors which would have to be adjusted for at the Borealis site studied here, to make this point. If intelligent automotive technologies were made to work in difficult geographical areas such as the Arctic, heavy infrastructure developments would be required for fully autonomous vehicles to operate. As the case study reveals, AVs and ITS will have to rely on webs of social and technical connectivity and require vast investments in infrastructures and communication networks to function properly. This is also shown in other studies of autonomous vehicle street tests (Marres, 2019; Hopkins and Schwanen, 2018). However, this type of work and investments are often underplayed in the current narratives of AVs futures.

This points to a third challenge relating to the fact that not all places in the world would have the same capability to develop these required infrastructures. For instance, it is known that the agency of infrastructures and built environments in Arctic and polar regions have more easily been overlooked, partly because these regions are less densely populated (Schweitzer et al., 2017). New additions to communications networks and infrastructures in such areas may however have more profound social implications and maintenance may be more demanding in terms of financial and human resources, thus pointing to the need for governmental support in order to build and maintain these infrastructures. Thus, we see pertinent challenges related to both scalability and justice concerns.

This brings us to conclude concerning the question: what does the analysis reveal about what it means to transform societies, infrastructures and vehicles towards more computerized configurations and how to govern future intelligent automotive technologies? So far, visions and innovation activities in the field of AVs have been dominated by market-driven, expert-focused discourses that may limit the range of alternative AV futures (Hopkins and Schwanen, 2018). Street and roads testing have often been associated by a lack of engagement with societal contexts and concerns. They have been identified as having profound limits to responsiveness in innovation related to inadequacies in social learning, the inability to involve diverse sets of actors in testing and by a lack of accountability towards the populations enlisted in tests (Stilgoe, 2017; Marres, 2020).

In the Borealis pilot some efforts were made to include local stakeholders, such as the fishing industry and road users when developing the test site. The site was chosen because of its role as a socio-economically important stretch of road, as a key transport corridor for seafood export from Norway to European and Asian markets. Thus, while there are clear commercial interests and market logics at play, the fact that the pilot was governed by public sector actors and engineers

committed to solving real societal problems shaped the innovation activities in a different direction than most AV and ITS pilots.

The pilot was framed as a test site that may situate Norway as an innovation centre with regards to automation, in the same way as many other countries trying to attract business by showing off industrial strengths and ambitions in this area. However, those involved in the Borealis pilot strived to deconstruct more established truth spots (Gieryn, 2006) that previously had lent credibility to claims about AV and ITS futures in Europe and the US. For instance, when AVs have been tested on the roads in San Francisco or in London, these testing sites have been portrayed as both anywheres – placeless places with underlying patterns that can be found in most cities – while at the same time being field sites with strategically important qualities (such as Lombard street in San Francisco, being called "the crookedest street in the world" by one of Google's founders (Stilgoe, 2020: 21). By conducting, developing and testing AV and ITS innovations in remote areas in the Arctic, it also become evident that not all test sites easily represent such lab-like anywheres. This has made us ask, whether certain truth spots may be able to displace knowledge claims from other (less truthful) places?

'Truth-spots' have traditionally been related to scientific or other knowledge claims and not, as in our case, claims within engineering used to demonstrate that technologies work. However, interpreting a truth spot, as a proof of concept, thus more in line with a 'proof-spot', which lends credibility to knowledge claims about the workability of technologies, demonstrates the importance of developing knowledge and experiences outside more traditional (often urban) truth-spots. We argue that such proof-spots are crucial for understanding how vehicles, self-driving or not, may drive in more extreme environments. However, it remains to be seen if new proofs from the Arctic in the form of challenges to level 5 automation can displace well established visions and expectations of autonomous driving based on work elsewhere.

Through the in-depth exploration of a large scale intelligent automotive technology pilot project that seek to structure and stimulate innovation by piloting new sociotechnical arrangements *in situ*, undertaken here, our ambition was to explore how place contributes to and challenges the credibility of knowledge claims, visions and expectations about AVs and ITS more generally. We regard the evaluative capacities of this road test as modest, but not without merit, as we see evidence of new and interesting articulations of social, cultural and political aspects related to intelligent automotive technologies being developed by the involved actors. However, seeking to understand how testing and the social relate by investigating how testing "operates *on* social life, through the modification of its settings" (Marres and Stark, 2020: 423) should be an important endeavour for future Science and Technology Studies also in other areas.

Acknowledgements

This project was funded by the Research Council of Norway's TRANSPORT2025 program, under grant number 283354. We would like to thank Wiebe Bijker and two anonymous reviewers for their valuable comments to previous versions of this article.

References

Anderson JL, Asche F and Garlock T (2018) Globalization and commoditization: The transformation of the seafood market. *Journal of Commodity Markets* 12: 2-8.

Borup M, Brown N, Konrad K and Van Lente H (2006) The sociology of expectations in science and technology. *Technology analysis & strategic management* 18(3-4): 285-298.

Blyth PL (2019) Of Cyberliberation and Forbidden Fornication: Hidden Transcripts of Autonomous Mobility in Finland. *Transportation Research Part D: Transport and Environment* 71: 236-247.

Bulkeley HA, Broto VC and Edwards GA (2014) *An urban politics of climate change: experimentation and the governing of socio-technical transitions*. London: Routledge.

Combs TS, Sandt LS, Clamann MP and McDonald NC (2019) Automated vehicles and pedestrian safety: exploring the promise and limits of pedestrian detection. *American journal of preventive medicine* 56(1): 1-7.

Duarte F and Ratti C (2018) The Impact of Autonomous Vehicles on Cities: A Review. *Journal of Urban Technology* 25(4):3-18, DOI:10.1080/10630732.2018.1493883

EC (2017) Autonomous cars: a big opportunity for European industry. *Digital Transformation Monitor,* January 2017. Available at: https://ec.europa.eu/growth/tools-databases/dem/ Autonomous%20cars%20v1.pdf (accessed 10.1.2022).

Engels F, Münch AV and Simon D (2017) One site—multiple visions: visioneering between contrasting actors' perspectives. *NanoEthics* 11(1): 59-74.

Engels F, Wentland A and Pfotenhauer SM (2019) Testing future societies? Developing a framework for test beds and living labs as instruments of innovation governance. *Research Policy* 48 (9): 103826.

Ganesh MI (2020) The ironies of autonomy. *Humanities and Social Sciences Communications* 7(1): 1-10.

Gieryn TF (2006) City as truth-spot: Laboratories and field-sites in urban studies. *Social studies of science* 36(1): 5-38.

Gieryn TF (2018) *Truth-Spots*. Chicago, IL: University of Chicago Press.

Gross M and Hoffmann-Riem H (2005) Ecological restoration as a real-world experiment: designing robust implementation strategies in an urban environment. *Public Understanding of Science* 14(3): 269-284.

Guston DH (2014). Understanding 'anticipatory governance'. *Social studies of science* 44(2): 218-242.

Haugland BT (2020) Changing oil: self-driving vehicles and the Norwegian state. *Humanities and Social Sciences Communications* 7(1): 1-10.

Haugland BT and Skjølsvold TM (2020) Promise of the obsolete: expectations for and experiments with self-driving vehicles in Norway. *Sustainability Science, Practice and Policy* 16(1): 37-47.

Hodson M and Marvin S (2009) Cities mediating technological transitions: understanding visions, intermediation and consequences. *Technology Analysis & Strategic Management* 21(4): 515-534.

Hopkins D and Schwanen T (2018) Automated Mobility Transitions: Governing Processes in the UK. *Sustainability* 10(4): 956.

Hopkins D and Schwanen T (2021) Talking about automated vehicles: What do levels of automation do? *Technology in Society* 64: 101488.

Jasanoff S (2016) *The ethics of invention. Technology and the human future*. New York: Norton & Company.

Koetsier J (2020) Elon Musk: Tesla Will Have Level 5 Self-Driving Cars This Year. *Forbes*. Available at : https://www.forbes.com/sites/johnkoetsier/2020/07/09/elon-musk-tesla-will-have-level-5-self-driving-cars-this-year/#387795012d1d (accessed 12.07.2020).

Laurent B and Tironi M (2015) A field test and its displacements. Accounting for an experimental mode of industrial innovation. *CoDesign* 11(3-4): 208-221.

Leonardi P M (2010) From road to math; the co-ecolution of technological, regulatory, and organizatiional innovations for automotive crash testing. *Social studies of science* 40(2): 243-274.

MacKenzie D and Wajcman J (1999) *The social shaping of technology*. Buckingham, UK: Open university press.

Marres N (2020) What if nothing happens? Street trials of intelligent cars as experiments in participation. In: Maasen S, Dickel S and Schneider C (eds) *TechnoScienceSociety: Technological Reconfigurations of Science and Society*. Cham: Springer, pp. 111-130.

Marres N (2019) Co-existence or displacement: Do street trials of intelligent vehicles test society? *The British journal of sociology* 71(3): 537-555.

Marres N, Guggenheim M and Wilkie A (2018) *Inventing the social*. Manchester, UK: Mattering Press.

Marres N and Stark D (2020) Put to the Test: Critical Evaluations of Testing. *British Journal of Sociology* 71(3): 423-443

Milakis D, van Arem B and van Wee B (2017) Policy and society related implications of automated driving. *Journal of Intelligent Transportation Systems* 21(4): 324-348

Milakis D, Kroesen M, and van Wee B (2018) Implications of automated vehicles for accessibility and location choices: Evidence from an expert-based experiment. *Journal of Transport Geography* 68: 142-148.

Mladenović MN, Lehtinen S, Soh E and Martens K (2019) Emerging Urban Mobility Technologies through the Lens of Everyday Urban Aesthetics: Case of Self-Driving Vehicle. *Essays in Philosophy* 20(2):1526-0569.

Mutter A (2019) Mobilizing sociotechnical imaginaries of fossil-free futures–Electricity and biogas in public transport in Linköping, Sweden. *Energy Research & Social Science* 49:1-9.

Naber R, Raven R and Kouw M (2016) Scaling up sustainable energy innovations. *Energy Policy* 110: 342-354.

NHTSA (2020) Automated vehicles for safety. Web page. Available at: https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety (accessed 10.07.2020).

Nowotny H (2003) Democratising expertise and socially robust knowledge. *Science and public policy* 30(3):151-156.

Pinch T (1993) "Testing-One, Two, Three... Testing!": Toward a Sociology of Testing. *Science, Technology, & Human Values* 18(1): 25-41.

Pinch T J and Bijker W E (1984) The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social studies of science* 14(3): 399-441.

Pollock N and Williams R (2010) The business of expectations: How promissory organizations shape technology and innovation. *Social Studies of Science* 40(4): 525-548.

Ryghaug M, Ornetzeder M, Skjølsvold TM and Throndsen W (2019) The role of experiments and demonstration projects in efforts of upscaling. *Sustainability* 11(20): 5771.

Ryghaug M and Skjølsvold TM (2021a) Transforming society through pilot and demonstration projects. In: Ryghaug M and Skjølsvold TM (eds) *Pilot Society and the Energy Transition*, Cambridge: Palgrave, pp. 1-22.

Ryghaug M and Skjølsvold TM (2021b) The Co-production of Pilot Projects and Society. In Ryghaug M and Skjølsvold TM, Pilot Society and the Energy Transition. Cambridge: Palgrave, pp. 23-62.

SAE international (2016) Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. *SAE International (J3016)*.

Schot J and Steinmueller WE (2018) Three frames for innovation policy: R&D, systems of innovation and transformative change. *Research Policy* 47(9):1554-1567.

Schweitzer P, Povoroznyuk O and Schiesser S (2017) Beyond wilderness: towards an anthropology of infrastructure and the built environment in the Russian North. *The Polar Journal* 7(1): 58-85.

Shladover SE (2018) Connected and automated vehicle systems: Introduction and overview. *Journal of Intelligent Transportation Systems* 22(3): 190-200.

Skjølsvold TM (2014) Back to the futures: Retrospecting the prospects of smart grid technology. *Futures* 63: 26-36.

Skjølsvold TM and Ryghaug M (2015) Embedding smart energy technology in built environments: A comparative study of four smart grid demonstration projects. *Indoor and Built Environment* 24(7): 878-890.

Skjølsvold TM, Ryghaug M, and Throndsen W (2020) European island imaginaries: Examining the actors, innovations, and renewable energy transitions of 8 islands. *Energy Research & Social Science* 65: 101491.

Soteropoulos A, Berger M and Ciari F (2019) Impacts of automated vehicles on travel behaviour and land use: an international review of modelling studies. *Transport reviews* 39(1): 29-49.

Sovacool BK, Kester J, Noel L and de Rubens G (2019) Contested visions and sociotechnical expectations of electric mobility and v2G innovation in five Nordic countries. *Environmental Innovation and Societal Transitions* 31: 170-183.

Sperling D, van der Meer E and Pike S (2018) Vehicle Automation: Our Best Shot at a Transportation Do-Over? In: Sperling D (ed) *Three Revolutions.* Washington, DC: Island Press, pp. 77-108.

Stayton E and Stilgoe J (2020) It's Time to Rethink Levels of Automation for Self-Driving Vehicles. *IEEE Technology and Society Magazine* 39(3):13–19.

Stilgoe J (2017) Seeing like a Tesla: How Can We Anticipate Self-Driving Worlds? *Glocalism: Journal of Culture, Politics and Innovation* 3: 1-20.

Stilgoe J (2018) Machine learning, social learning and the governance of self-driving cars. *Social studies of science* 48(1): 25-56.

Stilgoe J (2020) *Who's driving innovation? New technologies and the collaborative state.* London: Palgrave.

Usenyuk S, Hyysalo S and Whalen J (2016) Proximal design: Users as designers of mobility in the Russian North. *Technology and culture* 57(4): 866-908.

Usenyuk-Kravchuk S, Klyusov N, Hyysalo S and Klimenko V (2020) DEPENDING ON USERS: The Case of Over-Snow Motorized Transport in Russia. *ICON* 25 (2):76-102.

Van de Poel I, Asveld L and Mehos DC (eds) (2017) *New perspectives on technology in society: Experimentation beyond the laboratory*. London: Routledge.

Van Lente H and Rip A (1998) The rise of membrane technology: from rhetorics to social reality. *Social studies of science* 28(2): 221-254.

Wetmore J (2003) Driving the dream: The history and motivations behind 60 years of automated highway systems in America. *Automotive History Review* 7: 4–19.

Williams R and Edge D (1996) The social shaping of technology. *Research policy* 25(6):865-899.

Notes

1 Although we are aware that pilot projects, test beds, experiments, demonstration projects and field tests and trials have distinct features and can be defined more precisely, they are often used interchangeably depending on empirical focus and disciplinary backgrounds (Engels et al., 2019). We have therefore chosen to try to use the term "pilot project" consistently throughout this paper, as this is how the project that we are studying are most often labelled, without drawing sharp boundaries to other similar terms.

2 In between levels 0 and level 5 this categorization gives the following levels: *one*, features which provide warnings and/or momentary assistance; *two*, when some functions respond using information about the driving environment, but the driver must be ready to take control; *three*, when cars are fully autonomous under certain traffic conditions; *four*, when cars also perform all safety-critical driving functions within a certain number of driving scenarios (SAE International, 2016).

3 The possibility to reach level five autonomy has been disputed (Stilgoe, 2018). Still, if the goal is level four autonomy, we would expect that a thorough and diverse testing is necessary in order to enable the technology to essentially bracket the context in which the vehicle is embedded.

4 A configuration where two or more trucks are linked in convoy, using connectivity technology and automated driving support systems (so that vehicles automatically maintain a set, close distance between each other).

5 Some cities are also viewed as more worthwhile experimenting on than others, for reasons related to economic, regulatory, cultural bearings or other aspects pertaining to how they are structured and what significance they are thought to have nationally and internationally.

6 https://www.vegvesen.no/Europaveg/e8borealis/inEnglish downloaded 22.10.19

7 The pilot project was tied to several other smaller trial projects in the region, where our interviewees reported having included public consultations. Our empirical material also includes a focus group interview with users (truck drivers that used the road, the truck drivers' association as well as the Mayor of the municipality). Road users and representative organisations were not the primary target for the focus in this article.

Policy Concepts and Their Shadows: Active Ageing, Cold Care, Lazy Care, and Coffee-Talk Care

Marie Ertner

IT-University of Copenhagen, Denmark, saramarie@itu.dk

Brit Ross Winthereik

IT-University of Copenhagen, Denmark, brwi@itu.dk

Abstract

In this article, we explore the form of care known as 'active ageing' by attending to its expression in care policies and within a Danish care home. We argue that active ageing policies gain their efficacy through reference to 'the good life', which is something the policies frame as ensuing if the elderly take on an active lifestyle. In the care home, the concept of active ageing gains its efficacy through its relation to other concepts of care, such as 'lazy care'. The importance of the article lies in its demonstrating the dependence of policy concepts on other concepts (established or emerging), which lie in its shadow yet do important political work. Attending to shadow concepts is useful if trying to understand the inner mechanics of popular concepts in care policy, as well as the norms and resistance to which they give rise.

Keywords: care policy, practice, active ageing, shadow concepts, politics of ageing

The sound of ceramics and metal clinking in the dining room signals that lunch is near. As always, Ellen, the resident from room 20 B, is there well before lunch, setting the table with plates and cutlery for everyone. Since she moved into the care home, setting the table for the daily communal lunch has been part of her everyday routine, because, as she says, she likes to be active and help. The care personnel are pleased with her initiative, and make sure to acknowledge her work by thanking her for setting the table in front of everyone else. The other residents in the lunchroom watch Ellen's active buzzing around and when asked to take part in the celebration of her selfinvented routine, Finn exclaims in obvious frustration and despair …, "I can't even take care of myself sometimes; I'm not like super-Ellen!" (Excerpt from fieldnotes, Ertner, 2012)

Introduction

'Active ageing' is a core concept in policies in Denmark and many other countries in the western part of the world as both an ideal and a goal in establishing good elderly care. It is also a key concept for international policy bodies such as the EU and the WHO. In Denmark, active ageing remains a central value in care programs targeting ageing and/or vulnerable citizens. Activation, development of personal potential, and individual responsibility are central to the concept and clearly

articulated in the Danish Social Service Law,[i] which is the legal basis for all Danish elderly care policies. But how has the concept of active ageing been implemented? And can it in fact be expected to become omnipresent in public care organisations and infrastructures? Within academia, the concept has been heavily critiqued by scholars who have raised concerns about the implications of the blanket application of active ageing policies at the expense of more differentiated ideals and understandings of ageing and activity (Katz, 2000; Lassen, 2014; Venn and Arber, 2011). In this article we build on this research to explore further what the policy of active ageing considers 'good' care, to leverage a more general discussion of how policy concepts gain their efficacy.

To do so, we dive into research within STS, where relations between policy and care practices are key objects of study, and care is understood as a situated practice that is social, political, and material, negotiated in a complex interplay with policy and institutional routines. This framework has also pushed for other forms of engagement with care work, and critique thereof, than the approach presented by much ageing research. Rather than treating policy as singular and detached from care work, and as a potential wrong-doer in the face of 'local practices', STS researchers seek to develop approaches that keep both policy and practice present in the analysis. This involves considering policy and care work as interconnected. In the following we take this interconnection as our point of departure to gain a better understanding of the activities and agencies to which the concept gives rise.

Theoretically, we combine work from within STS on policy and care practices with anthropological research on concepts as both abstractions and practices. This framework allows us to attend to policy as more than a political ideal aiming to discipline and prescribe, but, rather, as a lively entity that gains its efficacy through relations with social and material entities in 'concept complexes'. Critique, in this view, becomes less about deconstructing a singular policy idea to posit a criticism of its application, and more about exploring concepts 'at work'. This way we hope to inform an understanding of policy development that reflects policy concepts as lively and work against the 'hardening' of dominating and taken-for-granted concepts.

Contemporary discussions within STS and anthropology have emphasised the potential of ethnography to intervene in the worlds studied (Ballestero and Winthereik, 2021; Zuiderent-Jerak and Jensen, 2007). Jespersen et al. (2012) argue that it may do so by inciting a 'loosening' or 'releasing' of everyday categories through attending to the micro-processes of everyday life where categories are contested. According to Winthereik and Verran (2012), such loosening of categories is exactly the aim of what they term 'good faith' analyses. Constructing ethnographies in good faith is a matter of embodying an irresolvable tension between different versions of reality, which is needed when things present a multiplicity (Mol, 2002; Mol and Law, 2004) rather than adding up to a consistent whole (Law and Mol, 2002). The analysis presented here consists of empirical vignettes that engage with the micro-processes of everyday life in a Danish care home for the very elderly, where conceptions of ageing, activity and care are enacted, contested and negotiated. The vignettes are constructed to convey the multiplicity of care options in the context of active ageing policies, to loosen up (rather than doing away with) the concept of active ageing and unsettle its certainty when pronouncing certain people and practices 'good' or 'bad'. During ethnographic fieldwork in the care home, we became aware not only of the different ways in which 'active ageing' was being enacted, but also, more specifically, how it formed relations with other concepts of care, something which seemed important when trying to understand 'the politics of policy practices' (Gill et al., 2017) within the realm under observation. This article explores the implementation of a specific care policy, namely active ageing, through attention to its relations with other concepts of care in everyday practices within a care home – the shadow concepts of care.

We begin by situating the concept of active ageing, then present our theoretical and methodological framework, methods and ethical considerations. Through empirical vignettes, we illustrate interchanges between different concepts of care, and their implications for social and affective relations within the care home. We then discuss

our proposition that 'policy contains its Others within', and the implications of this insight for policy implementation.

Situating active ageing

As a country with universal welfare for its citizens, Denmark has one of the world's largest public sectors in relation to its population, with the state providing free access to healthcare via taxation (Andersen, 2008; Evans et al., 2018). Healthcare is the largest single area of national expenditure (Walker, 2008) and, with the number of retirees growing, birth rates falling, and years of economic regression taking their toll on national budgets, healthcare has been under pressure. Active ageing has been a prominent concept in Danish policy on ageing and elderly care for more than two decades, and has also become the key strategy in international and global policy in the field (Walker, 2008). Yet, despite its being a central concept for several international policy bodies, including the WHO and the EU, there is no single definition of the concept (Lassen and Moreira, 2014; 2020).

It is worth noting that active ageing policies have been heavily critiqued within ageing research[ii], with one point of contention being the associated tendency of problematising older adults as unproductive consumers of welfare resources (Foster and Walker, 2015; Lassen and Jespersen, 2015). For example, Evans et al. describe how general healthcare policies emphasise the imperative that older people remain 'free' of public services, drawing, in a neoliberal fashion, on an entrepreneurial-economic rhetoric which encourages them to make self-directed, 'responsible' choices that ensure they avoid the consumption of scarce public resources (Evans et al., 2018: 5). Some have pointed to the tendency of such policies to responsibilise older people in terms of their own health, promoting successful, positive, healthy, active paradigms (Evans et al., 2018; Katz and Calasanti, 2015), and introducing a productivist focus to care (Walker, 2008). Yet scholars of ageing have also argued that such policies overlook intersecting issues such as social inequality, health disparities, and age relations (Katz and Calasanti, 2015), and generate social exclusion and stigma (Lassen and Moreira, 2014). Others have shown how the active ageing discourse works to produce inappropriate recommendations for physical activity, which do not always meet the needs of the ageing population, and has a subsequent effect of assigning 'folk devil status' to the elderly population as a burden on society (Pike, 2011: 222). In a similar vein, active ageing policies have been shown to neglect the actual bodies of older people (Holstein and Minkler, 2007: 16) and interfere with 'healthy' bodily needs such as napping (Venn and Arber, 2011), creating 'busy bodies' and people struggling to 'reclaim their bodies, subjectivities and everyday lives from their management by activity' (Katz, 2000: 148). At the heart of these critiques is an understanding of active ageing as a policy concept with certain negative effects in practice.

The Copenhagen care reform, *Active and safe all life through*,[iii] was the dominant local policy reform framing elderly care during the time of our studies in the care home and has provided the foundation for elderly care policy in the municipality since then. The tendencies pointed out by scholars of ageing, such as the construction of ageing around the dichotomy of active and passive elderly, can be identified in the reform. Moreover, a neo-liberal logic can be seen as a pervasive in the way ageing is rendered a biomedical object for improvement and intervention through, for example, a focus on the physical rehabilitation and enhancement of the individual's functional capacity, autonomy, and self-care. However well-intended, the program provides few opportunities to reflect what Vicky Singleton (2005) has identified as 'promises' and 'vulnerabilities'. Singleton (2005) uses these notions to characterise policy that entails contradictions and tensions and transgresses traditional boundaries. She argues that vulnerable policies are promising in the sense that they are open to difference and ambiguity, and thus avoid the hardening of everyday categories (Singleton, 2005: 771). This gives rise to the question of how to think about vulnerability in context of the Copenhagen care reform.

The reform extends the notion of 'being active' from the locus of functionality and mastering everyday duties such as cleaning and shopping, to include attention to loneliness and inclusion in social activities and communities[iii]. In this light,

care is both a matter of tending to basic physical requirements for living and of quality of life. Care is thereby not merely an effect of elderly people taking it upon themselves to be active but is distributed across different actors, such as the municipality, care personnel, families, and friends, and as related to different situations such as leisurely activities and daily meals.

According to the critique by researchers on ageing, policies that equate good care with activating care encourage the opposite of good care by promoting ageist, stigmatising, and excluding narratives (e.g. Lassen and Moreira, 2014; Katz, 2000; Venn and Arber, 2011). The policy paradigm, according to critics, has little relation to care. We are sympathetic with the points of critique and the problematisation of active ageing policies, but also see promising contradictions and ambiguities in the specific local reform program, which appears to be less singular than it is framed in critical ageing research. Seeing ageing policy as multifarious, and regarding the concept of active ageing as a way of unpacking the policy's promises and vulnerabilities, became an important point of departure for our analysis of the active ageing policy in practice.

Fieldwork

During 2012, the first author carried out ethnographic fieldwork in a care home on the outskirts of Copenhagen, the capital of Denmark, over a period spanning June to September in 2012. During these months between two and five visits were paid to the care home per week, from around 9am to 3pm, covering lunch, physiotherapy, coffee breaks and so on.

The first author had a personal relationship with people in the home through her own family. The personal ties between the ethnographer and one of the residents in particular infiltrates the ethnography in the sense that observations are partly an outsider's views of the inside, and partly an insider's view. Rather than seeing this situation as a bias to be reduced, we see this double connection to the care home as a condition of what McGranahan describes as 'the ethnographic': a culturally grounded way of being in and seeing the world (McGranahan, 2018: 2), and an embodied, intellectual, and moral positionality (Ortner, 2006: 42) which interweaves research and personal life. More than that, as the fieldwork commenced while the first author was visiting the care home as a relative, it is also an instance of what Muncey (2010: 2) describes as auto-ethnography: a narrative that "emerges out of the iterative process of doing research, while engaging in the process of living a life". More specifically, in our case this narrative emerged from the experiences of frequent visits to the care home described by the first author to her supervisor, the second author, when conducting doctoral research on IT-design in a community of elderly people elsewhere in Copenhagen (Ertner, 2015).

Emerging from our conversations was an interest in the notion, very popular at the time, of active ageing: how it was enacted and with what implications for the care of people in the last years of their lives. Clearly, the concept was applicable to the elderly people in the case featuring in my doctoral research on an innovation and design project seeking to develop digital meeting places for so-called socially and physically active older people. But how did it fare in a care home for dying elderly people?

To attend to active ageing as something that is part of care practices as both actor and as shaping the situation, we turned to Annemarie Mol's (2002) notion of praxiography, introduced in her book *The Body Multiple*: a detailed and complex, practice-oriented ethnography of atherosclerosis focusing on the co-performance of things and knowledge of them:

> Because as long as the practicalities of *doing* disease are part of the story, it is a story about practices. A praxiography. The "disease" that ethnographers talk about is never alone. It does not stand by itself. It depends on everything and everyone that is active while it is being practiced. This disease is *being done*. (Mol, 2002: 31)

Comparably, our praxiographic approach studies the practicalities of doing, or enacting active ageing within a care home. As such, we do not attend to active ageing as a 'single thing', isolated and detached from other things and events. We are interested in active ageing as something that may

take various forms when it is done in various ways in practices that summon a range of different objects, persons, places, concepts, and actions. As such, we do not seek to develop a coherent argument about the essence of active ageing, but to inquire into variations and differences between and among concepts of care across care work and in policy documents. Fieldwork involved the first author's visiting the residents in their private accommodation in the care home, but more importantly it involved being present in communal areas and taking part in the mundane activities of everyday life. In fact, most of the time spent in the field involved participating in care work such as cleaning, escorting residents to various activities within the care home, helping at the daily lunch gatherings, serving food, and sitting in with the residents over lunch. Taking part in practical activities gave the first author a better sense of the daily routines in the care home and helped her become part of them, sometimes offering informal encounters and conversations with residents about topics meaningful to them. Besides ethnographic observation, she also interviewed the director of the care home and talked to care personnel. Because of the intensity of her presence, and the intimate situations in which she often found herself, it was not possible to make recordings. So, at the end of the day, she noted down her observations in fieldnotes, which comprise the empirical basis of this paper. The study has not received ethical clearance from a research ethics committee, since this is not required for ethnographic research projects in Denmark. Ethical considerations and reflections have been part of the research process in different ways. As Muncey reminds us, the ethics of narrative and storytelling involves considerations of respect for stories, in this case, close attention to the question of who can be harmed by the auto-ethnographic vignette (Muncey, 2010: 106). We fully acknowledge our narrative privilege and the inaccessibility of an academic writing style to the people whose life in the care home we are representing.

Theoretical framework: Policy, care, and concept complexes

Within STS and Critical Policy Studies, both policy and care are seen as open-ended, socio-material processes, with policy being negotiated and constantly undergoing change as it is implemented. Indeed, David Mosse (2004), as part of his work on development cooperation, offers a perspective on policy that suggests it is nothing more than a starting point that is always translated in practice. In Mosse's (2004) view, any criticism of faulty policy implementation must take its departure from this understanding, which opens up a more nuanced approach to what counts as success and failure in policy implementation (Jensen and Winthereik, 2013). In the context of care for older people, this implies studying the situated, material, and social practices of policy: the infrastructure of care, which includes attending to how policy is enacted both by care providers and ageing persons. Yet, for researchers studying policy in practice, its concepts can be quite hard to 'get into view' (Jensen, 2004; Jensen, 2010), something that can be ascribed to the many different translations of it that happen in practice, and to the multiple ways there are of 'knowing governance' (Voß and Freeman, 2016). Those who make policies, implement them, or are their recipients, 'arrive at' governance quite differently; if a policy is a shared resource for action it may be due more to its qualities as a boundary object than because different groups of people interpret it in exactly the same way.

Care has been defined as an affectively charged and selective mode of attention (Martin et al., 2015), and 'an affective state, a material vital doing, and an ethico-political obligation' (Puig de la Bellacasa, 2011). Maria Puig de la Bellacasa's (2011, 2017) work on matters of care in technoscience is of special importance to us. She discusses the potential of viewing socio-technical assemblages as matters of care, rather than as matters of fact or matters of concern, as proposed by Latour (2012). This, she argues, directs attention to, and raises awareness of, the ethico-political and ontological dimensions of care. Engaging with a 'thing' such as a policy concept, as a matter of care, is not a matter of critically deconstructing that policy concept (here active ageing); rather it

involves visions of 'cutting' the shape of it differently. Instead of doing away with the policy concept, this move enriches and affirms its reality by adding further articulations, recognising its 'liveliness', and generating 'more interest' (Puig de la Bellacasa, 2017).

Cultivating sensitivity towards the contrasting and ethico-political implications of different versions and practices of care allows care to be seen as having various effects. Depending on how they are assembled and done, caring practices can bring harm and hurt, just as much as they can nurture and heal. Care can also be found in the most unlikely places (Law, 2015). Feminist technoscience scholars have argued that studying care in practice requires that critical attention be paid to 'the dark sides' of care in order not to take for granted its seeming innocence; rather, both harmful and nurturing aspects should be open to exploration (Gill et al., 2017). In a similar vein, policy has been described as representing 'technologies of legitimation' (Harrison and Mort, 1998). Both policy and care are characterised as political practices, with opposing dynamics, that distribute relations of power and generate categories of difference (Gill et al., 2017). Thus, a central aim of research is to attend to and engage in the politics of policy practices, meanwhile addressing how to "think with the tension between the scales of policy and situated care practices and imagine methods that may hold these scales in tension or allow them to go-on-together in difference" (Gill et al., 2017:14; see also Verran). These studies, and the notion that policy can be understood and studied as practice, frame our own approach to understanding enactments of active ageing.

We juxtapose readings of policy documents with ethnographic vignettes to attend to affects, relations, materials, and unsettling constituents of active ageing that are otherwise hidden, neglected, or marginalised by formal policy notions (Singleton and Mee, 2017:131). Importantly, this juxtaposition shows that both policy and care are contested. They are practical achievements that can be explored symmetrically, which means that policy is not 'above' care; rather, it is interwoven with care as an intricate part of everyday life.

To understand how this interweaving works, we needed a framework that would allow concepts to be seen as practical achievements. Marilyn Strathern's (2011) notion of 'concept complexes' helped us incorporate the idea that both policy concepts and care, and the relations between them, are formed conceptually and practically, through their relations to other things. Strathern (2011) argues for a concept of concepts that recognises the relations they have with other entities in the world. Such relations are never stable, and if we see concepts as explanatory models, as theories that are somehow outside that which they describe, it will hinder our understanding of unfamiliar practices. So, because there is a limit to how much new understanding a concept can afford, we need to see concepts as themselves malleable and dynamic (but not infinitely flexible) and related to other concepts in practice.

Using ethnographic description, Strathern (2011) takes us through the conceptual architecture of the concept of 'borrowing' by describing its relations to two other concepts: 'stealing' and 'sharing'. The purpose is to demonstrate the limitation of all concepts and the value of reaching their limits. Her analytical practice is one of "playing off different conceptual worlds against one another" (Strathern, 2011: 14), suggesting that the creative potential of working with concepts and their limits is that it allows us to work with conceptual complexes. As she quotes, "a visible institution or practice is never simply identical with itself but always carries with it its invisible double or shadow, which can turn back upon it so that one crosses over and becomes the other" (Jiménez and Willerslev, 2007: 528-529, cited in Strathern, 2009: 12). According to Strathern (2011: 12) concepts and shadows always 'journey together', but their relation must be established anew every time through ethnographic description. Thus, her analysis demonstrates continuity between concepts that may seem foreign to each other. She demonstrates that, in a sense, they 'happen' simultaneously.

What we take (borrow) from Strathern (2011) is firstly that the relations between policy and care do not have to be oppositional. Not only is policy translated as it is implemented; it is also part of a complex, meaning that policy concepts are

related to many other things and contained and expressed in mundane practices such as table-setting and bed-making. This helps us understand how mundane practices and articulations that are otherwise not obviously connected to policy paint a picture in which a policy concept and the practice form some sort of alliance. The hands and policies that care for the elderly thus also bring about new relationships between policy and care. In the following section we introduce three ethnographic vignettes to describe how such relationships take shape.

Cold care – erasing actions of self-care

> "This is the coldest place I've ever been", says Frida, a female resident of 93, who recently moved from her own home to the care home. I have visited Frida nearly every morning over the last month. Most mornings she greets me with a smile, but not today. Her dark mood contrasts with the bright sunshine that pours in through the big windows, casting its light on the sofa-chair where she is sitting. An incident that morning has put Frida in a bad mood, I quickly learn. She tells me that a member of staff had come into her room with her breakfast and morning pills. When she discovered that Frida's bed was not made, she reprimanded her. "Why have you not made your bed, Frida? You are to make your own bed, you know." (Frida mimics the carer with an angry expression on her face). She looks at me with piercing eyes and reflects upon the incident. "Her having to say that to me made me so sad. You see, I felt like I had lost my mind. I always like to keep my quilt turned back for a while to air it, you know, I think it is more hygienic. But I began to think that maybe I had become like some of the others in here who have nothing more inside their heads, since she had to talk to me like that, like I was a child. I must be like them now, I thought. This really is the coldest place I have ever been."

As in the opening vignette, we see how the notion of active ageing incorporates specific expectations of residents with respect to being active. Following the policy concept, the care worker is dispensing good care as she encourages the resident, who has not made her bed, to be actively engaged in the maintenance of her own home. For a carer looking for indications of an active person, the unmade bed serves as a sign of a lack of active agency. Just as in Finn's sitting and waiting for someone to set the table and the food to arrive, a passive attitude towards caring for one's own bed and keeping a tidy home is undesirable. Yet the rather specific notion that an unmade bed indicates lack of agency on the resident's part is contested by the resident herself, Frida, who later explains that turning back her bed is part of her morning routine. Seeing the situation through Frida's eyes, we come to learn that not making the bed is part of an active, intentional strategy of self-care performed through first airing the bed and then later making it.

It may seem a banal incident; however, the carer's activating comment does a lot in terms of judging Frida's actions and distributing authority and agency between the two of them. The carer's comment implies a judgement of the unmade bed as a case of neglect of self-care. Resident Frida thus becomes a person who neglects to care for herself, a passive older person who needs to be activated to take responsibility for her own bed-making, not someone who may negotiate the meaning of 'active'. The situation does not just revolve around the bed; at stake is also Frida's possible disinclination to take responsibility for matters of personal hygiene and cleanliness more generally. In consequence, Frida is rendered a passive, untidy, and irresponsible person. She articulates that the comment resulted in strong emotional reactions and the feeling of having 'lost her mind', which makes sense given the different specificities of the context. In a care home, loss of abilities, sanity, and thereby authority is something that happens very visibly – sometimes gradually, other times rapidly – on an everyday basis to many of the residents. Many of the residents found it difficult to live so closely with this fact, which instigated speculations and doubt concerning their own status. Will I become like that, too? When? Is it me already?

In this light, the specific comment by the carer has consequences that reach further than the mere question of bed-making routines. According to Frida, it removes her sense of agency and authority, rendering the carer the authority in the home, with the mandate to take charge of things and act upon the unsatisfactory situation. Having

her agency erased by the carer's comment makes Frida doubt her own mental state and sanity, and simultaneously describe the care home as a cold place, and the care she receives as cold.

Seen from this perspective, an outsider's view might shift judgement, positioning the carer as not doing a good job, but this is not our point. The episode must be understood in a context where continuous demands for efficiency, budget costs, and austerity policies require care-home workers to do more in less time. Time is a limited resource, and the policy value of active ageing has become a central value in care work. The carer is acting in accordance with the pervasive policy of active ageing by seeking to perform activating care. In that sense, her care is good, and there may be many situations where such an approach would have a more positive impact. Yet Frida's story highlights that there are situations where the policy's idea of a clear split between active and passive older people sometimes leads to 'cold' encounters between carers and residents in the care home: cold in the sense that activation leads to judgements rather than mutual understanding and a recognition of the resident's actual agencies and meaningful actions.

The vignette shows a situation at the limit of what counts as active ageing; active ageing is a made bed, not bedclothes left to air. The situation tells us something about one shadow of active ageing, which emerges through Frida's reference to some of the other residents, namely those who "have nothing inside their head … like a child". This is a state of which she is clearly fearful. When active care involves practices of deleting agency and translating active practices into neglect and passivity, other concepts of ageing emerge, which in this case feed on notions of mental disability or even death, loss of mind, and insanity. These notions work as shadows that give meaning, often implicitly and in unspoken ways, to activating practices.

Lazy care – neglecting responsibilities to care

In the previous vignette, a resident experienced a carer's activating care as an accusation of lack of self-care. In other situations, accusations are made by the residents against the care workers, whom they complain are lazy. As one care home director told me:

> The recent message from the municipality is that we must motivate instead of serving. That means that we need to keep our hands behind our backs in order not to help the residents do things they can do themselves. And it's true. Take Ellen, for instance. When she moved in here, I went into her room and automatically started to make her bed. Then she said, "Excuse me, I actually do that myself". And it really improves her quality of life to do things for herself. Others complain and ask, "Have the personnel become lazy?" But the fact is, they are perfectly capable of doing it themselves.

Here, the care home director is talking about how staff members seek to implement active ageing in their routines. She explains that it is not an easy thing to do, that in fact she must consciously prevent herself from automatically dispensing care in 'the old way' by carrying out tasks for the residents, such as making their beds. She explains that the staff do this by using strategies of 'holding their hands behind their backs' to allow residents to play an active role in caring for themselves and their home environment, since this ultimately provides a better quality of life. Yet the residents do not always appreciate these efforts, and indeed complain about them, she says.

Residents respond by noting that the personnel "have become lazy". Accusations of laziness suggest that the care personnel are neglecting to perform actions that are part of their work. The director does not take such comments too seriously, however, since, as she says, the residents are perfectly capable of doing many things themselves. In that sense, the complaints would appear to be evidence of laziness on the residents' part. Indeed, the judgement of laziness is projected back and forth between residents and personnel, changing the meaning of care. The personnel, guided by municipal policy on active ageing, see care as the commitment to help older people to do things themselves, to motivate and encourage them to care for themselves. On the other hand, residents experiencing personnel 'keeping their hands behind their backs' see this new approach to care as a lack of care, as lazy care. So how

may we understand the projection of opposing notions onto the same action? Is one expression, one version of care, more correct than the other?

In the words of the director, what counts is the fact that residents can do things themselves, which makes their accusations of laziness on the part of the staff not something to worry about too much. The residents' complaints seem to hinge on something other than capabilities, rather coming from an experience of not feeling 'cared for'. The accusation of 'lazy care' thus changes the premises of the situation by drawing attention to how care is experienced by residents, instead of their competences and capabilities as assessed by the personnel. In that sense, the complaint of 'lazy care' by some residents may be understood as a response to a situation in which new notions and practices of care have been introduced in the care home. Through the notion of lazy care, residents actively challenge and negotiate the meaning of care and the authority to make judgments about who should be more active or not. Indeed, the query, "Have the personnel become lazy?" does much more than merely articulate a complaint about staff, it shifts the very infrastructure of responsibility to render care and the agency to make judgements about inactivity. Calling the personnel lazy in response to activating forms of care can be seen as an act of regaining power, and *actively* participating, this time not in bed-making, but in defining the meaning of the concept of care. Care in this sense, would be the opposite of hands being held behind the back to stimulate self-care; it would be hands-on, active involvement in care, and creating the experience in residents of being cared for.

Active ageing is depicted as a relationship between care personnel and older people where care is provided in the form of activating, encouraging, motivating gestures. Here elderly people are at the receiving end, following instructions, and growing in their independence and selfcare: a positive relationship conditioned by the older person's willingness to accept the care responsibility as theirs. When active ageing is reformulated as lazy care, other types of relations between personnel and residents emerge, and the relationship can be characterised as more like a battle, with the batting around of judgements of laziness, and contestation over what counts as care and who is responsible for its provision. This quote from the care home director clearly illustrates the thesis outlined above: that active ageing strategies sometimes journey with the shadow concept of 'lazy care'. In these situations, active ageing is not simply a positive and generative relationship, but a mode of relating that also encompasses implicit, sometimes explicit, judgements of laziness, neglect, or lack of responsibility for care. Thus, the notion of lazy care transpires in relations where opposition and contestation over what counts as care are at play, but not directly and mutually explored, voiced, and negotiated.

Coffee-talk care – caring for quality of life

A central concept within the Copenhagen active ageing policy reform is that of social activity, which was also a central concern of both care personnel and residents in the care home; however, it was not a straightforward matter. What it meant and required to be social was contested, and relations between social activity and care were continually being shuffled and negotiated, with different outcomes in terms of forms of care. While active ageing policies stipulate that social activity is very important for health and overall quality of life, municipal policy presents a sharp contrast between coffee talk and care: "Coffee talk, friendship and good neighborship has never and will never be a responsibility of the municipality."[iv] Although social activity is a central aspect of the policy paradigm, the central actors in social situations are figured to be the older people themselves, and various mechanisms work to exclude care workers from engaging in social activities with care home residents. For instance, as the allocation of time and resources is based merely on measurements of residents' capabilities, no time is set aside for care workers to socialise with them. Similarly, eating from the lunch menu provided by the care home is prohibited for employees, and the enforcement of this rule has resulted in the care-workers and residents eating separately, since care-workers bring their own lunches which they eat in the office during their break.

For the care-workers, socialising is a central aspect of their professional work, and in their view, it is a central factor in the residents' quality of life. Given the limited or non-existent time for carers to facilitate social activities, they have sought to create communal activities among the residents. Thus, to implement social activities in the daily routines, a common lunch has become mandatory. However, it is not easy to make social activity happen in the way envisioned, and both personnel and residents are frustrated that the atmosphere in the lunchroom is far from presenting a vision of social synergy and comradely encounters between the residents. As one care worker observed, "We feel it is important that they get out of their rooms and be a bit social. But they complain and say, "But nobody says anything." "Well, you don't say anything either", we say. When we are there, they talk; when we leave, they leave too."

Viewed from a policy perspective, the residents' resentment about, and resistance to, being socially active together could be seen as the opposite of active ageing – an example of withdrawal and passivity, an unhealthy attitude – hence something that should be counteracted with activity-inducing strategies. Indeed, the carers are often frustrated that the residents only engage socially when they are around. Lunches in the dining room are, therefore, often consumed in silence, and residents talk about the awkward atmosphere and tension. Other residents complain that they are not invited by the personnel for coffee.

> Frida: The living room is always empty, it's weird. I thought that being in a care home meant sitting together and drinking coffee. But I have never been invited for coffee. In here you have to care for yourself.
> Ethnographer: Can you not go to the living room and have coffee with some of the other residents?
> Frida: Sometimes I try to talk to the others, but you can't. It just gets completely, "Good day, man, axe handle" (an old Scandinavian expression indicating that a conversation is so lacking in meaning that it borders on the absurd).

Being invited for coffee is something very different from 'being active', according to policy. Frida had certain expectations of life in a care home which have not been fulfilled: sitting together and drinking coffee and being invited for coffee were among them. For many of the residents, the care workers must be part of social activities for them to be acknowledged as valuable social relations. Having someone there who is able to facilitate meaningful conversations, ask questions, and keep the dialogue going is important. While the policy makes a sharp distinction between coffee talk and care, for most residents being socially active with others is only possible when a care-worker is present to enable communication. When care-workers are taken out of that equation, but the ideal of socially active care home residents remains, it results in awkward, uncomfortable moments for residents that hamper any sense of being socially capable individuals, and of the care home as a social community. This serves as a reminder that coffee-talk is an important aspect of care seen from the perspective of residents.

Discussion: Shadow concepts as loosening agents

Current care policies present activation as care but ageing research has shown how active ageing policies sometimes come into conflict with care locally. Taking a departure point in the Copenhagen care reform, we find that the policy contains a hard dichotomy between active and passive ageing, as its critics have argued. However, we have seen that there are also promising contradictions, transgressions, and resistances to the concept, not least in the implementation of it. In our analysis of fieldwork material in the care home we saw attempts at loosening up the otherwise hard dichotomy of active and passive forms of ageing. We have identified shadow concepts as such loosening agents as they seem to contest the hardness of the concept of active ageing. Our empirical vignettes gave concrete examples of what such loosening work looks like in practice, when a resident reflects on how she prefers to make her bed in particular ways and at particular times, or another resident contests the required acknowledgement of somebody who is able to set tables for communal lunch. Like the UK health policies described by Singleton, practices of active

ageing sometimes create active citizens, but as we have shown they also generate notions of inadequacy, incompetence, laziness, and passivity.

We see the emergence of shadow concepts as potential openings for developing situated forms of policy implementation. If policy is recognised as thoroughly vulnerable and uncertain, as Vicky Singleton (2005) suggests, such moments of contestation could be treated as opportunities for collectively exploring, voicing, and negotiating different concepts of care. This would require personal, analytical, creative, and social resources, as well as training and time, but might lead to a form of 'activation' that would be beneficial for differentiated care practices that encourage the growth of social relations between people like 'Super-Ellen' and her fellow residents. More generally, policy development and its implementation require more than 'the right policy concept'. We, therefore, do not so much seek to critique and deconstruct the concept of active ageing, as we wish to point to the lack of recognition of the many ways in which it sprouts new concepts in use. Such multiplication is not necessarily a problem for the efficacy of a policy concept; rather, it should be considered an opportunity for decision makers to develop a better understanding of a policy's unplanned and unintended ontological effects that are, nevertheless, part of policy practices. What, then, would more careful policy and care practices be? Our proposition is that they would exhibit recognition of this 'liveliness' of policy concepts and their shadows in practice. Acknowledging and dealing with this aspect of policy concepts would emphasise the need for practical, material, and pedagogical resources to allow care workers to revisit ideals and develop practices that are sensitive to keeping the boundaries between different concepts open, negotiable, and ambiguous to allow for inclusion and differentiation.

Drawing on Marilyn Strathern's (2011) notion of concept complexes, we were able to extend the insight that active ageing policy is translated in practice by showing that it is not only dynamically and materially implemented, but also part of a conceptual complex that contains active ageing and various 'shadows' in practice. Analysing the shadow concepts of active ageing in practice alerted us to the ways in which notions of active ageing are related to other concepts. Tuning in on the enactments of these shadow concepts, as we have done, has showed us some of the darker sides of active ageing policy in practice. We find that thinking about relations between policy concepts and their Others in practice is highly relevant to grasping the relations between policy and care. If we are to understand how policy works in care practices, we cannot focus only on the policy concepts in themselves but need to attend to the nexus of other concepts, or shadows, through which policy is made to work in practice.

Researchers studying policy and care relations have called for ways to "think with the tensions between the scales of policy and situated care practices and imagine methods that may hold these scales in tension or allow them to go-on-together in difference" (Gill et al., 2017: 14). By tracing shadow concepts of active ageing policies in practice, our analysis seeks to connect mundane practices and things in the care home – bed-making, communal lunch arrangements, and coffee-drinking – with care policy, in order to examine the relations that develop as a result. Exploring policy and care relations by tracing concepts and shadow concepts can be one way of giving voice to otherwise marginalised and neglected experiences with care in the field (Puig de la Bellacasa, 2011), and a way to produce more 'careful' policies of ageing and care practices.

Conclusion

Based on auto-ethnography and the notion of concepts as practice, this article finds that as ageing policy is practiced, the concept of active ageing multiplies into various other concepts. In our analysis we have considered these other emerging concepts as shadow concepts: companions to a dominant concept that loosen up this concept in practice. The shadow concepts we found were 'cold care', 'lazy care', and 'coffee-talk as care'. Attention to policy concepts as working through complexes of concepts in practice, and therefore as 'lively', enables recognition and further exploration of how formal policy is implemented and received by 'users' in practice, for example, how a policy is resisted. This can help decision makers in

their adjustments of policies in ways that consider social and material translations, which can make space for autonomy and agency in the places where policy is meant to take effect.

A note on ethics in the study

Ethics in relation to both participation and consent were central concerns in this study. How should ethics be secured in relation to research when most of the participants are not able to process descriptions, oral or verbal, of a research project, even less to comprehend the purpose and implications of giving consent? As the American Association of Anthropology puts it, "Given the open-ended and often long-term nature of fieldwork, ethical decision making has to be undertaken repeatedly throughout the research and in response to specific circumstances" (ASA, 2011: 2). In contrast to what can be termed 'checklist ethics' concerned with rules and standards, our ethical commitment pertains more to ethics approaches that underscore the importance of situated reflection and negotiation of ethics with research participants. Such approaches, referred to as empirical, situated, or relational ethics, are concerned with acting in responsible, accountable, and reflexive ways throughout the course of fieldwork and research (Zigon, 2020; Willems and Pols, 2010).

In Denmark, people usually move to care homes towards the very end of their lives. This means that only the most frail and vulnerable of older people live in these institutions. Within a care home there are, therefore, many residents who suffer from different things such as cognitive or physical conditions or fatigue, which affect their ability to engage in conversations. Most residents were 80+ and many of them experienced reduced hearing, neurodegenerative diseases such as dementia, and other age-related frailties. This poses several problems in relation to following commonly prescribed formats of consent and developing 'patient perspectives' (Pols, 2005). Some residents were unable to understand our purpose in being in the care home, engage in longer dialogues, or verbally convey their perspectives and views in ways that were comprehensible to others. However, avoiding talking about active ageing policies in practice did not seem an ethical or 'good' solution. As Pols (2005, 210) points out, "analysing talk as an act of representation ignores the various performative aspects of talking that link the talking to a specific situation". As we did not want to exclude 'silent residents' from inquiries and representation, taking part in practical activities and communal situations in the care home became a core method for developing insights into how 'active ageing' was enacted in various situations, which meant that we chose to direct our research gaze as much at situations and everyday practices, as at individual residents and our conversations with them.

Oral consent was negotiated with research participants on an ongoing basis whenever possible. In other situations, it was not even possible or ethically viable to engage in conversations about consent. In these cases, and in general, ethics was pursued through situated and relational reflection over the sensitising awareness expressed by Jarrett Zigon:

> [E]thics as ongoing attunement is not about adhering to pre-established criteria or grammar, and neither is it about finding the slot of shared meaning. Rather, to the extent that language is a modality of ethical attunement, it is that call, that demand, that pull, that allows the possibility to dwell once again with others in the world between us. (Zigon, 2020: 1009).

Following this ethos of research goes far beyond consent forms, as it requires acknowledging that responsibility for the other is a commitment that stretches across time and space in our being relational (Zigon, 2020: 1010). Both care workers and residents who were formally interviewed and directly involved as informants were informed about the research. We considered how to balance the criteria of informed consent with sensitivity towards the often difficult mental and cognitive conditions of our informants. We wanted to avoid overburdening them with technical terms and loads of information that they could perceive as personally irrelevant, but at the same time not underestimate their need and capacity to understand the purpose of the project and our use of their data. In order to adapt this information to each informant, the ethnographer had

one-on-one conversations with them, adapting the choice of words and degree of detail to the mental and cognitive conditions of the particular informant. The conversations were highly dialogical and steered by the questions and concerns of the informants. We generally chose to use as few technical terms as possible, to avoid confusion.

To secure the confidentiality of the informants, all information has been kept safely stored and only shared between the authors of this paper. All names are pseudonyms, and to secure the anonymity of the informants, the name and location of the care home have been kept confidential, and empirical descriptions that could reveal their identities have been avoided.

Acknowledgements

Several colleagues have commented on previous versions of this paper, and we are very grateful to all of them. In particular, we would like to thank Marisol de la Cadena, Astrid Pernille Jespersen, Aske Juul Lassen, the anonymous reviewers and the editors. We owe great thanks to residents and employees at the care home, for letting us be a part of their everyday.

References

Andersen HT (2008) The emerging danish government reform–centralised decentralisation. *Urban Research and Practice* 1(1): 3–17. DOI: 10.1080/17535060701795298.

Association of Social Anthropologists of the UK and the Commonwealth (2011) *ASA ethical guidelines*. www.theasa.org

Ballestero A and Winthereik BR (2021) *Experimenting with Ethnography: A Companion to Analysis*. Durham & London: Duke University Press.

Ertner SM (2015) *Infrastructuring Design : An Ethnographic Study of Welfare Technologies and Design in a Public-Private and User Driven Innovation Project*. PhD Thesis, IT University of Copenhagen, software and systems section.

Evans AB, Nistrup A and Pfister G (2018) Active ageing in Denmark; shifting institutional landscapes and the intersection of national and local priorities. *Journal of Aging Studies* 46: 1–9. DOI: 10.1016/j.jaging.2018.05.001.

Foster L and Walker A (2015) Active and Successful Aging: A European Policy Perspective. *The Gerontologist* 55(1): 83–90. DOI: 10.1093/geront/gnu028.

Gill N, Singleton V and Waterton C (eds) (2017) *Care and Policy Practices*. London: Sage. Sociological Review Monographs Series 65: 2.

Harrison S and Mort M (1998) Which Champions, Which People? Public and User Involvement in Health Care as a Technology of Legitimation. *Social Policy & Administration* 32(1): 60–70. DOI: 10.1111/1467-9515.00086.

Holstein MB and Minkler M (2007) Critical gerontology: reflections for the 21st century. In: Bernard M and Scarf T (eds) *Critical Perspectives on Ageing Societies*. Bristol: The Policy Press, pp. 13–27.

Jensen CB (2004) Researching partially existing objects: What is an electronic patient record? Where do you find it? How do you study it? *The Centre for STS Studies, Aarhus*.

Jensen CB (2010) *Ontologies for Developing Things: Making Health Care Futures Through Technology*. Leiden: Sense Publishers.

Jensen CB and Winthereik BR (2013) *Monitoring Movements in Development Aid: Recursive Partnerships and Infrastructures*. Cambridge, MA: The MIT Press.

Jespersen AP, Petersen MK, Ren C and Sandberg M. (2012) Guest Editorial: Cultural Analysis as Intervention. *Science Studies* 25(1): 3–12.

Jiménez AC and Willerslev R (2007) An anthropological concept of the concept?: reversibility among the Siberian Yukaghirs. *Journal of the Royal Anthropological Institute* 13(3): 527–544. DOI: 10.1111/j.1467-9655.2007.00441.x.

Katz S (2000) Busy Bodies: Activity, aging, and the management of everyday life. *Journal of Aging Studies* 14(2): 135–152. DOI: 10.1016/S0890-4065(00)80008-0.

Katz S and Calasanti T (2015) Critical Perspectives on Successful Aging: Does It "Appeal More Than It Illuminates"? *The Gerontologist* 55(1): 26–33. DOI: 10.1093/geront/gnu027.

Lassen AJ (2014) Billiards, Rythms, Collectives, Billiards at a Danish Activity Centre as a Culturally Specific Form of Active Ageing. *Ethnologia Europaea: Journal of European Ethnology* 44(1): 57–74.

Lassen AJ and Jespersen AP (2015) Ældres hverdagspraksisser og aldringspolitik. Om synkroniseringsarbejdet imellem hverdag og politik. *Kulturstudier* 6(1): 79. DOI: 10.7146/ks.v6i1.21242.

Lassen AJ and Moreira T (2014) Unmaking old age: Political and cognitive formats of active ageing. *Journal of Aging Studies* 30: 33–46. DOI: 10.1016/j.jaging.2014.03.004.

Latour B (2012) Why Has Critique Run out of Steam? From Matters of Fact to Matters of Concern. *Critical Inquiry* 30(2): 225–248.

Law J (2015) Care and killing Tensions in veterinary practice. In: Mol A, Moser I, and Pols J (eds) *Care in Practice: On Tinkering in Clinics, Homes and Farms*. Bielefeld: transcript Verlag, pp. 57–72. DOI: 10.14361/transcript.9783839414477.57.

Law J and Mol A (2002) *Complexities: Social Studies of Knowledge Practices*. Durham: Duke University Press.

Martin A, Myers N and Viseu A (2015) The politics of care in technoscience. *Social Studies of Science* 45(5): 625–641. DOI: 10.1177/0306312715602073.

McGranahan C (2018) Ethnography Beyond Method: The Importance of an Ethnographic Sensibility. *Sites: A Journal of Social Anthropology & Cultural Studies* 15(1): 1–10. DOI: 10.11157/sites-id373.

Mol A (2002) *The Body Multiple : Ontology in Medical Practice*. Durham & London: Duke University Press.

Mol A and Law J (2004) Embodied Action, Enacted Bodies: the Example of Hypoglycaemia. *Body & Society* 10(2–3): 43–62. DOI: 10.1177/1357034X04042932.

Mosse D (2004) Is Good Policy Unimplementable? Reflections on the Ethnography of Aid Policy and Practice. *Development and Change* 35(4): 639–671. DOI: 10.1111/j.0012-155X.2004.00374.x.

Muncey T (2010) *Creating Autoethnographies*. LA, London, New Delhi: SAGE. DOI: 10.4135/9781446268339.

Ortner SB (2006) *Anthropology and Social Theory: Culture, Power, and the Acting Subject*. Durham: Duke University Press.

Pike ECJ (2011) The *Active Aging* Agenda, Old Folk Devils and a New Moral Panic. *Sociology of Sport Journal* 28(2): 209–225. DOI: 10.1123/ssj.28.2.209.

Pols J (2005) Enacting Appreciations: Beyond the Patient Perspective. *Health Care Analysis* 13(3): 203-21. DOI: 10.1007/s10728-005-6448-6.

Puig de la Bellacasa M (2011) Matters of care in technoscience: Assembling neglected things. *Social Studies of Science* 41(1): 85–106. DOI: 10.1177/0306312710380301.

Puig de la Bellacasa M (2017) *Matters of Care: Speculative Ethics in More than Human Worlds*. Minneapolis & London: University of Minnesota Press.

Singleton V (2005) The Promise of Public Health: Vulnerable Policy and Lazy Citizens. *Environment and Planning D: Society and Space* 23(5): 771–786. DOI: 10.1068/d355t.

Singleton V and Mee S (2017) Critical compassion: Affect, discretion and policy-care relations. *The Sociological Review 65 (1).*

Strathern M (2011) Sharing, Stealing and Borrowing simultaneously. In: Strang V and Busse M (eds) *Ownership and Appropriation*. Oxford: Berg Publishers, pp. 23-41.

Venn S and Arber S (2011) Day-time sleep and active ageing in later life. *Ageing & Society* 31: 197-216. DOI: 10.1017/S0144686X10000954.

Verran H (2013) Engagements between disparate knowledge traditions: Toward doing difference generatively and in good faith. In: Green L (ed) *Contested Ecologies*. Cape Town: HSRC Press, pp. 141– 162.

Voß J-P and Freeman R (eds) (2016) *Knowing Governance: The Epistemic Construction of Political Order.* Palgrave Studies in Science, Knowledge and Policy. Basingstoke: Palgrave. DOI: 10.1057/9781137514509.

Walker A (2008) Commentary: The Emergence and Application of Active Aging in Europe. *Journal of Aging & Social Policy* 21(1): 75–93. DOI: 10.1080/08959420802529986.

Willems D and Pols J (2010) Goodness! The empirical turn in health care ethics. *MEDISCHE ANTROPOLOGI* 22(1): 161.

Winthereik BR and Verran H (2012) Ethnographic Stories as Generalizations that Intervene *Science & Technology Studies* 25(1): 37-51.

Zigon J (2020) *Morality: An Anthropological Perspective*. New York: Routledge.

Zuiderent-Jerak T and Jensen CB (2007) Editorial Introduction: Unpacking 'Intervention' in Science and Technology Studies. *Science as Culture* 16(3): 227–235. DOI: 10.1080/09505430701568552.

Notes

i Ministry of Social Affairs and the Interior, Denmark, 2019, §1)

ii We use the term 'ageing research' to cover a wide body of research in age and ageing, such as gerontology, and cultural gerontology

iii Copenhagen municipality reform program *Aktiv og tryg hele livet* (2011) [Active and safe all life through]. This policy was referred to as a reform program because of the explicit transition to an active ageing paradigm. The reform program can be found here (in Danish) https://www.kk.dk/sites/default/files/agenda/a1bdf595b507bede1e0569d2fe75121690dd4448/7-bilag-3.PDF. Since then, other policies have followed under the same banner of active ageing; 2015-2018 *Live strong all life through*, 2019-2022 *Keep up all life through*.

iv Copenhagen municipality reform program *Aktiv og tryg hele livet* (2011:8) [Active and safe all life through].

Constructing 'Doable' Dissertations in Collaborative Research:
Alignment Work and Distinction in Experimental High-Energy Physics Settings

Helene Sorgner

University of Klagenfurt, Austria/helene.sorgner@aau.at

Abstract

Many young scientists are trained in research groups, yet little is known about how individual doctoral dissertations are carved out of collaborative research projects. This question is particularly pronounced in high-energy physics, where thousands of physicists share an experiment's apparatus, data, and the authorship of publications. Based on qualitative interviews with researchers working at CERN's Large Hadron Collider, this paper analyses what makes a PhD dissertation 'doable' in this context. Describing the levels of work organisation, the challenges, and the actors involved in constructing 'doable' dissertations in collaborative research, I argue that doctoral dissertations are the emergent product of alignment work performed throughout the PhD. Individualisation is achieved by temporally, qualitatively and formally distinguishing dissertations from work on collective publications. I discuss how these processes shape the roles of students and advisors, and the content and value of dissertations in collaborative research.

Keywords: alignment work, collaboration, doctoral students, dissertations, authorship, physics

Introduction

The observable increase in the number and size of research collaborations across the sciences (Milojević, 2014; Wuchty et al., 2007) seems to be in conflict with traditional academic career and reward systems focusing on individual achievements (Mangematin, 2001). This includes the work of PhD students, who contribute substantially to collaborative knowledge production (Larivière, 2012). Although Science and Technology Studies (STS) has long had an interest in the socialisation of students as members of research communities, there is no dedicated study on the practices that shape doctoral dissertations in collaborative environments. PhD dissertations based on collaborative research need to satisfy seemingly opposing requirements. As an academic qualification, the dissertation should constitute an independent and original research contribution, yet contributing to research in practice means supporting the ongoing work of a collective. Against the backdrop of this structural tension between collaborative research practices and individual attribution

of results, this paper asks *how doctoral dissertations – as individually attributed research outcomes – are made doable in collaborative research.*

Contemporary experimental high-energy physics presents an extreme case of collaborative research, where thousands of physicists share an experiment's apparatus, data, and the authorship of publications. About one third of the researchers involved in the experiments at the European Laboratory for Particle Physics' (CERN) *Large Hadron Collider* are doctoral students. Based on an analysis of interviews with graduate students, post-docs, and PhD advisors in experimental high-energy physics, this paper describes the practices involved in constructing dissertations that contribute to collective research goals while being attributable to an individual student. I refer to dissertations satisfying both requirements as 'doable', drawing on Fujimura's (1987) concept of 'doable problems'. Constructing doable problems in collaborative research requires 'alignment work' (Jackson et al., 2011) between various levels of work organisation. Describing the levels of work organisation, the challenges, and the actors involved in constructing 'doable' dissertations in collaborative research, I argue that doctoral dissertations are the emergent product of alignment work performed throughout the PhD. Given that academic qualifications rest on the attribution of work to a single author, I also describe how dissertations are distinguished from collaborative work. These processes have implications for the respective roles of students and their advisors. Practices of distinction also shape the contents and value of dissertations vis-à-vis other products of collaborative research, particularly the collective publications of results. My work contributes to studies on knowledge production and doctoral training across epistemic cultures (Delamont et al., 2000; Knorr Cetina, 1999) and demonstrates that a focus on dissertations offers a magnifying lens on the internal dynamics of collaborative research.

Doctoral students in collaborative research

Existing work in social studies of science has conceptualised doctoral training as a process of socialisation into culturally specific forms of knowledge production. As such, the PhD involves transmitting tacit problem-solving skills (Delamont and Atkinson, 2001) and a field-specific habitus (Delamont et al., 1997; Traweek, 1988). Doctoral training and the format of students' contributions reflect a research community's specific epistemic practices and work organisation. Compared to the humanities and social sciences, PhD students in natural sciences work less independently (Laudel and Gläser, 2008), often as members of research groups with a clear division of labour (Delamont et al., 2000). Research groups in turn have multiple and sometimes conflicting functions, serving as sites of academic training and career building as well as of (collaborative) knowledge production (Hackett, 2005).

Studies focusing specifically on the contributions of doctoral students to collaborative research are few and far in between. The most comprehensive comparative study (Delamont et al., 2000) found that in laboratory-based research groups, research problems are typically passed on from one generation of doctoral students to the next. Students do not have much choice in their topics, theoretical frameworks, or research methods, as these are determined by the advisor and the group. Advisors take care to choose experiments that can be expected to deliver publishable results within the timeframe of the PhD, and assign back-up problems to students, in case an initial project does not work out (Campbell, 2003). Although publications based on a student's work will usually be co-authored by their advisor and other collaborators, existing authorship conventions ensure that the main contributor can be identified (Laudel, 2001).

More recently, STS research has focused on how external factors such as changes in research governance affect epistemic and social practices in research groups (e.g. Fochler et al., 2016; Müller, 2014), including the construction of 'interesting' research problems (Rushforth et al., 2019). It has been argued that tighter funding regimes entail a 'projectification' (Ylijoki, 2016) of research, based on third-party funding with clearly defined deliverables and timeframes (Whitley et al., 2018). Doctoral students are increasingly hired as members of project-specific research groups, where they may be required to 'tailor' their disser-

tations to the demands of research funders (Möllers, 2017). Depending on the supervisory styles of PhD advisors (Louvel, 2012) and the ability of research groups to create a 'protected space' for PhD students (Degn et al., 2018) students may be more or less required to align their research external productivity goals.

Given that these observations mainly concern smaller research groups based at a single laboratory, it is unclear how well they map onto large-scale collaborations. In experimental physics, where collaborative research is the norm and dissertations are best characterised as post-hoc collections of a student's contributions to a team effort, the trend towards 'projectification' of doctoral training may be resisted (Torka, 2018). Contemporary experimental high-energy physics is collaborative in quite a radical sense, as no single step in the research process – from planning and building the technical infrastructure to taking, reconstructing and analysing data – could be achieved by an individual, a team or even a large research institution (such as CERN) alone. Moreover, high-energy physics experiments are known for their egalitarian and consensus-oriented style of self-governance (Knorr Cetina, 1995; Shrum et al., 2007). Experimental results are always published in the name of the entire Collaboration[1] (listing up to 3000+ authors in alphabetical order) running the experiment. This convention of collective authorship (Biagioli, 2003; Galison, 2003) recognises the broad range of contributions and extensive internal review required for any publication (Graßhoff and Wüthrich, 2012), and establishes the Collaboration as a collective epistemic subject (Knorr Cetina, 1999). Although the collectivisation of results and reputation prevents internal struggles for authorship, it raises the question of how individual achievements are adequately recognised within and beyond the Collaborations (Birnholtz, 2006; European Committee for Future Accelerators, 2015). This question also concerns dissertations, which require individual authorship, implying that students' contributions need to be actively distinguished from collective research outcomes.

The process of constructing doctoral dissertations in high-energy physics Collaborations differs from the same process in laboratory-based research groups in at least three significant aspects. The first is the convention of collective authorship, which troubles the identification and attribution of individual contributions. Second, due to the wide range of tasks involved in experimental research, PhD students often contribute to the work of several different groups within their Collaboration. We may ask how the availability of many different potential projects and supervisors shapes students' contributions and affects the respective roles of students and PhD advisors, in comparison to the research groups described above. Third, the peculiar timelines of high-energy physics experiments present a potential challenge for constructing dissertations. A single cycle of data-taking and analysis may take several years. One such process also involves the work of several different groups, which means that its completion is beyond the control of any individual team or group leader. This raises the question of how dissertations, which need to produce individually attributable results within a given timeframe, are constructed despite the intrinsically collaborative nature and long timespans of research.

Doable problems and alignment work

To answer the questions raised above, I will use the sensitising concepts 'doable problems' and 'alignment' introduced in Fujimura's (1987, 1996) study on oncogene research. Given that a dissertation should produce a research contribution, we may conceptualise it as consisting of (one or several) 'doable' research problems. Fujimura argues that the 'doability' of research problems not only depends on their technical feasibility but is actively constructed as researchers align tasks at several levels of work:

> In fact, scientific work gets done and problems are solved when all the necessary parts at all levels of work organisation are collected and made to fit together. […] That is, articulation between levels is required to bring all the tasks at different levels of work organisation together into alignment to create a doable problem. Problems are more or less doable depending on how difficult it is to articulate among levels to create alignment. (Fujimura, 1987: 262).

In Fujimura's case study, the levels of work organisation in need of alignment are an experiment as a set of tasks, the laboratory where experiments are conducted, and the wider social world of cancer research and molecular biology. Fujimura mentions the case of a PhD student who had to give up his initial project one year before graduation, because the problem had been solved and published by a different research group. Instead of postponing his graduation, which future hiring committees might interpret as a personal failure, the student focused on a secondary problem to finish his dissertation on time (Fujimura, 1987: 262-264; Fujimura, 1996: 171-172). This example illustrates how constraints arising at a different level of work (the 'social world') instigate a researcher to re-organise their experimental work. The initial problem was not 'doable' as a dissertation project anymore, because it did not meet the requirements that a dissertation contain original research, and that graduate research should not exceed a certain period of time.

Jackson et al. (2011) extend Fujimura's notion of alignment to the temporal dimension and the challenges of multi-sited research in large-scale collaborations. The authors point out that to make collaborative research doable, researchers need to reconcile the different temporal structures or 'rhythms' emanating from organisations, infrastructures, phenomena, and researchers' own biographies:

> To resolve issues of temporal conflict and fit, participants build instruments and environments, reshape organisations and institutions, and recraft or reorient their personal lives. All of this constitutes what we refer to here as *alignment work*, understood as the complex set of actions and activities required to bring otherwise disparate rhythms into heterogeneous and locally workable forms of alliance. (Jackson et al., 2011: 251; emphasis added)

This concept of 'alignment work' draws attention to the material and biographical aspects of collaborative research, which are only implicit in Fujimura's conception. To stabilise levels of work organisation and enable the configuration of tasks and problems, the organisational, infrastructural, phenomenal and biographical dimensions of distributed scientific work need to be (at least temporarily) aligned. These dimensions provide temporally situated resources and constraints ('rhythms') for the construction of doable problems. Such resources and constraints include the availability of instruments and data at different sites (Bruyninckx, 2017); the life cycles of research objects (Dippel, 2019); the academic schedules of collaborators and their institutions; the recurring dates of major conferences (Ochs and Jacoby, 1997), and the individual time constraints of researchers' lives beyond the lab.

For the purpose of analysing doctoral students' research, I adapt Fujimura's concept of 'doable problems' and Jackson et al.'s concept of 'alignment work'. We may distinguish several levels of work organisation relevant to the construction of 'doable' dissertations in collaborative research, which are in turn structured by the infrastructural, phenomenal, organisational and biographical 'rhythms' described above, and subject to 'alignment work'. Work organisation takes place on and between these levels: the *level of individual tasks* done by the student (corresponding to Fujimura's 'experiment'), the *level of the group or team* working together on the same project (corresponding to the 'laboratory') and the *level of the epistemic community* (corresponding to the 'social world'). We may expect these levels of work organisation to be relevant to doctoral students' work in all disciplines where collaborative research is the norm.

Experimental high-energy physics presents a specific case, because research groups are joined into large research Collaborations. This means that beyond the individual and the group level, there are several formally distinguished levels of work organisation within the Collaboration that doctoral students' work is embedded in (cf. Fig. 1). Moreover, because the majority of active high-energy physicists are members of only a handful such collectives, the Collaboration is, in many ways, directly equivalent to the epistemic community or 'social world' for a student. Alignment with to the level of the epistemic community as described in this paper is thus specific to collaborative research where collaborators beyond the local research group may directly influence PhD researchers' work.

Materials and method

My analysis builds on 15 interviews conducted in the course of a research project on the social and epistemological conditions of knowledge production in high-energy physics experiments.[2] This sample contains two different types of semi-structured expert interviews. The first type are exploratory interviews with ATLAS Collaboration members at different career stages based at research institutions in Germany and the US. These interviews covered a wide range of topics concerning collaborative research. The second type are problem-centred interviews conducted with ATLAS and CMS Collaboration members who were selected for their familiarity with a specific research topic or organisational process.[3] Although initially corresponding to different research interests, both types of interviews provided insights on the construction of dissertations, as became evident during the first round of analysis.

For the 12 interviews of the first type (7 PhD students, 3 professors, 2 post-docs), I visited one US-American and two German university departments in 2018 and 2019.[4] These brief two-day research visits allowed for informal conversations with researchers during lunch and coffee breaks, which were helpful in contextualising my interviews. My sampling strategy was to gather a range of perspectives from within the same institution, which means that the researchers I interviewed were not necessarily working closely together (except for two professor/post-doc/student triangles: Philipp/Natalie/Judith and Toby/Cara/Sam, the professors in both cases being experienced group leaders and PhD advisors). The interviews focused on the development and organisation of a researcher's work within their department and their working group in the Collaboration, as well as the supervision and situation of doctoral students.

In the course of analysis of the research project's shared interview pool, I supplemented this sample with three more interviews of the second type, which my colleagues and I had conducted to learn about specific Collaboration-internal processes.[5] I selected these accounts from experienced senior researchers (10+ years of supervising students) because they illustrate important aspects of the integration of PhD students' work in their respective Collaborations. In these interviews, the supervision of PhD students was not initially addressed by the interviewer. That dissertations nevertheless became a topic indicates the significance of PhD students' work for collaborative research processes.

Most interviews were conducted in person, at researchers' workplaces or in one of the cafeterias at CERN. Two interviews were conducted via video call. Interviews lasted between 45

Table 1. Selected interviews

Interviews Type 1			
Group 1	Germany	Professor	Philipp
		Post-Doc	Nathalie
		Student	Judith
		Student	Anton
		Student	Brian
Group 2	Germany	Professor	Tim
		Student	Matilda
		Student	Gabriel
Group 3	USA	Professor	Toby
		Post-Doc	Cara
		Student	Sam
		Student	James
Interviews Type 2			
	France	Professor	Simon
	France	Professor	Paul
	UK/Germany	Professor	Karen

minutes and 2 hours. Upon obtaining the explicit consent of the interviewees, they were recorded and transcribed verbatim.[6] I analysed interviews using the Atlas.ti software, following the principles of Grounded Theory (Charmaz, 2006; Corbin and Strauss, 2008) in the manner of the 'flexible coding' approach (Deterding and Waters, 2021). After an initial round of close reading and thematic coding, I identified the negotiations involved in constructing dissertations to be an emerging topic and focused selectively on references to such processes. The analytic category of 'alignment work' emerged from iterative open coding and comparative analysis of these passages. Starting from the observation that interviewees often indicated 'misalignments' between individual and collective projects or the necessity for 're-aligning' a student's work to that of a group, I noticed that also seemingly unproblematic cases of 'deciding on a topic' or 'being assigned a task' may be understood as instances of alignment work, as I will describe below.

The organisation of research in the ATLAS Collaboration

My case study focuses on PhD students in the ATLAS Collaboration, a research organisation building, running and maintaining the eponymous particle detector at CERN. The ATLAS Collaboration currently comprises research groups based at 181 research institutions from 41 countries. Of the more than 3000 researchers actively involved in ATLAS, about 1200 are doctoral students.[7] ATLAS is the largest of the four experiments recording and analysing the decay products of proton-proton collisions produced by the Large Hadron Collider (LHC). Its main scientific goals, shared with the CMS experiment, are to confirm and study the Higgs boson, and to discover hitherto unknown phenomena ('novel physics'). As protons collide and produce energy, new particles (such as the Higgs boson) are created and decay into other particles (such as electrons, photons or muons). From the traces of decay products registered by the detector, the original particle produced in the collision can be statistically inferred. To do so, the relevant data need to be selected, processed and calibrated, and the objects of interest need to be reconstructed and distinguished from noise and background processes. A 'physics analysis', the research process that leads to potentially new and publishable results, is only the last step in a long line of technical and analytic tasks. Physics analyses may be 'measurements' of properties of known particles or 'searches' for new particles and phenomena.

The main branches of the ATLAS Collaboration's internal organisation represent the activities necessary to run the experiment (including data preparation, software and computing, and 'hardware' work on the detector), with 'physics analysis' being one such activity. The branch of 'physics analysis' is organisationally divided into 'combined performance groups', which calibrate analysis methods and study their efficiency, and 'working groups' focusing on specific searches and measurements (Fig. 1). A prominent working group in ATLAS, such as the Higgs boson group, may have several hundred members and is further divided into subgroups investigating specific 'decay channels' of the Higgs. One subgroup, for example focusing on Higgs bosons decaying into two b-quarks, is made up of several analysis teams.

Students become involved in the ATLAS Collaboration through their affiliation with an institution that hosts an ATLAS group. The student's advisor and a few other researchers and students at the same department constitute the student's 'local group'. PhD students are typically based at their home institutions, working from their local offices and collaborating with other ATLAS members remotely. If their home institution has enough funding, PhD students may also spend between a few months to a year at CERN.

For all the students I interviewed, original contributions to at least one physics analysis – ideally resulting in a publication – were required to obtain a PhD in experimental particle physics. This means that the PhD student will be a member of an analysis team embedded in a subgroup of a working group in ATLAS. The student's main analysis project would usually be related to the research foci of their advisor's local group, and their analysis team and working group would often (but not necessarily) include local colleagues. A local post-doc would then supervise

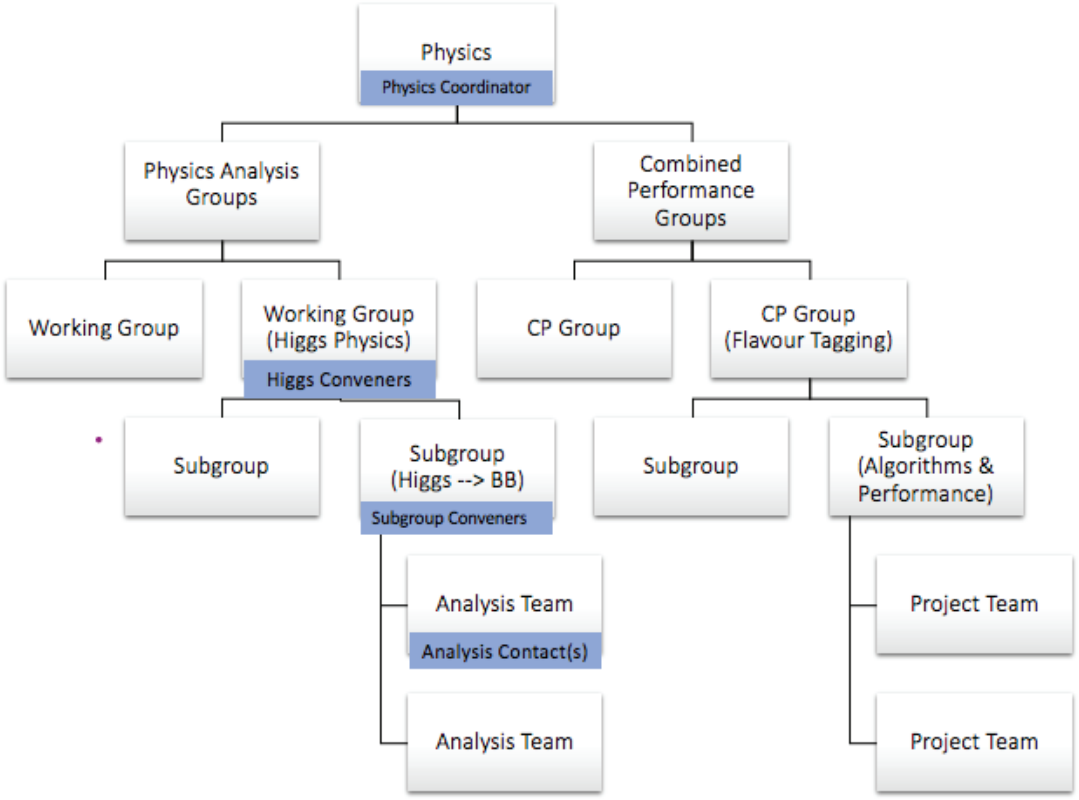


Figure1. Levels of Work Organisation in ATLAS

the student on the job, with the advisor receiving regular updates. As in other natural science disciplines (Delamont et al., 2000), PhD advisors steer the research of their students on a strategic level and are not involved in its hands-on aspects.

A PhD student is also expected to contribute to other activities of the experiment besides physics analysis. To be included on the list of ATLAS authors, every new Collaboration member must complete a 'qualification task'. The qualification task is defined as a purely 'technical' contribution, for example, to detector hardware (upgrading and testing a specific detector component), data preparation, or to 'combined performance'. Its completion should take the student about half a year of full-time work. Depending on the commitments of their local group and their personal interests, students may continue to contribute to technical activities beyond their qualification task. For this reason, PhD students in ATLAS often simultaneously work on several projects within the Collaboration, each coming with its own group and supervisors.

Constructing dissertations in the ATLAS Collaboration

The brief introduction to ATLAS research indicates that doctoral students perform several tasks within the Collaboration, and that besides their local advisors, group conveners and analysis contacts are involved in articulating those tasks. To illustrate how a student's tasks over time evolve into a dissertation, the following section describes this process from the perspective of two ATLAS PhD students. Both students were close to finishing their dissertation when I interviewed them and have graduated since. Their accounts allow a more comprehensive description of the several-year long process of constructing a dissertation in comparison to those of the advisors and coordinators discussed in the sections below, which zoom in on specific challenges within that process. The two cases are similar in that both dissertations were significantly based on 'technical' contributions. They differ in terms of how easily the requirement for the 'main project', i.e. a contribution to physics analysis, was fulfilled.

Natural choices and being lucky

Judith[8], an advanced PhD student at a German university, described the construction of her 'main analysis' as a somewhat organic process. She had developed an interest in particle physics during high school and proceeded to do a bachelor's and master's degree under the supervision of ATLAS physicists at her local university. Halfway through her master's programme, she was offered a PhD position with the same ATLAS group. In retrospect, her personal interests were perfectly aligned with those of the group at the local department, as well as the research agenda of ATLAS:

> So, I already did a very basic analysis in my master's thesis looking for exactly such a heavy particle […]. And then somehow it became a thing in ATLAS that they also wanted to do that analysis and then this was basically the natural choice, to say that we want to participate there. […] This has also grown historically, these [specific analyses] are something that the local group has been doing for a while. And then we just kind of went along with the course of events in ATLAS.

Judith refers to the project that she first started working on as a "natural choice" for her group, since there was an interest on the side of the Collaboration to do searches for such (unknown) heavy particles. The "historic" development of research at her local department coincided with the research priorities of the Collaboration following the confirmation of the Higgs boson. This created favourable conditions to further pursue an analysis project that she had started, in a rudimentary form, in her master's thesis.

The second part of Judith's dissertation belongs to the category of 'combined performance work', i.e., the study and optimisation of analysis methods used in ATLAS. Originally conceived as a qualification task to obtain ATLAS authorship, Judith worked on a method for identifying b-quarks (so-called 'b-tagging', a process in the category of 'flavour tagging') resulting from a specific decay throughout her PhD. Because of its novelty within the Collaboration, this work would eventually also result in a publication and turned into a major part of her dissertation:

> That was a real luxury. During the qualification task, we also published a [conference note] about it. So, it wasn't just a qualification task where you do something technical that maybe is integrated later on, but then you don't really contribute. For me, it really became a part of the dissertation, that was really cool, I was also lucky in a way.

According to Judith, it is not very common that qualification tasks result in contributions to ATLAS publications, or that students can base a solid part of their dissertation on these contributions. Judith was also "lucky" because the qualification task had resulted from a compromise. Initially, Judith had wanted to do a different project for her qualification task, which the group convenors rejected as being "too close to analysis" (interview Judith). Judith's dissertation eventually consisted of a general introduction to the theory and practice of high-energy physics at the ATLAS experiment, with a specific focus on the identification of Higgs-boson and b-quark decays; a description of the search for an unknown heavy particle, focusing on her contributions to the (already published) analysis; and a description of her work on b-tagging, some of which was documented for the first time in her thesis. Judith took longer than her initial project-based work contract to finish the dissertation, with a studentship funding the final year of her PhD. Shortly before graduation, she successfully applied for a post-doc fellowship at another German research institution.

Compromising to graduate

The story of another PhD student, based at a prestigious US-American research university, reveals that constructing a doable dissertation is not always a straightforward process. Sam had done some work on the CMS experiment as an undergraduate student and been recommended to a professor in the ATLAS Collaboration who later became her PhD advisor. For personal reasons, she decided to focus on projects that could be done remotely and stayed in the US throughout her PhD. Her qualification task was similar to Judith's in that it also studied the efficiency of a 'b-tagging' algorithm. Although only intended to earn her the status of an ATLAS author, this task developed into a project taking over the greater part of her PhD. It involved the production of particular data samples, which Sam started taking

responsibility for, serving as a 'software contact' for all groups requiring these samples. Approaching the final year of her PhD, Sam had contributed substantially to software maintenance and 'combined performance' work in ATLAS, but still needed an analysis to form the centrepiece of her dissertation. Like Judith, she chose a search that aligned with the research at her department and the knowledge she had gained from working on 'b-tagging'. A few months into working on a task that had been suggested by one of the conveners, she had to find out that this task had already been accomplished by another student: "They were nearly done and [the conveners] just hadn't kept track of who was doing what." To salvage her dissertation, Sam then joined a new search led by a post-doc from her local group:

> The reason I picked the supersymmetry search I'm working on isn't because it's the most compelling physics beyond the standard model search. It's because I want to graduate and it's a final state involving [b-quarks]. […] I originally picked one just based on the physics I knew and that search was too full. So, then I picked one that wasn't quite what I wanted but there was room for me.

Sam's story illustrates that although there is more than enough data and work for everyone in ATLAS, this work is not easily distributed. Despite the formal hierarchy of coordinator roles, group coordinators cannot simply assign tasks to individual researchers, only suggest. Moreover, the more exciting analyses may attract more researchers than there are tasks required for preparing a publication, and group conveners may sometimes "lose track" of who is doing what. This experience of a search being "too full" made Sam chose a smaller group doing a novel analysis, minimising the risk of redundancy, but at the loss of her own enthusiasm for the project.

At the time I interviewed her, Sam did not yet know whether the results of this analysis would be available in time for her graduation:

> I'm not 100 percent certain if the data will make it into my thesis, because I think we're going to unblind our results right around when I do my defence. But I already discussed with [my advisor] and some of the faculty from my committee and they decided that would be OK. Because I do have data in my other projects […].

Sam did not want to postpone her graduation because she had been accepted to a job placement program for the tech industry. Her dissertation eventually consisted of an introduction to LHC physics with the ATLAS experiment, a description of her work on b-tagging, a description of her software support work and a description of her contributions to the supersymmetry search and its expected results, based on simulated data.

The cases of Sam and Judith exemplify several elements of constructing doable dissertations in ATLAS. Students are asked to become members of local research groups as potential contributors, based on the skills they have demonstrated in earlier work. Even when students have an initial research interest, the project they end up working on emerges from (re-)aligning their interests to the tasks available in analysis groups and the current research priorities of the Collaboration. Constraints for constructing doable problems may arise on the level of the group (finding a task that contributes to collective projects but has not been done), on the level of the Collaboration (e.g., the internal distinction between analysis and qualification tasks, current research priorities), but also on the individual level (personal competences and preferences, graduating at a certain time or securing additional funding). There are certain expectations concerning the contents of a dissertation, but advisors and advisory committees do have some leeway in deciding whether a student's contributions to collaborative work meet those expectations. The two accounts also indicate the alignment work performed by students, advisors and other coordinators to construct doable problems at different stages of the dissertation. In the following two sections, I will zoom in on these practices and describe instances of alignment work between the individual, group and Collaboration-wide levels of work organisation in ATLAS. The first section illustrates how the alignment of collective and individual rhythms and resources creates opportunities for doable dissertations. The second section describes the alignment work shaping a student's contribution within an analysis team.

Arranging alignment with collective priorities

As mentioned above, the timelines of most research processes in ATLAS exceed the duration of a single dissertation. A major challenge thus consists in fitting students' individual contributions into Collaboration-wide schedules. Particular measurements and searches for new particles are planned years in advance, based on the anticipated data output, which so far has exceeded expectations. Some of these anticipated results have been defined as 'milestones' for the experiment, because they represent significant advances in particle physics. The need for a result to "go out" to secure the scientific credibility of the Collaboration, and the need of individual students to make substantial contributions and graduate, may conflict in these cases. Paul, a senior researcher based in France, mentioned this conflict while describing his own role as a coordinator of a high-profile analysis in the CMS Collaboration:

> Clearly the big analyses like ttH observation [the observation of the production of a Higgs boson and a top quark-antiquark pair], it's an analysis of a Collaboration of 4000 people, so it's a measure that you have to do for the outside world. But we can do this measure thanks to the work of the Collaboration, but mainly thanks to the work of the PhD students. This kind of big analysis, the analysis has to go out, independently of the timeline or the graduation for a PhD.

Paul addresses a tension that is inherent to the work in the Collaboration. Although research projects depend on the labour of many individual PhD students (and post-docs), collective research processes do not respect individual timelines such as work contracts or graduation dates. The more prestigious an analysis and the more researchers are involved in it, the more likely it is that it will take longer than the expected duration of a PhD to be completed. Conversely, when students join such an effort too late, their chances to make significant contributions before the results need to be 'out' are diminished.

Constructing doable dissertation projects thus requires advisors to plan carefully on behalf of their students. PhD advisors need to anticipate the opportunities when enough data have been collected and students may still be expected to make a significant contribution. Simon, a professor at a French research institute, described his strategic considerations when hiring a new PhD student in the following way:

> You have to see how much it will match with the expected publications. So, for example, I will take a PhD student in HH [studying collision events where a pair of Higgs bosons is produced] for Run 2. Because we finished to take data end of this year. So, he will start in 2018 and he will finish 2021. So, to justify the funding we say, next year he will improve the bbgammagamma [Higgs bosons decaying to two b-quarks and two photons] analysis and the year after I will do a combination with the other channel with CMS, do the interpretation with theorists. And I think one year later it [would] be problematic. So, one year before it's too early to start, to be really involved in the publication. One year later we have only a bit more data here but not significantly more than before, so, it's not sure there will be a publication.

The opportunity for a doable dissertation is created through aligning several organisational and infrastructural rhythms (Jackson et al., 2011). Simon was looking for a student to join his group right at the end of 'Run 2', the LHC's second data-taking period (2015 to 2018). During this time, the accelerator produced collisions at the unprecedented centre-of-mass energy of 13 TeV and ATLAS recorded an even higher number of collision events than anticipated. This provided ample data to be analysed over the second 'long shutdown' of the LHC, before data-taking would expectedly resume in 2021.[9] From the Collaboration's point of view, it is beneficial to prepare a publication only when the full dataset has been analysed. For this reason, starting much earlier than at the end of Run 2, when data-taking is still underway, would disadvantage a student. A doable dissertation furthermore needs to fit into the three-year-funding cycle for research projects structuring academic work across disciplines in European countries. In France, this three-year cycle also applies to the individual funding of PhD students (Louvel, 2012). The topic and the start of a PhD project need to be chosen in such a way that contributions to ATLAS publications can be expected within three years. Simon also explained

that whether such a pre-aligned dissertation succeeds will eventually depend on finding the right PhD candidate, who is capable and interested in completing the assigned tasks within the three-year timeframe.

Collective priorities may also require a student to align their work at a later stage, when they have already begun their analysis work. This, however, need not be detrimental to the student's interests. Corresponding to the cycles of data taking, there are times when fewer results can be expected, as researchers are asked to wait with publications until the full intake of data has been accomplished. As the following account from a PhD student at a German university illustrates, the Collaboration manages such droughts by "slipping in" smaller projects to maintain a steady flow of results and publications:

> We will have this [centre-of-mass energy of] 13 TeV for another year, and that's why [the ATLAS management] didn't want us to publish a whole lot of papers with last year's data, because then nobody would have time to add this year's data and publish based on those. [...] But of course we didn't want to say that we don't publish at all, so they said that there will be a few exceptions […]. And my analysis just somehow slipped in there, because [the working group conveners] also trusted that my local supervisors, my professor and my post-docs, they'd make sure that this won't take too long. (Gabriel)

Gabriel at the time was working mainly by himself, repeating an analysis that had already been done during Run 1 of the experiment. The promise of a timely result allowed Gabriel to begin working towards publication, even though his analysis was based on incomplete Run 2 data. The conveners of his working group chose it as one of the analyses that would fill the gap in the publishing cycle when most of the results based on the previous dataset (from Run 1) were already out, and data production for Run 2 was still underway. Gabriel's advisor negotiated a slightly later deadline than the group conveners had envisaged, but it was clear that the analysis should be out within the year. In this case, aligning the student's work to collective priorities was also beneficial to Gabriel, who could complete his main analysis earlier and start writing up the dissertation during the third year of his PhD.

Improvising alignment with group-level work

Once projects have been assigned and deadlines have been agreed on, the coordination of individual tasks among the group working towards a publication presents another challenge. Within the area of physics analysis, working group and sub-group conveners are expected to "keep track of who is doing what" (Interview, Sam) across analysis teams, while 'analysis contacts' oversee the coordination of tasks involved in a single analysis or publication project. These coordinators are responsible for integrating the work of individuals in collective research projects and thus play a vital role in constructing doable dissertations.

Cara was a post-doc at a US-American university at the time I interviewed her and served as an analysis contact in a search for a supersymmetric partner particle of the Higgs boson. She mentioned the example of a PhD student who came up with an ambitious, but only potentially doable idea:

> So, we knew exactly what we wanted to do with the paper, and I think everyone was on board with that. And then this student came out with his advisor and said, 'Oh, this is an improvement that we could add'. And we said, 'Great idea. But it's gonna be very challenging to have this in. You know, in the timescale that we need to have this in.' At this point we had a bunch of students that had to graduate on this analysis. We couldn't just have two more months to have a nice improvement on top. […] And the student worked for a very long time. He's a very good student. But it came to a point where it wasn't done yet. And we couldn't keep waiting for it.

Cara explains that constructing a doable research problem requires taking the group's interests into account. The analysis group as a whole had agreed to work towards a specific publication, and the tasks had been defined and distributed among the group members accordingly. The envisaged deadline reflected that other doctoral students on the team soon needed results to be able to graduate. Although this particular

student's proposal for an improvement seemed promising to the analysis contacts, it turned out to be too time-consuming. Resolving the dissonance between the group's schedule and the student's individual contribution required alignment work:

> What happened at the time is we came up with another thing for him to work on, that we kind of had decided to have in the paper. […] And the advisor wasn't totally convinced at first, he said 'but this seems like too much of a small thing for him to have ownership of'. So, we then put another twist in that project. […] So it was something that, you know, I had to sit down with the other analysis contact and we had to be like, 'OK, we need to come up with something that he can claim ownership of and it can't just be a small task'. (Cara)

In this case, alignment work involved identifying another contribution that could be added to the paper within the remaining time, and then negotiating the specific contents of that contribution with the student's advisor. Cara's account also highlights that doctoral students' contributions to collective papers should not only consist of "small tasks". Cara defined such "small tasks" in terms of their duration: "You can have to find a little project that is like, a week long, where they do a little study, but it ends up being like a sentence in the paper." A contribution intended to form part of a student's dissertation would have to be more substantial than such a "little study". This is because the dissertation, unlike a collective paper, will be attributed to the student as an individual. In this case, the analysis contacts achieved sufficient substance to satisfy the student's advisor by "adding another twist" to the task. Between these two constraints — the publication deadline and the expectation that the student's contribution should be worth having "ownership of"— the analysis contacts managed to construct a doable problem.

Disentangling alignment

Existing academic norms require a PhD dissertation to be an independent research contribution that can be attributed to a single author. This requirement seems to contradict the realities of collaborative research in high-energy physics, where students' work must be aligned with collaborative work and results are attributed to a collective. How are these contradictory requirements reconciled? In this final section describing my findings, I identify three strategies of individualising students' work, which are partly embedded in the practices described above. By way of these disentanglements, PhD students' work is temporally, qualitatively and formally distinguished from collaborative work and collective publications.

The *first disentanglement* is *temporal*. There is a time when a student does collaborative work within the group, and there is a time when a student is working on their dissertation. Typically, these phases are consecutive, as the "writing up"–phase takes place once the student's contributions to collaborative work are considered substantial enough to be converted into a dissertation.

> Actually, you're part of the Collaboration until — well, until you start writing up. ATLAS does not set that date, that's something for you and your advisor to agree on. […] Usually, when you're at the point of finishing a paper or an analysis, that's a good time, of course. […] There's a few rules in ATLAS, they think that they can dictate the students more, but in the end it's the professor who is responsible for what's in the dissertation. (Brian)

As this German PhD student explains, transitioning into the "writing up"–phase' may feel like leaving the Collaboration and (re-)entering a mode of work under the auspices of one's advisor. The main work context shifts back from the Collaboration to the local group. For students who spent some of their PhD on site at CERN, this transition would also involve a re-location to their home university.

Brian's account also highlights the persistent authority of PhD advisors. Several of my German interview partners indicated that students who run out of funding sometimes abandon an analysis before publication, or hand over to a younger colleague. This seems only possible if advisors may decide when a student's contribution qualifies for a dissertation, and if the contents of a dissertation are to some extent detached from the collective publication. Although originating in collaborative work, a dissertation is the only publication in high-energy physics that is always attrib-

uted to a single author, and normally also the only publication that a student will obtain single authorship of. According to one PhD advisor, it is their responsibility to ensure that a student's dissertation satisfies the criteria of independence and originality "despite" its origin in a collaborative effort:

> The publication normally isn't the same as the dissertation. […] Here's the issue: [The dissertation] is defined as an independent scientific achievement that has not been done by anyone else. This means that you need to ensure that despite the collaboration in the working group, the contribution of the PhD student is scientifically independent, and that it will pass as a doctoral dissertation. That is my job […], in the end, it is my responsibility to say, 'this is a doctoral dissertation'. (Philipp)

The *second disentanglement* between dissertations and collaborative work thus proceeds via a *qualitative* distinction between routine work and original or independent work, or between *small* and *big* tasks. As exemplified by Cara's story above, advisors and coordinators consider the requirement of scientific independence when negotiating a student's contribution to a collective paper. In Cara's story, the student's advisor actively ensured that the student's contribution would be worth "having ownership of". This indicates that the need for distinction is anticipated and criteria of independence and originality are already applied when constructing doable problems for students. Just *how* substantial, original, and independent a student's work will be seems to be a matter of negotiation. It also depends on the advisor's expectations and local conventions at the student's home institution. Although the advisors I interviewed gave some examples of actual and hypothetical contributions that students may 'write up' in their theses (such as developing a new algorithm or applying a new statistical method), the criteria remain situational. What tasks are worth doing for a student is decided individually, as part of the alignment work between research goals on the level of the Collaboration, group-level projects and the student's individual interests, skills and constraints.

A *third disentanglement* from collaborative work takes place on a *formal* level. ATLAS has a strict policy allowing only results that have passed the Collaboration's internal review process to be published or presented in public, but an exception is made for PhD dissertations (Charlton et al., 2009). For example, PhD dissertations may contain figures of results that have not been approved (yet), but these figures must not show the label reserved for official ATLAS results. In practice, this means that students need to re-do the plots they have produced for a publication and mark them as preliminary results or 'work in progress'. The writing up–phase allows students to pursue ideas and approaches that could not be realised within the working group or included in a paper. Here, students have the opportunity to create contributions that are genuinely their own, as long as their results do not contradict those of official ATLAS publications. Students are also allowed to present their work at smaller workshops and national conventions. However, since these contributions are not subject to the collective review process, they will not be considered to be official ATLAS results and typically not be referred to in other ATLAS publications. A formal and qualitative distinction is made between the work that students create as part of the collaborative process, and the work that is their own, but merely validated as part of a dissertation.

The formal distinction between collective publications and dissertations suggests that dissertations only have value on the individual level, as a means of obtaining an academic title. However, in some of my interviews, another function of dissertations was described, namely the documentation of the technical and methodological state of the art: "Usually (the PhD) was the best knowledge of the thing at this time. And at least in my lab, the part of the PhD which is a technical part is documented. […] So, it's a document which is always useful" (Interview Simon). This value of the dissertation as documentation originates in the process of disentanglement just described, which implies that the technical contributions and innovations of doctoral students are often not included in collective publications, or not described in detail. The "independent scientific achievements" (Interview Philipp) that are only documented in

dissertations may, however, be taken up in collaborative research projects later on.

Alignment work thus shapes dissertations in two distinct ways. Fulfilling the requirement that a dissertation consist of contributions to research in high-energy physics, dissertations result from aligning students' work with collective processes. The specific problem a student works on is a result of what can be made doable within an ATLAS group at this particular point in time. To fulfil the requirement that this contribution is an independent achievement, students and their advisors can take advantage of the overflows and excess produced through alignment work. The necessity of creating alignment with group-level and Collaboration-level processes excludes some ideas, contributions and approaches as outside the (momentary) scope of collective publications. This work can then be performed by students in a more independent manner as part of their dissertation. In this way, the content of a dissertation is created directly and indirectly through alignment work: Directly through the efforts of constructing doable problems, and indirectly through excluding some contributions from collective publications, such that they can be claimed individually.

Discussion – how are dissertations made doable?

My paper set out to investigate the tension between the notion of a scientific doctorate as an individual achievement, and the practical and organisational realities of collaborative research.

Based on an analysis of interviews with experimental particle physicists, my answer to the question *how doctoral dissertations are made doable in collaborative research* is two-fold: Dissertations are made doable by aligning students' work to collaborative research processes, as well as reflexively disentangling and proactively distinguishing students' contributions from collective research outcomes. Constructing dissertations in collaborative high-energy physics neither resembles the execution of a pre-conceived research project nor the post-hoc assembly of contributions into a written document but is best described as an emergent process of articulating and performing tasks that will result in distinguishable outcomes.

This process requires *alignment work* across levels of work organisation, performed by several different actors. Due to the long timespan of experimental research in high-energy physics, potentially doable contributions need to be identified in advance, considering the rhythms of instrumentation, data-taking, and planned publications, such that students' work is aligned with collective research goals on the level of the entire Collaboration. This type of alignment work is mainly performed by advisors, sometimes in coordination with group conveners. Constructing doable problems also requires an ongoing and flexible articulation of tasks that fit into group-level work. This type of alignment work is performed by group coordinators, together with students and their advisors. It requires flexibility and a capacity for improvisation when new ideas come up and individual tasks take longer than expected. On the part of students, it requires resilience when promising ideas are given up in favour of problems that are more consistency within the group's collective schedule.

To satisfy the requirement that dissertations showcase students' ability to do independent and original work, students' work is *temporally, qualitatively and formally distinguished* from the collaborative projects they have contributed to. "Writing up" dissertations is temporally separated from work on publications. What students "write up" are typically details and contributions that did not make it into collective publications due to constraints on time and space. Alignment work therefore shapes dissertations both directly, by constructing doable contributions for students, and indirectly, through defining some problems as outside the scope of collective publications, which can then be explored by students independently. The status of single authorship for dissertations formally distinguishes students work from collective publications. That dissertations are not listed as official ATLAS publications might signal that they are less epistemically significant or mere add-ons to collectively validated work. However, as described above, dissertations also provide a detailed documentation of analysis techniques and other technical contributions that is not

otherwise publicly available. In this sense, the need for distinction of dissertations from collective work that seems to devalue dissertations might also result in making them more valuable to the collective, as technical documentations and repositories for new approaches.

Concerning the role of advisors in large-scale collaborations, my findings indicate that PhD advisors continue to play a significant role in the construction of doable dissertations despite the formally hierarchical management of research processes. When hiring PhD students, advisors need to identify potentially doable problems, considering collective research priorities and expected publications. Advisors may take on an active role in creating tasks for their students within collaborative research processes, negotiating with coordinators, and advocating for their advisee's work. They may also support students with additional funding, so a student need not abandon an analysis prematurely. It is the advisor's and advisory committee's prerogative to decide that a student's research contributions are sufficient for graduation. Despite the broader range of potentially doable problems within a Collaboration and the availability of supervisors beyond the student's local group, the advisor's influence on dissertations is thus comparable to that of group leaders in laboratory-based research groups (cf. Delamont et al., 2000; Campbell, 2003). One plausible explanation is that advisors mediate between the organisational dimension of dissertation work (i.e., the local institution's requirements for the PhD) and the Collaboration. Since the requirements for an academic qualification are locally defined, local advisors remain the final authority on its contents.

Concerning the role of students, the personal and biographical dimension of constructing doable dissertations becomes most evident. Students may have personal preferences, such as where to live and how much time to spend on their PhD, which influence the process of constructing a dissertation, for example through a selection of tasks that allow remote work or earlier graduation. Students who pursue careers outside academia may opt for a more pragmatic approach and an earlier separation from collaborative research. Here, the wide range of research processes and potential contributions available in a Collaboration seems to allow students in high-energy physics more flexibility concerning the content and duration of their dissertations than their colleagues in laboratory-based research groups have, and a more active role in alignment work, particularly at the later stages of the PhD.

The effects of external constraints on dissertations, in particular project-based funding, may be mitigated through alignment work, depending on how flexible local funding arrangements are and whether additional sources of funding are potentially available. Students who enjoy greater personal and institutional resources might, in turn, find it easier to write dissertations that are both well-aligned with collaborative research goals and considered to be original contributions.[10] However, to answer the question of whether changes and differences in PhD programme structures or funding arrangements also impact the construction of dissertations in high-energy physics, a more systematic comparison of these practices (either across time or across research groups subject to different arrangements) would be required.

Experimental high-energy physics certainly presents a boundary case of collaborative research. Some of the alignment processes described above will only exist in large-scale research collaborations, where collaborators and constraints beyond a student's immediate group directly influence the doability of individual research problems. Furthermore, alignment work between group-level and individual-level work is virtually absent in most of the humanities and many social science disciplines, where solitary work and single-authored publications are the norm. However, in humanities and social sciences, changing expectations such as an increased demand for journal publications are also transforming the formal requirements on PhD students' work, with cumulative dissertations and co-authored articles becoming more acceptable. Investigating how alignment work shapes dissertations, such that they fulfil the requirements of academic institutions as well as those of the respective epistemic community, would thus be insightful for STS research interested in the dynamics of contemporary research more generally. In particular, the specific mecha-

nisms of distinguishing doctoral students' work and ensuring its independence and originality deserve closer scrutiny, given the observable trend towards more and larger research collaborations across disciplines. My analysis shows that dissertations emerge over time as a product of alignment work, based on the resources and constraints provided by the infrastructural, organisational and biographical dimensions of scientific work. They also show that a dissertation's content, format and epistemic value are shaped by formal and qualitative criteria of distinction, which are proactively applied in alignment work. This second observation indicates that beyond establishing a coherent collective (Boisot, 2011; Galison, 2003; Knorr Cetina, 1995), large-scale research collaborations also need to develop mechanisms for distinguishing individual contributions, which might be just as significant in shaping epistemic practices.

Acknowledgements

I am indebted to our interview partners, my advisor Martina Merz and my colleagues Sophie Ritson, Barbara Grimpe, Daria Jadreškić and Markus Tumeltshammer for helpful comments and discussions of this paper throughout the research and writing process. My research has benefited from collaboration with ATLAS physicists Peter Mättig and Christian Zeitnitz, who helped me approach interview partners, negotiated access to internal documents, and validated my findings. Earlier versions of this manuscript have been presented at the sociology of science seminar at TU Berlin and the doctoral students' seminar at the Department of Science Communication and Higher Education Research, University of Klagenfurt. I would also like to thank two anonymous reviewers and the editor, Alexandra Supper, for feedback and suggestions that have substantially improved the article. This research was funded by the Austrian Science Fund (FWF) [I 2692-G16].

References

Biagioli M (2003) Rights or Rewards? Changing Frameworks of Scientific Authorship. In: Biagioli M and Galison P (eds) *Scientific Authorship: Credit and Intellectual Property in Science*. New York, NY: Routledge, pp. 253–279.

Birnholtz JP (2006) What Does it Mean To Be an Author? The Intersection of Credit, Contribution, and Collaboration in Science. *Journal of the American Society for Information Science and Technology* 57(13): 1758–1770. DOI: 10.1002/asi.20380.

Boisot M, Nordberg M, Yami S and Nicquevert B (eds) (2011) *Collisions and Collaboration: the Organization of Learning in the ATLAS Experiment at the LHC*. Oxford: Oxford University Press.

Boisot M (2011) Generating Knowledge in a Connected World: the Case of the ATLAS Experiment at CERN. *Management Learning* 42(4): 447–457. DOI: 10.1177/1350507611408676.

Bruyninckx J (2017) Synchronicity: Time, Technicians, Instruments, and Invisible Repair. *Science, Technology, & Human Values* 42(5): 822–847. DOI: 10.1177/0162243916689137.

Campbell RA (2003) Preparing the Next Generation of Scientists: The Social Process of Managing Students. *Social Studies of Science* 33(6): 897–927. DOI: 10.1177/0306312703336004.

Charlton D, Gianotti F, Hinchcliffe I, et al. (2009) ATLAS Policy leading to approval of Physics results: Version 3.3. ATLAS Collaboration.

Charmaz K (2006) *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis.* London: Sage.

Corbin JM and Strauss AL (2008) *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. 3rd ed. Los Angeles, Sage.

Degn L, Franssen T, Sørensen MP and de Rijcke S (2018) Research Groups as Communities of Practice—a Case Study of Four High-Performing Research Groups. *Higher Education* 76(2): 231–246. DOI: 10.1007/s10734-017-0205-2.

Delamont S and Atkinson P (2001) Doctoring Uncertainty: Mastering Craft Knowledge. *Social Studies of Science* 31(1): 87–107. DOI: 10.1177/030631201031001005.

Delamont S, Parry O and Atkinson P (1997) Critical Mass and Pedagogic Continuity: Studies in Academic Habitus. *British Journal of Sociology of Education* 18(4): 533-549.

Delamont S, Atkinson P and Parry O (2000) *The Doctoral Experience. Survival and Success in Graduate School: Disciplines, Disciples and the Doctorate*. Washington, D.C: Falmer.

Deterding NM and Waters MC (2021) Flexible Coding of In-depth Interviews: a Twenty-first-century Approach. *Sociological Methods & Research* 50(2): 708–739. DOI: 10.1177/0049124118799377.

Dippel A (2019) Die Schraube, der Marder und der Bug: Zeitlichkeit und Materialität im Experimentieren am Beispiel ethnografischer Feldforschung über Physik. *Schweizerisches Archiv für Volkskunde* 115(1): 7–26. DOI: 10.5169/seals-842278.

European Committee for Future Accelerators (2015) *Memorandum on the Evaluation of Experimental Particle Physicists*. Available at: http://cds.cern.ch/record/2014643/files/ecfa-291_ECFA-HEP-evaluation.pdf (accessed 2 August 2018).

Fochler M, Felt U and Müller R (2016) Unsustainable Growth, Hyper-Competition, and Worth in Life Science Research: Narrowing Evaluative Repertoires in Doctoral and Postdoctoral Scientists' Work and Lives. *Minerva* 54(2): 175–200. DOI: 10.1007/s11024-016-9292-y.

Fujimura JH (1987) Constructing `Do-able' Problems in Cancer Research: Articulating Alignment. *Social Studies of Science* 17(2): 257–293. DOI: 10.1177/030631287017002003.

Fujimura JH (1996) *Crafting Science: A Sociohistory of the Quest for the Genetics of Cancer*. Cambridge, Mass: Harvard University Press.

Galison P (2003) The Collective Author. In: Biagioli M and Galison P (eds) *Scientific Authorship: Credit and Intellectual Property in Science*. New York, NY: Routledge, pp. 325–353.

Graßhoff G and Wüthrich A (eds) (2012) *MetaATLAS: Studien zur Generierung, Validierung und Kommunikation von Wissen in einer modernen Forschungkollaboration*. Bern: Bern Studies in the History and Philosophy of Science.

Hackett EJ (2005) Essential Tensions: Identity, Control, and Risk in Research. *Social Studies of Science* 35(5): 787–826. DOI: 10.1177/0306312705056045.

Jackson SJ, Ribes D, Buyuktur A and Bowker GC (2011) Collaborative Rhythm: Temporal Dissonance and Alignment in Collaborative Scientific Work. In: *Proceedings of the ACM 2011 conference on Computer supported cooperative work - CSCW '11*, Hangzhou, China, 2011, pp. 245–254. ACM Press. DOI: 10.1145/1958824.1958861.

Jones GA, Kehm BM and Shin JC (eds) (2018) *Doctoral Education for the Knowledge Society: Convergence or Divergence in National Approaches?* 1st ed. 2018. Knowledge Studies in Higher Education. Cham: Springer. DOI: 10.1007/978-3-319-89713-4.

Knorr Cetina K (1995) How Superorganisms Change: Consensus Formation and the Social Ontology of High-Energy Physics Experiments. *Social Studies of Science* 25(1): 119–147. DOI: 10.1177/030631295025001006.

Knorr Cetina K (1999) *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, Mass: Harvard University Press.

Larivière V (2012) On the Shoulders of Students? The Contribution of PhD Students to the Advancement of Knowledge. *Scientometrics* 90(2): 463–481. DOI: 10.1007/s11192-011-0495-6.

Laudel G (2001) Collaboration, Creativity and Rewards: Why and How Scientists Collaborate. *International Journal of Technology Management* 22(7/8): 762. DOI: 10.1504/IJTM.2001.002990.

Laudel G and Gläser J (2008) From Apprentice to Colleague: the Metamorphosis of Early Career Researchers. *Higher Education* 55(3): 387–406. DOI: 10.1007/s10734-007-9063-7.

Louvel S (2012) The 'Industrialization' of Doctoral Training? A Study of the Experiences of Doctoral Students and Supervisors. *Science & Technology Studies* 25(2): 23–45. DOI: 10.23987/sts.55274.

Mangematin V (2001) Individual Careers and Collective Research: Is There a Paradox? *International Journal of Technology Management* 22(7/8): 670. DOI: 10.1504/IJTM.2001.002984.

Merz M and Sorgner H (2020) Komplexe Organisationen zum Sprechen bringen. In: Donlic J and Strasser I (eds) *Gegenstand und Methoden Qualitativer Sozialforschung*. 1st ed. Leverkusen: Verlag Barbara Budrich, pp. 51–67. DOI: 10.3224/84742326.04.

Milojević S (2014) Principles of Scientific Research Team Formation and Evolution. *Proceedings of the National Academy of Sciences* 111(11): 3984–3989. DOI: 10.1073/pnas.1309723111.

Möllers N (2017) The Mundane Politics of 'Security Research'. *Science & Technology Studies* 30(2): 14–33. DOI: 10.23987/sts.61021.

Müller R (2014) Postdoctoral Life Scientists and Supervision Work in the Contemporary University: A Case Study of Changes in the Cultural Norms of Science. *Minerva* 52(3): 329–349. DOI: 10.1007/s11024-014-9257-y.

Ochs E and Jacoby S (1997) Down to the Wire: the Cultural Clock of Physicists and the Discourse of Consensus. *Language in Society* 26: 479–505.

Rushforth A, Franssen T and de Rijcke S (2019) Portfolios of Worth: Capitalizing on Basic and Clinical Problems in Biomedical Research Groups. *Science, Technology, & Human Values* 44(2): 209–236. DOI: 10.1177/0162243918786431.

Shrum W, Genuth J and Chompalov I (2007) *Structures of Scientific Collaboration*. Cambridge, Mass: MIT Press.

Torka M (2018) Projectification of Doctoral Training? How Research Fields Respond to a New Funding Regime. *Minerva* 56(1): 59–83. DOI: 10.1007/s11024-018-9342-8.

Traweek S (1988) *Beamtimes and Lifetimes: The World of High Energy Physicists*. Cambridge, Mass: Harvard Univ. Press.

Whitley R, Gläser J and Laudel G (2018) The Impact of Changing Funding and Authority Relationships on Scientific Innovations. *Minerva* 56(1): 109–134. DOI: 10.1007/s11024-018-9343-7.

Wuchty S, Jones BF and Uzzi B (2007) The Increasing Dominance of Teams in Production of Knowledge. *Science* 316(5827). American Association for the Advancement of Science: 1036–1039. DOI: 10.1126/science.1136099.

Ylijoki O-H (2016) Projectification and Conflicting Temporalities in Academic Knowledge Production. *Teorie vědy / Theory of Science* 38(1): 7–26.

Notes

1 To distinguish the organisations running high-energy physics experiments from research collaborations in a general sense, the former will be referred to as "Collaborations" with a capital C.

2 This research has been conducted in the context of the interdisciplinary Research Unit *The Epistemology of the Large Hadron Collider* and its sub-project 'Producing Novelty and Securing Credibility: LHC Experiments from the Perspective of Social Studies of Science'.

3 For a more detailed description of our approach to interviews, see (Merz and Sorgner, 2020).

4 While the experiences of PhD students reflect different models of graduate education in Germany and the US (Jones et al., 2018), these differences become less significant as soon as US students have passed their course requirements, become members of research groups and start working on their dissertations. At this point, doctoral students orient their work towards the Collaboration, and the various groups in which their projects are embedded become the main work contexts for US-American and German students alike. My interviews and analysis have focused on this phase of the PhD for the US-American students.

5 I thank Sophie Ritson, who conducted two of these interviews, for pointing out their relevance to me.

6 Participants were approached via email, informed about the research interests of the project, and provided with a copy of the consent form in advance (asking for the permission to record the interview, describing the use and storage of data, and the rights of the interviewee to remove consent and end the interview at any time).

7 https://atlas.cern/discover/collaboration, accessed November 30, 2021. For a detailed description of the (early) ATLAS collaboration from a management studies perspective, including the design of the detector and the scientific aims of the experiment, see (Boisot et al., 2011).

8 All names have been changed to preserve interview respondents' anonymity. Quotes from interviews originally conducted in German have been translated by the author.

9 Due to the delays incurred during the COVID-19 pandemic, the start of Run 3 eventually had to be postponed to 2022.

10 Regarding this observation, a limitation of my study is that most of my interview respondents are members of relatively influential 'local groups'. PhD students who are members of groups with fewer resources and connections might be less integrated in their Collaboration and experience less support for their work overall, resulting in very different challenges for constructing doable dissertations.

‘If You’re Going to Trust the Machine, Then That Trust Has Got to be Based on Something’: Validation and the Co-Constitution of Trust in Developing Artificial Intelligence (AI) for the Early Diagnosis of Pulmonary Hypertension (PH)

Peter Winter

School of Sociology, Politics and International Studies (SPAIS), University of Bristol, United Kingdom / peter.winter@bristol.ac.uk

Annamaria Carusi

Interchange Research; and Department of Science and Technology Studies, University College London (UCL), United Kingdom.

Abstract

The role of Artificial Intelligence (AI) in clinical decision-making raises issues of trust. One issue concerns the conditions of trusting the AI which tends to be based on validation. However, little attention has been given to how validation is formed, how comparisons come to be accepted, and how AI algorithms are trusted in decision-making. Drawing on interviews with collaborative researchers developing three AI technologies for the early diagnosis of pulmonary hypertension (PH), we show how validation of the AI is jointly produced so that trust in the algorithm is built up through the negotiation of criteria and terms of comparison during interactions. These processes build up interpretability and interrogation, and co-constitute trust in the technology. As they do so, it becomes difficult to sustain a strict distinction between artificial and human/social intelligence.

Keywords: Artificial Intelligence, technology development, early diagnosis, trust, collaboration, validation

Introduction

In this article, we consider the central question of trust in Artificial Intelligence (AI) technologies for medical diagnosis. As AI becomes increasingly integrated into existing workflows and implemented to support diagnosis and treatment, clinical experts will find it difficult to understand how AI algorithms have been validated: this is where the problem of trust arises (Scheek et al., 2021). For many clinical and technical experts (such as computer and data scientists), trust is a matter of explainability and transparency of the algorithm, or the justification of the outputs of an algorith-

mic model (Tonekaboni et al., 2019; Barda, 2019; Cutillo et al., 2020). One way to broach these issues of trust is through the development of guidance that aims to foster responsible and trustworthy applications of AI (Bærøe, 2020). Examples include AI4People (Floridi et al., 2018), Asilomar AI principles and the Independent High Level Expert Group on Artificial Intelligence (AI HLEG) set up by the European Commission (2019). Altogether, guidance and initiatives associated with developing trustworthy AI have in common ethical frameworks (principles and guidelines) to improve morally good outcomes. In particular, the AI HLEG argue that AI should be designed and developed in ways that build in interpretability from the start through assessment lists – a work process which assumes that trust can be accomplished through a rigorous application of pre-identified evaluation criteria.

Yet, despite these efforts, levels of acceptance of healthcare AI remain low: several studies have come to the conclusion that there is a lack of trust among clinical experts towards these kinds of technologies, which as a consequence, has led to low acceptance and use (Topol, 2019; Strohm, 2019; Cabitza et al., 2020; Sreedharan et al., 2020; Nagendran et al., 2020). Topol (2019) shows that the lack of data and proof is eminently to blame – indeed, he argues that there is a lack of research investigating the validation and readiness of Machine Learning (ML) models in clinical settings, prompting distrust on the assumptions underpinning many validation tests that have been assessed in the laboratory. Taking this idea of the validity of ML models one step further in the context of AI development in biology and medicine, Littmann et al., (2020) states that it is collaboration itself which leads to AI research that is more scientifically valid, in that it is more correct and reproducible. One could similarly compare such thought on collaboration with the work of Elish and Watkins (2020: 6) in Science and Technology Studies (STS) who take stock of the 'sociotechnical' engagements between different human experts and their algorithms and the work of building trust in new technology. We claim that an important source of trust is the collaboration between AI developers and clinical experts, and we aim to show how forms of collaboration support the collective construction of validation and interpretability, which ultimately grounds trust in the technology.

This article aims to give a detailed account of how collaboration informs the co-emergence of trust and validation in a setting where three AI algorithms are being developed for use in real-world clinical settings. In addition, we show how validation that looks towards real-world settings is not something that occurs at the endpoint of the development process. Instead, it occurs *throughout* the development process and is built into the application. This occurs primarily through collaboration with clinical experts from the initial stages of development and concrete practices of repurposing healthcare data. While validation may often be viewed as something that comes at the endpoint of algorithm development, the grounds upon which validation is based starts at the outset of collaboration and continues through the development process across contexts, practices and technologies. This approach is particularly relevant to our discussion on the practical efforts and meaningful selection of criteria for comparison in AI development. For example, as will be explored later, clinical experts who participate in the practice of selecting, testing, and refining criteria (e.g. labels, codes, or variables) for comparison are the ones who are able to interrogate the outputs of validation, whereas a clinician who has not been involved in that process may not comprehend or interpret the outputs in the same way and may open up the potential for "blind trust" in the technology (Gaube et al., 2021: 1).

The subject of trust has a wider relevance for social scientists interested in collaboration and development of new (AI) technologies, and will provide critical insight in an area imbued with high claims, promise and technological expectations (see Rajpurkar et al., 2017, 2018; Perry, 2017; Ming, 2018). This article will draw from the overlapping fields of STS and Computer Supported Cooperative Work (CSCW) on collaboration in the context of technology development, to treat collaboration as a set of work practices that are invoked at particular times for building trust towards algorithms. The first part of the article considers the relationship between trust and collaboration in general. The next part of the

article deals with the basic notions related to validation in the technical and social science literatures. The focus then shifts to the background of the study and its associated methods are described. The central section of the article opens by showing how the first three dimensions of AI development (*Querying Datasets*, *Building the Software,* and *Training the Model)* play their role in the validation process. In the following sections, we take this further to consider the different kinds of collaborative practices for building trust in the validation process, specifically reflecting on the practical efforts and meaningful selection of criteria for comparison. After this salient presentation of data, we move on to discuss how validation is a collaborative endeavour, foregrounding our position that validation starts at the outset of collaboration and continues through the development process across contexts, practices and technologies. The final section of the article concludes with the notion that AI requires constant monitoring and refinement which are a far cry from providing a 'technological fix' for problems in society and healthcare in particular.

Trust and collaboration

The topic of trust has received a great deal of attention in research into how technologies are deployed to support tasks and decisions. How to trust the outputs of technologies is particularly acute when their development and use crosses across different expertises and disciplines. In these contexts, trust emerges through particular collaborative tasks between people with different expertise, as seen in multidisciplinary teams (MDTs) who jointly make diagnosis or treatment decisions (Van Baalen et al., 2017; Van Baalen and Carusi, 2019). A similar line of thought is followed by Elish (2018: 369) who argues that trust in AI technologies can also be built by including or "looping" in stakeholders (such as clinicians) from the very beginning and throughout the development process. Such collaborations are mediated by a 'local champion', a clinical expert involved in the development of the technology who does "vital trust-building work" throughout the hospital and the wider clinical community (Strohm, 2019: 58). The field of CSCW has developed a substantial and highly relevant body of work that explores trust in various contexts, and frequently focuses on the role of trust as a key aspect of collaboration between people, but also in relation to processes and technologies which directly impact how expert judgements are made (Fitzpatrick and Ellingsen, 2013). Here, trust is commonly conceptualised as linked to features of interpersonal relationships between people and often remains implicit with familiarity/lack of familiarity being a basis for trust/mistrust in human-human interactions (e.g., Jirotka et al., 2005; Carusi, 2009). Trust may also be conceptualised as generated 'in action', built up in some form of situated or contextual practical engagement of a work routine, often in contexts when people have a responsibility to build trust in new technology (e.g., Clarke et al., 2006a, 2006b; Oudshoorn, 2008; Kuutti and Bannon, 2014; Papangelis et al., 2019).

When interpersonal and practical trust-building becomes a mediator for the development of new technologies (i.e., algorithms), people become deeply embedded in technical and non-technical processes, and other temporalities. Here, technology development is characterised as a complex and active form of sociotechnical production with experts being influenced by a variety of parameters, pressures, and politics that make up the social construction of complex technologies (Mackenzie, 1990; Laurent and Thoreau, 2019). Mackenzie (1990), in particular, demonstrated how the accuracy of a technology can be constructed and shaped by both technical engagement and the perspectives of social actors involved in its process of development. In contexts of collaboration, these interactions may be seen as the often 'invisible work' that goes into technology development - although the people who perform such interactions are quite visible, the work they do is relegated to the background (Star and Strauss, 1999: 20). According to Star and Strauss (1999: 10), one important form of work which is often invisible in making technologies work is the concept of 'articulation work' – a type of work that happens after breakdowns or unanticipated contingencies as it is "work that gets things 'back on track' in the face of the unexpected". Pallesen and Jacobsen (2018: 173) suggest articulation work can also be understood as the work of coor-

dinating between different sites of the experiment in collaborative research, in addition to being a salient concept for situated problem-solving. In other words, experts can bridge social worlds and thus 'mesh together' these very different social worlds to get work 'done'. Taking this approach, articulation work could also be seen (and needed) to support a type of 'sociotechnical infrastructure' that scaffolds medical and organisational work (Star, 1999). Star and Strauss' (1999) notion of 'invisible work' has also started to become an important analytical tool for understanding data work in healthcare (Bonde et al., 2019; Bossen et al., 2019; Bossen and Piras, 2020). In this context, invisibility may refer to the invisible nature of collaborative work performed by actors around practices of data; a process which plays a key role in ensuring the truthfulness and correctness of data to support clinical practice (Bjørnstad and Ellingsen, 2019).

Together, we might see these as two kinds of trust that complement each other: the interpersonal trust experts acquire when interacting with experts from different disciplinary backgrounds on the one side, and the practice-orientated trust experts acquire when they participate in developing the tool, technology or instruments. We seek to convey the idea that both types of trust work are forms of invisible work because they too often remain implicit and hidden in scientific accounts of validation. Taking this into account, the concept helps us to identify and surface the invisible work of trust, as well as also to become attentive to, how the mundane work of collaborative research and data practices are generative of validation.

Our research suggests that trust in healthcare AI is co-constituted by collaborators from throughout the development process, and that this underpins validation. This point about AI and the fact that trust, validation and the technical characteristics themselves are co-constructed is significant in a broader debate where AI tends to be seen as a 'technological fix' able to solve multiples issues, including the problem of trust in the ability of institutions to solve complex problems. According to Katzenbach (2019), AI is accepted in particular areas, like healthcare, transport, and social media, as a kind of technological fix for solving specific problems. For example, Katzenbach (2019) recognises that autonomous vehicles can help reduce traffic accidents, and sees the potential of using AI for detecting misinformation and hate speech online. Specifically, however, he argues that this talk about 'AI fixing things' is misleading because it obfuscates the importance of human labours and social relations that these technologies are built upon. For this reason, the objective of this article brings to light not only the technology's inherent technical properties, but also the role social processes such as collaboration play in the construction of trust in AI development.

With this article, we want to bring trust and validation together: we propose that collaboration plays a part in the generation and maintenance of trust relationships (between people and technologies) which directly impact how expert judgements are made and accepted. In the next section we suggest that the focus on validation as a technical solution to trust has left underappreciated the collaborative, social aspects of the validation process. These are the focus of the social science literature on validation, which proposes that validation is as much about people's social interactions with technology and each other as it is about any technical feature of the technology. As we will later show, the process of selecting and negotiating the criteria that go into evaluating the technologies, and considering it 'validated' are useful for building in trust in judgements made about the technology and its outputs.

Validation

Technical literature

In the technical literature, algorithms are required to pass some form of quality control in the form of a validating test (or set of tests or criteria) in the demand for trusted or trustworthy systems (Alpaydin, 2016; Tonekaboni et al., 2019; Barda, 2019; Cabitza et al., 2020). These tests are often based on a comparative performance of the technology, comparing its performance with other performances considered to be a 'gold standard', such as a human expert producing confirmed findings in a diagnostic report (e.g., Gulshan et al., 2016; Esteva et al., 2017; Rajpurkar et al., 2018; Annarumma et al., 2019). Accordingly, the perfor-

mance of the algorithm against the gold standard is often expressed in statistical terms (e.g., 'accuracy', 'sensitivity', 'specificity') and by some kind of expert who is able to make a judgement about its performance, such as having high predictive accuracy, for example Rajpurkar et al's (2017, 2018) CheXNeXt algorithm being able to make accurate predictions at a level that "exceeds the average radiologist performance" (Rajpurkar et al., 2017: 2). However, such claims can prompt considerable scepticism and distrust across scientific and medical communities, as was the case with radiologist Oakden-Rayner (2017, 2018) who initiated a critique of the CheXNeXt model (along with the help of Rajpurkar and his team) to verify the accuracy of its predictions. The conclusion of that critique was that the images had not been labelled correctly, nor did the labels reflect clinical practice having the potential to produce meaningless predictions (Oakden-Rayner, 2018).

Oakden-Rayner's critique contains important epistemological questions that deserve consideration: questions about how comparisons can be made (especially between algorithms and human experts), and how data is labelled (who labels the data, who inspects the data and whether experts with relevant clinical experience are considered). Labels and codes or criteria for comparing performances come to matter greatly when it comes to validation because they are based on the so-called 'ground truth' of features that the algorithm has learned in the training data – the labels, annotations, or codes in this instance constitute the ground truth or ground for comparison (e.g., Gulshan et al., 2016; Esteva et al., 2017; Oakden-Rayner, 2018; Cabitza et al., 2020; Scheek et al. 2021).

In addition, such validation is commonly represented as consisting of two isolated approaches: internal and external (Topol, 2019; Cabitza and Zeitoun, 2019; Nagendran et al., 2020). The focus of most healthcare AI development is a form of internal validation, carried out within computer science laboratories and tested on retrospective datasets. External validation is usually referred to as the clinical validation of AI systems and tested on prospective datasets of entirely new data ('in the wild') (Cabitza and Zeitoun, 2019). As Nagendran et al., (2020) point out, most validation studies are tested on retrospective datasets only, with the number of prospective datasets tested in real-world clinical settings extremely low (only 6 out of 81). Cabitza and Zeitoun (2019: 161) also distinguish between 'statistical', 'relational', 'pragmatic' and 'ecological' validity. Statistical validity is claimed by them to be objective, 'intrinsic' and 'essential' to the system. However, relational, pragmatic and ecological validity consider the context of the algorithm in one or other way. For instance, either with respect to usability or pragmatic consequences (for example, how data is handled), or with 'ecological' consequences, (for example, with respect to work settings). Nonetheless, however technical these different forms of validation may seem to social scientists, they are important concepts in understanding how experts consider validation as a technical practice and something that comes at the endpoint of technology development.

Social science literature

Social science literature on model validation provides us with the capacity to investigate validation practices and trust practices *in the making*. This literature on validation in science has provided us with a sustained analysis of the confusions and uncertainties that accompany validation (Randall and Wielicki, 1997; Shackley et al., 1998; Küppers and Lenhard, 2005; Sundberg, 2006; Winsberg, 2010; Morrison, 2015). Science policy scholars have produced in-depth analyses of the validation of chemical or environmental models, showing the extent of uncertainties and disagreement on the model's validity, relevance and bias (Oreskes et al., 1994; Oreskes, 2004; White et al., 2010). A major reason for this would seem to lie in the nature of how evidence is subjected to different standards of 'proof' and different ways of thinking about proof in different sectors – a far cry from the supposed 'objectivity' of models or the quantitative nature of empirical data (Oreskes, 2004). For an STS view on this matter, see Mackenzie's (2001) work on *Mechanizing Proof* and how experts negotiate data to be worked with and construct 'proof' of the correctness of a program or software design. 'Proof' that the model or software is in absolute sense 'correct' or 'dependable' is very much a social process of iteration (e.g.,

doing testing, returning to the nature and use of data, redefining the test, repeating the test, finding the design fault, and so on) (Mackenzie, 2001: 43). At some point in this cycle, experts come to an understanding that their software is often reasonably reliable because of how humans interact with the technology and by testimony to a 'trustworthy agent' to whom they may turn (Mackenzie, 2001: 307). Other STS literature has also made the same points about the validation of models while exploring different factors affecting scientists' reasoning and choices (Sundberg, 2006; Böschen, 2009, 2013; Carusi et al., 2012; Carusi, 2014, 2016; Thoreau, 2016; Boullier et al., 2019; Laurent and Thoreau, 2019). Böschen (2009, 2013), in particular, has distinguished between what he calls four 'evidential cultures' and two of these are most relevant in this context. First that a 'restrictive evidential culture' rests primarily on experimental methods in controlled laboratory settings using models to establish causality, but often orient scientists to particular drawbacks of the phenomenon being tested (e.g., having limited available data on which to test the comparability of results). Second, that a 'holistic evidential culture' may be combined with other tests and different forms of knowledge to evaluate the phenomena. This time there is less interest in capturing causal phenomena and more of a move towards capturing complex elements of an ecosystem or the larger system of people's lives and cosmologies. This holistic culture chimes with the notions of pragmatic validity and ecological validity of other studies discussing validation (Cabitza and Zeitoun, 2019).

However, the most important contributions of STS researchers in the analysis of validation for this article derives from research on the implementation of an AI algorithm for the early detection of sepsis ('Sepsis Watch') (Elish, 2018; Elish and Watkins, 2020; Sendak et al., 2020). Concerned with the validation of Sepsis Watch, these authors present validation as an integral component for establishing the trust of clinicians and point out how existing epidemiological or 'gold standard' definitions of sepsis were found to be inadequate at predicting the risk of sepsis in real-time cases in the clinical setting (Elish and Watkins, 2020: 18). What they found in the clinical setting was a negotiation and refinement of criteria and variables where trust had been manifested in the process. Trust of the sepsis algorithm was by no means dependant on some technical neutrality of the model, but a series of key activities that brought clinicians and statisticians together, promoting a potent combination of empirical observation, refinement and repair. The emphasis on real-time validation and the ongoing collaborative work of clinicians and statisticians shows that the algorithm came to be trusted through technical demonstrations of efficacy rooted within social relationships.

The central argument of these articles is that validation is as much about people's social interactions with technology as it is about any technical feature of the technology; it is inextricably socio-technical. The technology is not an inert thing passively being acted upon until it reaches a point where it is deemed 'validated'. Rather, it actively mediates interactions and fosters interpersonal trust and practice-orientated work, and through these, the creation of scientific knowledge and technical results, such as its accuracy (Mackenzie, 1990) or proof (Mackenzie, 2001; Laurent and Thoreau, 2019). The criteria according to which validation will be assessed are not pre-defined, but emerge during the process (Carusi, 2014). This makes for a technology that is more likely to be accepted by potential users, and actually embedded in their real-world context.

Taken together, these studies recognise the importance of validation on clinical experts' trust of models. However, there is still much work to be done in investigating the process of validation. This is especially the case when validation is associated with AI models in healthcare, which iteratively involves contesting and selecting criteria or classifications for comparison. This article does just this by paying deeper attention to the voices involved in the process of validation and making explicit the conditions under which their reasoning operates. It extends the previous STS literature by showing how the collaborations that give rise to AI co-produce the criteria that act as grounds for comparison which underlie validation practices.

Our study: AI in the clinic

Our study explored the development process of three AI technologies for the early diagnosis of pulmonary hypertension (PH). PH is a rare, progressive and life shortening lung disease that is often diagnosed at an advanced stage. Diagnosis for PH is assisted by a myriad of testing technologies (such as right heart catheterisation, blood tests and medical imaging). However, such technologies are often deployed too late in the disease process, and therefore may yield a late diagnosis with limited treatment outcomes or poor markers for prognosis (Kiely et al., 2013). Because of this problem of late diagnosis, clinicians and researchers around the world are looking to AI as a route to an earlier diagnosis for PH in order to bring about better life expectancy and quality of life for patients (e.g., Kiely et al., 2019; Swift et al., 2020)

The first AI being developed is a 'screening' algorithm to detect patients 'at risk' of PH trained on Hospital Episode Statistics (HES) data. The second algorithm being developed is an 'imaging' algorithm which uses Magnetic Resonance Images (MRI) of the heart in order to detect disease features of PH. The third algorithm being developed is the 'biomarker' algorithm to detect signs or signatures of PH in blood samples trained on biomarker data related to PH, to be included in the screening algorithm. At the time of our study, all three algorithms were at the proof-of-concept stage with the intention of being deployed and used in the context of a UK PH Referral Centre at a major NHS Teaching Hospital. Thus, we are studying three proof-of-concept projects in the early development phase, highlighting the invisible work of the sociotechnical infrastructure (Star, 1999), ideally for organising, supporting, and elevating the next steps of each project in order to facilitate their route into clinical trials.

Methods

This article is based on qualitative interviews with seven participants involved in developing three proof-of-concept AI algorithms for the early diagnosis of PH. Participants included: two PH clinicians, one consultant PH nurse, one radiologist, one computer scientist, one data scientist, and one biomedical scientist to fully take into account the sea of discourses, ideas, scientific criteria, and concepts that shape validation and trust in AI development. In total there were six face-to-face interviews in workplace offices. One of these interviews was a joint interview conducted with the computer scientist and radiologist both working together to develop the imaging algorithm. Data was collected between 17/05/2019 - 22/10/2019. Recordings were transcribed and uploaded to NVIVO 12 to help manage, code and analyse themes that emerged from the transcripts. Taking an inductive approach to thematic analysis (Braun and Clarke, 2006), the theme of validation explicitly emerged across the group of research participants with decisions involving validation understood to be inherently tied to trust: interpersonal interactions and various computer supported practices involved in validation demanded consideration.

Our fieldwork was conducted on three developing algorithms (mentioned above). These algorithms were small scale pilot projects or proof-of-concept projects being developed to show the viability of AI to tackle challenges of early diagnosis, projecting hopes of a 'technological fix' (Katzenbach, 2019). As such, the projects involved small numbers of people, and often just two or three people were the main developers and sometimes one person would be working on two, or even all three of the algorithms. Accordingly, our interview numbers are not high. This will affect the generalisability of our findings. We might say that the proof-of-concept nature of the projects we studied and our own study are limited in similar ways. Despite the relative intimacy of our research domain, our research produced some important insights concerning how these collaborations operated to establish trust and to set criteria for validating the performance of the AI applications.

Results

Laying the ground for validation: querying datasets, building software, and training the model

Validation is often represented as the final stage of technology development (Alpaydin, 2016). However, a significant amount of interpersonal and practice-orientated trust work, and a large proportion of training/testing activities occur ear-

lier in the development process. These opportunities to build trust in the technology are crucial for technology development, but often remain 'invisible' and go unnoticed and unaccounted for, relegated to the background (Strauss and Star, 1999; Oudshoorn, 2008). Here, one needs to think of the previous three work activities (*Querying Datasets*, *Building Software*, and *Training the Model)* that take place before a formal validation phase, a perspective that shows how each activity lays the ground for validation. Whilst we have argued elsewhere how these three activities help to demystify the algorithm and 'de-trouble' transparency issues (Winter and Carusi, *forthcoming)*, we argue in this article how each activity can also be said to present interpersonal and pragmatic opportunities for building trust towards the final validation experiment. These activities lay the foundation for how trust and validation co-emerge in the sociotechnical infrastructures of diagnostic work through their negotiation and refinement of criteria and are explained in the following.

Querying datasets is concerned with how external or internal datasets are curated. It brings into play questions around how the datasets have been labelled or coded and by whom and whether they have sufficiently included clinical experts, which may lead to imprecise datasets and to inaccurate tests. As Oakden-Rayner (2017, 2018) reminds us, dataset quality is crucial in relation to the way in which criteria such as labels on medical images lay the ground for validation, namely how the labels are used to validate its performance. In our study, a radiologist developing the imaging algorithm echoed this concern by highlighting the difference in quality between datasets that have been collected prospectively and retrospectively:

> When evaluating very large cohorts with thousands of patients, people will question, unless it's a prospective study, 'how do they know that person actually had that condition?' And if it's from a clinical database, how was that really done? If all patients went through a multidisciplinary team meeting with recorded outcomes, that's very robust. But when data is collected retrospectively without an MDT diagnosis or similar assessment this can leave uncertainty as to the validity of the data.
> (Participant 4, Radiologist)

The quote expresses the radiologist's concern about the quality of prospective datasets and retrospective datasets. For the radiologist, if a label can be traced through to a prospective study in which the radiologist is either directly involved in the labelling of data, or is familiar with the experts who have participated in its labelling, the dataset is considered "very robust". However, if labels have come from a retrospective study where the labelling is not first-hand, the labelling process is less certain, because the radiologists are not sure of the processes used by the experts in applying the labels, asking for example, "how do they know that person actually had that condition?", and "how was that really done?". The radiologist's trust is anchored in previous interactions with expert members of the MDT and serves as the basis for the radiologist's perception of the quality of the dataset, and in this sense, is a form of interpersonal trust (Jirotka et al., 2005; Carusi, 2009; Van Baalen and Carusi, 2019).

Despite the lack of certainty regarding how labels were applied in a retrospective dataset, these datasets are used for technology development. Retrospective datasets provide the raw material for reconstructing and interpreting diagnoses, as seen in the quote below:

> Retrospective data labelling has its limitations and it's going to require us to go back into it and look at the scans and make a retrospective diagnosis on some cases because it comes from a number of different acquisition methods, different radiographers, and in the case of derived measurements different software [...] So coding is very, very important […] it needs work for people to go back in and classify patients retrospectively sometimes.
> (Participant 4, Radiologist)

Consequently, our focus on practice calls attention to the lengthy struggles clinical experts may face with research materials to reconstruct them in a way that facilitates their diagnosis, for example through labelling or coding key features of interest and aligning them with their own clinical experience and local work practice. This treatment of retrospective datasets demonstrates how practical work of querying and relabelling features on images is required for the radiologist to trust the

dataset. CSCW scholars will recognise this as one type of data work that takes place to elucidate the emerging requirements for management and work system design (Bossen et al., 2019). Indeed, the complexity of 'repurposing' data to serve secondary purposes beyond the practices of its initial use (Bonde et al., 2019; Bossen and Piras, 2020) challenges our radiologist to work with conflicting qualities or ambiguities of data and the activities needed to ensure the 'correctness' of data (Bjørnstad and Ellingsen, 2019). Importantly for this article, the radiologist's reconstruction of diagnoses through negotiation and refinement of diagnostic criteria reminds us of how trust can actually be engendered in a practical situation (Clarke et al., 2006a; 2006b; Oudshoorn, 2008; Papangelis et al., 2019) and moreover, calls attention to their 'articulation work' (a form of invisible work) that "gets things back on track" when unanticipated situations arise (Star and Strauss, 1999: 10).

Building the software means the building of a classification software. It is an activity that continues to lay the ground for validation because it draws on the experience of clinical experts who participate in the negotiation or refinement of appropriate criteria (e.g., diagnostic labels/codes or other variables) for software building. As part of this process, clinical experts start learning how the software arrives at its classifications, how the software is assessed, and how to participate in future refinements of criteria.

Training the model takes the last activity further by inviting clinical experts to assess the training outputs of the algorithm in an imagined clinical context. Clinical experts are included in the critical assessment of the software's outputs and participate in discussions about whether outputs are relevant or plausible, using their clinical experience to change or refine existing criteria included as features of the model.

However, as we will see in the second half of the article, this process of establishing what could count as criteria for comparison is never static or fixed (Carusi, 2014). Rather, it continues throughout the whole of the development process and sets up the algorithm for a formal validation test. The next section continues to look at this process, particularly focusing on the method of internal validation and the collaborative work involving AI developers and clinical experts in setting up the criteria for comparison between the algorithm's results on different or unused datasets. Building on the previous discussion about the negotiation and refinement of labels in 'Laying the Ground for Validation', we investigate how criteria and variables under retrospective conditions have to be retemporalised for clinical contexts accordingly by bridging or 'meshing' the nexus between external validation and internal validation.

Different forms of validation

As we have previously discussed, there are two main steps to validation: testing against retrospective datasets and testing prospectively (Topol, 2019; Cabitza and Zeitoun, 2019; Nagendran et al., 2020). The focus of most AI development is on retrospective datasets, which is a form of internal validation, carried out within (mostly) computer science laboratories in universities or industries. External validation is the testing of the AI application against entirely new data, 'in the wild', as it is not the data in the same dataset as the algorithm was trained on. In our study, AI developers invited clinicians into the laboratory to assess the performance of the algorithm on the retrospective datasets: work that bridged the gap between internal and external validation and allowed both AI developers and clinical experts to gain an understanding of *how* validation was carried out. The involvement of the clinical experts in bridging the gap between internal and external validation shows how knowledge can be co-produced and how the knowledge from the laboratory needs to be related to the real-world (Boullier et al., 2019). The bridging between two different settings for validation purposes continued the process of establishing appropriate criteria for comparison, showing how criteria continue to be negotiated and refined in ongoing iterations of tests (Carusi, 2014), and in the process, how trust and validation co-emerge. This bridging process begins with the clinical expert's first encounter with the results of the first internal validation test.

Internal Validation

Here we join the computer scientist and radiologist in an interview about their method of internal validation for the imaging algorithm in the labo-

ratory. Internal validation here involves the computer scientist and radiologist pursuing the goal of setting up a meaningful comparison between the algorithms results on unused segments of the imaging dataset, and then later on refining the criteria for comparison between the algorithm and the radiologist**.** When asked how they went about validation for the imaging algorithm, the computer scientist replied:

> We use cross-validation. Basically, we partition the data set into ten parts, ten partitions, then we use nine of them for training and one for testing and then just rotate. So that is one of the classical methods in machine learning to validate a method when we have limited number of samples in a dataset. I think in the beginning it really gives us quite a good estimation of how much the algorithm can achieve compared to the current approach of manual segmentation.
> (Participant 5, Computer Scientist)

In their approach to validate the imaging algorithm, the computer scientist states that they are using the method of "cross-validation". The computer scientist explains how this specific validation process is informed and dominated by the separation of datasets into nine training sets ("we use nine of them for training") and one testing set ("one for testing") which are then rotated ("then just rotate"). The comparison that this approach relies on is with "manual segmentation". That is, it is with the diagnostic labels that have already been applied to the dataset and queried by clinical experts (as described above). When asked about how they arrived at this judgement of how good the validation was and who was involved, the computer scientist highlights the important part the labels play in providing 'ground truths':

> The data are all labelled with ground truths […] When we try to predict the label of the individual patient on that test set, we're doing the prediction pretending the label is not available. Then we use the ground truth labels to compare the predicted label and then we compute an error, so if this error is small then that's high accuracy.
> (Participant 5, Computer Scientist)

First, this quote shows how each label on a dataset of medical images constitutes a 'ground truth' – a process established earlier in the article by the radiologist's querying of datasets (e.g., the relabelling of features). Second, the performance of the imaging algorithm in arriving at the 'correct' detection of PH-related features is compared with the clinical labels embedded in the dataset. On the basis of this comparison, the size of the error between the computer's performance and the labelled dataset is computed. This becomes the metric of how well the algorithm performs ("so if this error is small then that's high accuracy"). This establishes the statistical validity of the algorithm (Cabitza and Zeitoun, 2019). Importantly for this article, this excerpt from the interview highlights how the objective of AI development is to build models that are accurate *enough* and highlights how accuracy is negotiated (Mackenzie, 1990) which for Laurent and Thoreau (2019: 165) is 'part and parcel' of technology development. Moreover, the picture of what can be deemed equivalent to what becomes clear in practice (Carusi, 2016): labels/codes become essential criteria and underpin judgements about the accuracy of validation tests (Scheek et al., 2021). Importantly, for this article, internal validation tests provide further opportunities to mediate practice-orientated trust between collaborators (Clarke et al., 2006a, 2006b; Oudshoorn, 2008; Kuutti and Bannon, 2014; Papangelis et al., 2019). The next section will illustrate how this trust building deepens, paying particular attention to how clinical experts generate meaning with respect to the labels/codes or variables in the model – a process which is particularly useful when it comes to the 'interpretability' of outputs and continues the bridging between internal and external validation.

Interpretability

In the previous sections, we showed how clinical experts query the quality of datasets. We argued how clinical experts play a crucial role in establishing the quality of its curation: this helps them better understand the criteria that they are dealing with (e.g., labels/codes), builds practice-orientated trust work, and lays the ground for validation tests (e.g., cross validation). We also argued in the previous section that clinical experts bridge between internal and external validation. The next section will illustrate in detail the action of this bridg-

ing, highlighting the role clinical experts play in interpreting the performance of the imaging and screening algorithms in the validation tests. In fact, this is a process which shows how internal validation is inscribed with a view from clinical experience, however implicit that view might be.

Designing an AI system with interpretability in mind from the start opens up opportunities for not only practical interpretation and interrogation (and questions around what the output is, or how the output is made to matter in different situations), but also for building trust. The quote below from the computer scientist developing the imaging algorithm highlights why this practical interpretation matters:

> I actually have an end user over there to ask me questions […] like Participant 4 to give some suggestions on how to visualise the features and so on. I think that's something fresh to me and that also inspires me to write like interpretable machine learning […]. I think those kinds of challenges are real only when you start to interact with the community. So only when I interact with a domain expert, with an end user, then the question will come in.
> (Participant 5, Computer Scientist)

The computer scientist highlights how their collaboration with the radiologist brings in a variety of benefits: 1) questions that the computer scientist may not have thought of; 2) interpretability, that is, a kind of translation between the performance of the algorithm and the context of the domain expert; and 3) reality in terms of the uses to which it could be put in the radiologist's world. Working with the radiologist is a chance for considering the outputs of the algorithm in a clinical context and thereby highlights the radiologist's potential for bridging between internal and external validation – thus continuing to highlight the articulation work of clinical experts who 'mesh together' otherwise divided tasks, users and different systems (i.e., internal vs. external) and remains invisible because of its implicit nature (Star and Strauss, 1999; Pallesen and Jacobsen, 2018: 173). Nevertheless, interpreting algorithmic outputs is essential for the ongoing validation of the software, as iterative querying and questioning by clinical experts anchors the performance of the algorithm to their real-world context of use. In turn, this connection between meaning and use lays the ground for comparison for validation tests and engenders trust.

We observed similar processes of establishing the interpretation of algorithm outcomes for real-world contexts through querying and interrogation in the development of the screening algorithm. The main collaboration here was between a bioinformatics company and clinician. Again, we see the importance of the algorithm's outcomes being something 'real' in the clinicians' world ("From Participant 1's point of view, they're like 'this is something that I can relate to, I can relate to that number": Participant 2, Data Scientist). According to the data scientist in this case, the process of selecting the most appropriate ICD-10 codes was for "making sure that your comparative group are somehow relevant", which "is really important" and that without this clinical insight into how patients are diagnosed in the real-world clinic means that "the model at the end is just so trivial". Together, the consequences of this interpretation work in the development of the imaging and screening algorithms iteratively feed into the *Training of the Model*. This is because the results of any validation test feeds into further refinement of the model of the domain enacted in the algorithm - a further example of the ongoing 'articulation work' of the software (Star and Strauss, 1999). It is a process which occurs in the ongoing cycle of iterations for testing models (Carusi, 2014) and an integral aspect of all software development (Mackenzie, 2001). It also highlights the role of clinical experts engaging in linking or 'meshing' otherwise divided social worlds of the laboratory and the clinic, and how an understanding of criteria (labels/codes/variables) are negotiated within these laboratory settings with a view to their clinical application. Again, this practice-centred approach adds to the formation of a context of trust where the broader context is taken into account (Kuutti and Bannon, 2014).

Trusting questions

Trust, as we have seen in the sections above, is threaded implicitly throughout the whole of the development process and consists of a set of

interpersonal interactions (Jirotka et al., 2005; Carusi, 2009; Van Baalen and Carusi, 2019) and practical engagement of the technology in question (Clarke et al., 2006a, 2006b; Oudshoorn, 2008; Kuutti and Bannon, 2014; Pallesen and Jacobsen, 2018; Papangelis et al., 2019). However, trust is spoken about explicitly when it comes to some final validation test or method. From our interviews, clinical experts considered trust and validation as closely associated. Clinicians, in particular, considered validation a proxy for trust and to be on the terms of those whose trust is required for acceptance:

> Do you trust the information that you've been given and how much validation do you require? And I think that's the important thing. That element of trust. […] If you're going to trust the machine, then that trust has got to be based on something. So, it can be blind faith. So maybe some people are fairly evangelical about things, you've got blind faith that actually that machine is really good, so I'll just go with that.
> (Participant 1, Clinician)

As the quotation from the clinician reveals, trust is evidently directed at validation. Ultimately what it means if something is 'validated' is that it is trustworthy. For this clinician, validation is open-ended, since they are aware that the demands of validation could vary ("how much validation do you require?"). However, it is still possible to distinguish between requiring some form of validation and "blind faith", which they also associate with being "evangelical" about machine learning. The clinician then goes on to talk about an attitude of curiosity which comes into play in understanding what is meant by validation:

> A lot of people have got a certain degree of curiosity about 'do I really believe that?', 'is that really true?' and it's like that when you see a patient. You can take everything a patient tells you at face value or you can try and interrogate that information to see whether or not it's right. And you need to recognise sometimes that you are not very good at extracting information. Sometimes the patient is not very good at giving you a clear story so there's always those sorts of balances, checks in the system.
> (Participant 1, Clinician)

The clinician highlights the key role of "curiosity" and draws an interesting analogy between themselves as a clinician working to understand what a patient tells them, and trying to understand what the algorithm's outputs are telling them. The clinician does not necessarily take the patient's descriptions or statements at face value - not because they do not believe or actively mistrust the patient - but because there are many reasons why there may be lack of clarity in a patient's account. For example, there may be many confusing factors in a patient's experience of the condition ("sometimes the patient is not very good at giving you a clear story"). A general sense of being curious about the patient's presentation of a condition is an essential component of diagnosis. The diagnostic puzzle brings out the non-judgemental but epistemically driven attribute of curiosity - though we might also see in this a kind of constructive questioning or scepticism.

For this clinician, curiosity is also about how to make sense of an algorithm. Among the questions they ask themselves are: "do I really believe that?", "is that really true?". In this way, the clinician extends their professional attitude towards patients to the outputs of machine learning: that is, the clinician does not simply and straightforwardly believe it. Much like the patient, the clinician is unlikely to take the algorithm's output at "face value"; but is instead likely to "interrogate that information". This clinician also recognises that just as there is sometimes a lack of clarity in the accounts of patients, there may be a lack of clarity in the outputs of the machine. Crucially it involves both an interpretation of what the patient/algorithm is 'saying', *and* a questioning of its truth, a potential withholding of belief. Here too the collective and collaborative aspect of clinical practice is at play, and the clinician refers to how the checks and balances of other colleagues often work in these situations to raise questions so that diagnosis can be revised and rectified ("there's always those sorts of balances, checks in the system"). The clinician's suspicion towards model outputs on the one hand whilst also acknowledging the different skills required for interpretation chimes with the findings of other studies on computational models and validation (Randall and Wielicki, 1997; Sundberg, 2006).

This questioning and interrogation also leads to a refinement of the whole validation process. This is clear in the quotation below from the data scientist's collaboration with the same clinician, describing their processes of checking the performance of the algorithm:

> What can I solve myself by looking at the data and then what can I raise to them to say 'this looks kind of strange?' I think that's what's hugely valuable […] if you can have a clinical expert to be part of the development procedure I found that to be just priceless because they and the team they saw all of the things that we did, they saw when we were worried, they saw when we were like 'no this actually looks okay now' and I think you can't put a price on the value of that in growing the trust.
> (Participant 2, Data Scientist)

Here again we have an indication of how important the collaboration is: for mutual understandability linked to mutual agreement regarding how it should be tested, for joint 'ownership' of the AI application, and for establishing trust as practice-orientated (Clarke et al., 2006a, 2006b; Oudshoorn, 2008; Papangelis et al., 2019). Nevertheless, for the data scientist, making changes and refinements of the algorithm's variables (after questioning and interrogation) resembles the beginning of a journey through which the clinician acquires the understanding that will eventually allow them to 'get it'. As the data scientist, stated:

> You need your key clinical champions to be part of it and to say 'I've been on this journey with this development and I get it, and I've contributed and I can see where it's going', I think it's so important.
> (Participant 2, Data Scientist)

As the quotation from the data scientist reveals, clinicians act as "clinical champions", thereby opening doors to the broader community. One example is participant 1 whose act as a champion for the screening algorithm and PH community makes sure that the bioinformatics team who are helping them develop the algorithm have access to the clinician's PH networks of clients and partner hospitals they need. The clinician's role as clinical champion is articulated in the quote below from the data scientist:

> Participant 1 invited us to an advisory board where we had about 8 of the different specialists from the 14 centres all across the UK. We presented the algorithm to them, we said 'this is what we're doing'. We invited their comments, we invited a lot of criticism to be honest and it was a very productive discussion and at the end we said: 'we're excited about this, but we need more information and evidence to be sure about it. Would you like to be involved as a collaborator?' and they said 'yes'. So, they've signed a letter for us, which we then gave to NHS digital.
> (Participant 2, Data Scientist)

In the words of Strohm (2019: 35) the clinician as a 'local champion' acts as a mediator of trust and forms a bridge between the computer/data scientists, the AI application in development, and the broader community. There would be very low prospects for external validation without this.

External validation with unseen data in the real-world

The whole process of development is geared towards external validation. These validations require an "independent cohort" (Participant 3, Biomedical Scientist) or "virgin population" (Participant 1, Clinician), that is, the AI algorithm is required to be tested in real-world clinical conditions on prospective data (Topol, 2019; Cabitza and Zeitoun, 2019; Nagendran et al., 2020). For the team that we interviewed developing the screening algorithm, external validation is a "really important" step towards identifying those patients who could be asked to come into the clinic for further tests, in the hope of arriving at earlier diagnosis. However, this process is highly challenging because it involves real people: not only data points in a data set, but people whose data has not been definitively classified and labelled in a clean dataset of 'ground truths', and also who may have a deadly disease:

> What we haven't done yet is a prospective validation which I think is really important. And I would say of all the patients today 'who do we flag as the most high risk?' and then follow them up, so wait a bit of time, so wait for six months, wait for twelve months and see 'did they actually get

> diagnosed or referred?' 'Are these people actually being managed?'
> (Participant 2, Data Scientist)

In the quote above, the data scientist highlights the limitations of data collection in clinical contexts for developing analyses and validating the screening algorithm. That is to say, the data scientist is hard pressed to tell us how they will actually steer this course:

> I think it's a slightly trickier validation because they could still be people who would be PH patients but just don't get referred in that period of time. But I think it's still useful. One of the things that we discussed which we haven't really ironed out yet is we could actually invite them to a clinic and some of the specialists could say 'oh I'd be really happy to bring them into a clinic if you flagged them'. But I think again we would need to think very carefully about what we would need to do in order to operationalise that and also again the risks and all the ethics in all that.
> (Participant 2, Data Scientist)

The data scientist's desire for prospective validation on the one hand, whilst also fearing what this challenge might indicate on the other, is perhaps not unusual and chimes with the desire for prospective validation in clinical contexts (Cabitza and Zeitoun, 2019). The data scientist, in particular, argues that the data collection corpus is still reliant on patients being referred to specialty centres, as initial referral patterns are constantly changing and have different patterns in different regions. Furthermore, in some instances referrals provide only general ICD-10 codes or basic patient information that almost inevitably fail to capture the holistic understandings that can be found in a MDT diagnosis that are critical for establishing ground truth labels and dataset credibility. For all the researchers we interviewed, it was critical to move onto prospective validation, so that ultimately a much broader range of patients could be screened for PH. This process would need to be constantly re-anchored into real-time outputs and closely examined and refined by diagnoses from actual clinical practice. For Elish and Watkins (2020: 50) this validation process is a type of 'feedback loop' which combines clinical expertise and machine learning prediction and, in effect, gives us an idea of how validation will occur in clinical practice as an accomplishment of sociotechnical work.

Discussion

The close collaboration between AI developers and clinical experts throughout the development process brings the AI application out of the laboratory into the 'real world' of its clinical users. This bridging between the laboratory and the clinic brings meaning to AI applications, making their key features interpretable in their context of use. This bridging also affects the way in which the performance of the algorithm is assessed and validated. In internal validation this occurs through comparing the outcomes of runs of the software against different segments of the dataset queried and labelled by clinical experts. Comparisons can be carried out in a number of different ways, and according to multiple different criteria; there are different grounds for comparison for different domains, uses and practices. Identifying which criteria should be used and how depends on *who* is making the assessments and for which purposes, and crucially on how the outcomes are found to be relevant or not, given meaning or not in the context of use (recall that clinician 1 when interpreting the performance of the algorithm in a validation test remarked: "this is something that I can relate to, I can relate to that number"), and how the algorithm is questioned and interrogated by clinical experts, according to their expertise and experience. Finding the grounds for comparison goes hand in hand with the interpretability of the algorithm's outcomes in that context. Like Elish and Watkins (2020), our analysis aligns with the concept of 'articulation work' as a form of invisible work which is necessary during innovation (Star and Strauss, 1999). In our analysis, this process of articulation begins by coordinating and embedding clinical experts into the work of AI developers, and once embedded, participate in iterative activities to get things "back on track", such as the querying of datasets and bridging between internal and external validation (Star and Strauss, 1999: 10).

Bridging performs a social and an epistemic function. We saw that the involvement of both AI developers and clinical experts results in the algorithm gaining meaning, interpretability and comparability in the real-world context of use. We then saw how trust and bridging between laboratory and clinic are in a somewhat circular relationship, which is however not a vicious circularity. One of our clinician participants put the central trust question thus: "if you're going to trust the machine, then that trust has got to be based on something". The 'something' it is based on is built up through collaborative practices in every stage of the development process, and even draws on close interpersonal interactions as part of their every day clinical practice. This makes the validation of the AI application's software jointly produced by everyone in the collaboration, a co-production that establishes the criteria for assessment of the performance of the AI algorithm in a way that is epistemically accessible for all involved, if not absolutely, in ways that are relevant to each expertise and use. Besides this crucial epistemic role, co-production plays a crucial social role in establishing sufficient acceptance for the validation process to proceed to the next stage, with the user-collaborators becoming 'local champions' of the application for the broader community (Strohm, 2019).

Criteria for assessment of the algorithm or criteria for judging its outputs have come to occupy a particularly significant position within the validation context. This has been related by many to be the 'evidential culture' that is required for making credible decisions, where criteria are not defined by a standard of proof or regulatory organisation but emerge in the social dynamics of co-production. It is a collection of human judgements about similarities between objects of interest where people use their experiences of a phenomena in the real-world and anticipate whether it is comparable, or sufficiently similar to tests such as computational models that predict risks (Böschen, 2009, 2013). By focusing on the process of how collaborators establish grounds for comparison which is the basis for validation and for trust, this article offers a novel contribution to this existing focus. Although the Sepsis Watch research allows us to understand how the development and interpretation of criteria for comparison can take place in clinical contexts (Elish, 2018; Sendak et al., 2020; Elish and Watkins, 2020), our research yields an understanding of the crucial role of co-producing the grounds for comparison on which assessments are based, which precede the AI system reaching clinical settings. This article does so by considering the process by which this is achieved where details and nuances matter and remain underexplored.

Even though the grounds for comparison are often expressed statistically, which metrics and which variables go into the statistics are determined in the context of collaboration, depending on their relevance, usefulness, etc. In addition, statistical validity is dependent on a number of important further trust practices in the domain: the trust of the clinicians who query the dataset, in the diagnostic practices of other clinicians; the trust of the computer scientist in the clinical experts querying the data set (a kind of trust by proxy); the trust that each of the collaborators have in the abilities and expertise of the other. Given the social and epistemic complexity of trust practices, it is clear that statistical validity is never standalone, but rests on the shifting sands of these practices. It is hard to find any feature of AI that is an intrinsic, 'objective' feature of an application, as even the statistical assessments are highly relational (cf. Cabitza and Zeitoun, 2019). Far from AI being a technical fix for problems faced in healthcare settings, an AI system that works in these contexts is produced through a complex interplay of social, epistemological and technological factors, that require sustained attention to bring to the surface invisible work and sociotechnical infrastructures underpinning the development process. Research into developing healthcare AI needs to broaden its focus to encompass the clinician's situated participation in sociotechnical work environments when it comes to processes of trust building. Doing this will allow us to reach a better understanding of the details in how trust is engendered, and indeed to assess the extent to which these trust practices are robust, given the real-world tasks that these intelligent systems but perform.

Conclusion

Following the process of developing AI applications for supporting diagnosis in clinical settings shows validation to be neither purely technical, nor simply the final step of the development process in a formal validation phase. Establishing what counts as validation occurs through an iterative and piecemeal process, that brings together people with multiple different expertises, and the real-world contexts in which those expertises are used to make complex decisions. The grounds for comparing the performance of the algorithm with other performances, so that it can be both meaningfully interpreted and evaluated by all those involved with it emerge simultaneously with the developing AI. In this way, epistemic accessibility is built into the algorithm, and traced into it. This allows trust to be built into the system and co-constituted by collaborators throughout the process, and not by some 'end point' realisation. This trust is multi-faceted, as it is engendered by interpersonal, multi-expert collaboration (e.g., computer scientist, data scientist, biomedical scientist) and practical interactions with the technology before it even gets to a formal validation phase. Rather than trust being produced by validation, trust supports meaningful validation. This is not a backward pipeline with the arrows simply going in the opposite direction: it is a form of trust which works in a complex system intertwining social, epistemic and technological aspects. AI development needs to get better at attending to this intertwinement.

Acknowledgements

We would like to thank the Wellcome Trust for the Seed Award that funded this research [Grant number: WT/213606 awarded to Dr. Annamaria Carusi]. We would also like to thank our collaborators who took part in this research, without whom this research would not be possible. Furthermore, we would like to thank the reviewers for their insightful comments on the manuscript, which helped us to strengthen our argumentation. Finally, we would like to thank all the participants who attended the AI in the Clinic Network Event (26/03/2021) whose helpful comments enhanced our thoughts for this article.

References

Alpaydin E (2016) *Machine Learning: The New AI*. Cambridge: MIT Press.

Annarumma M, Withey SJ, Bakewell RJ, Pesce E, Goh V and Montana G (2019) Automated Triaging of Adult Chest Radiographs with Deep Artificial Neural Networks. *Radiology* 00:1-7.

Asilomar AI Principles (2017). Principles Developed in Conjunction with the 2017 Asilomar Conference [Benevolent AI 2017]. Available at: https://futureoflife.org/ai-principles (accessed 07.01.2020)

Barda AJ (2019) *Design and Evaluation of User-Centered Explanations for Machine Learning Model Predictions in Healthcare*. PhD thesis, University of Pittsburgh, US.

Bærøe K, Miyata-Sturm A and Henden E (2020) How to Achieve Trustworthy Artificial Intelligence for Health-care. *Bulletin of the World Health Organisation* 1;98(4): 257-262.

Bjørnstad C and Ellingsen G (2019) Data Work: A Condition for Integrations in Health Care. *Health Informatics Journal* 25:526–535.

Bonde K, Danholt P and Bossen C (2019) Data-Work and Friction: Investigating the Practices of Repurposing Healthcare Data. *Health Informatics Journal* 25: 558–566.

Bossen C, Pine KH, Cabitza et al. (2019) Data-Work in Healthcare: An Introduction. *Health Informatics Journal* 25(3): 465-474.

Bossen C and Piras EM (2020) Introduction to the Special Issue on Information Infrastructures in Healthcare: Governance, Quality Improvement and Service and Service Efficiency. *Computer Supported Cooperative Work (CSCW)* 29: 381-386.

Böschen S (2009) Hybrid Regimes of Knowledge? Challenges for Constructing Scientific Evidence in the Context of the GMO-debate. *Environmental Science and Pollution Research* 16(5): 508–520.

Böschen S (2013) Modes of Constructing Evidence: Sustainable Development as Social Experimentation - The Cases of Chemical Regulations and Climate Change Politics. *Nature and Culture* 8(1): 74–96.

Boullier H, Demortain D and Zeeman M (2019) Inventing Prediction for Regulation: Modelling Structure-Activity Relationships at the US Environmental Protection Agency. *Science & Technology Studies* 34(2): 137-157.

Braun V and Clarke V (2006) Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3(2): 77-101.

Cabitza F and Zeitoun JD (2019) The Proof of the Pudding: In Praise of a Culture of Real-World Validation for Medical Artificial Intelligence. *Annals of Translational Medicine* 7(8):161.

Cabitza F, Campagner A and Balsano C (2020) Bridging the 'last mile' gap between AI implementation and operation: 'data awareness' that matters. *Annals of Translational Medicine* 8(7): 501.

Carusi A (2009) Implicit Trust in the Space of Reasons and Implications for Technology Design: A Response to Justine Pila. *Social Epistemology* 23(1): 25-43.

Carusi A (2014) Validation and Variability: Dual Challenges on the Path From Systems Biology to Systems Medicine. *Studies in History and Philosophy of Biological and Biomedical Sciences* 48:28-37

Carusi A (2016) In Silico Medicine: Social, Technological and Symbolic Mediation. *Humana-Mente Journal of Philosophical Studies* 30: 67-86.

Carusi A, Burrage K and Rordríguez B (2012) Bridging Experiments, Models and Simulations: An Integrative Approach to Validation in Computational Cardiac Electrophysiology. *American Journal of Physiology-Heart and Circulatory Physiology* 303: H144–H155.

Clarke K, Hardstone G, Hartswood M, Proctor R and Rouncefield M (2006a) Trust and organisational work. In: Clarke K, Hardstone G, Rouncefield M and Sommerville I (eds) *Trust in Technology: A Socio-Technical Perspective*. Dordrecht: Springer, pp. 1-20.

Clarke K, Hughes J, Martin D et al. (2006b) 'It's about time": Temporal Features of Dependability. In: Clarke K, Hardstone, G, Rouncefield M and Sommerville I (eds) *Trust in Technology: A Socio-Technical Perspective*. Dordrecht: Springer, pp. 105-121.

Cutillo CM, Sharma KR, Foschini L et al. (2020) Machine Intelligence in Healthcare – Perspectives on Trustworthiness, Explainability, Usability, and Transparency. *Npj Digital Medicine* 3(47): 1-5.

Elish MC (2018) The Stakes of Uncertainty: Developing and Integrating Machine Learning in Clinical Care. *Ethnographic Praxis in Industry Conference Proceedings* 2018: 364–380.

Elish MC and Watkins EA (2020) Repairing Innovation: A Study of Integrating AI in Clinical Care. *Data & Society*. Available at: https://datasociety.net/library/repairing-innovation (accessed 06.01. 2020).

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-Level Classification of Skin Cancer With Deep Neural Networks. *Nature* 542: 115–118.

European Commission (2019) *Ethics Guidelines for Trustworthy AI: High-level Expert Group on Artificial Intelligence*. Directorate-General for Communications Networks, Content and Technology. Publications Office, European Commission. Available at: https://data.europa.eu/doi/10.2759/177365 (accessed 09.11.2019).

Fitzpatrick G and Ellingsen G (2013) A Review of 25 Years of CSCW Research in Healthcare: Contributions, Challenges and Future Agendas. *Computer Supported Cooperative Work* 22: 609-665.

Floridi L, Cowls J, Beltrametti M et al. (2018) AI4People – An Ethical Framework for a Good Artificial Intelligence Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28(4):689–707.

Gaube S, Suresh H, Raue M et al. (2021) Do As AI Say: Susceptibility in Deployment of Clinical Decision-Aids. *Npj Digital Medicine* 4(31): 1-8.

Gulshan V, Peng L, Coram M et al. (2016) Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* 316: 2402.

Jirotka M, Procter R, Hartswood M et al. (2005) Collaboration and Trust in Healthcare Innovation: The eDiaMoND Case Study. *Computer Supported Cooperative Work* 14: 369-398.

Katzenbach C (2019) Busted: AI will fix it. In: *Alexander von Humboldt Digital Society Blog*, 29 October. Available at: https://www.hiig.de/en/busted-ai-will-fix-it/ (accessed 18.08. 2021).

Kiely DG, Elliot C, Sabroe I and Condliffe R (2013) Pulmonary hypertension: diagnosis and management. *British Medical Journal* 346(1): f2028.

Kiely DG, Doyle O, Drage E et al. (2019) Utilising artificial intelligence to determine patients at risk of a rare disease: idiopathic pulmonary arterial hypertension. *Pulmonary Circulation* 9(4): 1-9.

Kuutti K and Bannon LJ (2014) The Turn to Practice in HCI: Towards a Research Agenda. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA, pp. 3543–3552.

Küppers G and Lenhard J (2005) Validation of Simulation: Patterns in the Social and Natural Sciences. *Journal of Artificial Societies and Social Simulation* 8(4): 1-13

Laurent B and Thoreau F (2019) Situated Expert Judgement: QSAR Models and Transparency in the European Regulation of Chemicals. *Science & Technology Studies* 32(4): 158-174.

Littmann M, Selig K, Cohen-Lavi L et al. (2020) Validity of machine learning in biology and medicine increased through collaborations across fields of expertise. *Nature Machine Intelligence* (2): 18-24.

MacKenzie D (1990) *Inventing Accuracy. A Historical Sociology of Nuclear Missile Guidance*. Cambridge, Mas: MIT Press.

Mackenzie D (2001) *Mechanizing Proof: Computing, Risk, and Trust (Inside Technology)*. Cambridge, Mas: MIT Press.

Ming D (2018) This Algorithm Reads X-rays Better Than Doctors Do. In: *Vice News,* 13 December. Available at: https://www.vice.com/en_us/article/kzvkym/this- algorithm-reads-x-rays-better-than-doctors (accessed 16.12.2019).

Morrison M (2015) *Reconstructing Reality: Models, Mathematics and Simulations*. Oxford: Oxford University Press.

Nagendran M, Chen Y, Lovejoy CA, et al. (2020) Artificial Intelligence Versus Clinicians: Systematic Review of Design, Reporting Standards, and Claims of Deep Learning Studies. *BMJ* 368: m689

Oakden-Rayner L (2017) Exploring the ChestXray14 Dataset: Problems. In *Luke Oakden-Rayner Blog*, 18 December. Available at: https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/ (accessed 16.12. 2020).

Oakden-Rayner L (2018) CheXNeXt: An In-Depth Review. In *Luke Oakden-Rayner Blog,* 24 January. Available at: https://lukeoakdenrayner.wordpress.com/2018/01/24/chexnet-an-in-depth-review/ (accessed 16.12. 2020).

Oreskes N, Shrader-Freshette K and Belitz K (1994) Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science* 263: 641-646.

Oreskes N (2004) Science and Public Policy: What's Proof Got to Do With It? *Environmental Science & Policy* 7: 369-383.

Oudshoorn N (2008) Diagnosis at a Distance: The Invisible Work of Patients and Healthcare Professionals in Cardiac Telemonitoring Technology. *Sociology of Health and Illness* 30(2): 272-288.

Pallesen T and Jacobsen PH (2018) Articulation Work From the Middle – A Study of How Technicians Mediate Users and Technology. *New Technology, Work and Employment* 33(2): 171-186.

Papangelis K, Potena D, Smari WW et al. (2019) Advanced Technologies and Systems for Collaboration and Supported Cooperative Work. *Future Generation Computer Systems* 95:764-774.

Perry TS (2017) Stanford Algorithm Can Detect Pneumonia Better than Radiologists. In: *IEEE Spectrum Biomedical Blog,* 17 November. Available at: https://spectrum.ieee.org/the-human-os/biomedical/diagnostics/stanford-algorithm-can-diagnose-pneumonia-better-than-radiologists (accessed 16.12.2020).

Rajpurkar P, Irvin J, Zhu K et al. (2017) CheXNet. Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. Available at: http://arxiv.org/abs/1711.05225 (accessed 16.12. 2020).

Rajpurkar P, Irvin J, Ball RL et al. (2018) Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of CheXNeXt to Practicing Radiologists. *PLOS Medicine* 15: e1002686.

Randall DA and Wielicki BA (1997) Measurements, Models and Hypotheses in the Atmospheric Sciences. *Bulletin of American Meteorological Society* 78(3): 399–406.

Scheek D, Rezazade Mehrizi MH and Ranschaert E (2021) Radiologists in the Loop: The Roles of Radiologists in the Development of AI Applications. *European Radiology* 31: 7960-7968.

Sendak M, Elish MC, Gao M et al. (2020) The Human Body is a Black Box: Supporting Clinical Decision-Making with Deep Learning. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain, 27-30 January 2020, pp. 99–109.

Shackley S, Young P, Parkinson S and Wynne B (1998) Uncertainty, Complexity and Concepts of Good Science in Climate Change Modelling: Are GCM's the Best Tools? *Climatic Change* 38: 159-205.

Sreedharan S, Mian M, Robertson RA and Yang N (2020) The top 11 most cited articles in medical artificial intelligence: a bibliometric analysis. *Journal of Medical Artificial Intelligence* 3(3): 1-12.

Star SL (1999) The Ethnography of Infrastructure. *American Behavioral Scientist* 43(3): 377–391.

Star LS and Strauss A (1999) Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work. *Computer Supported Cooperative Work* 8: 9–30.

Strohm L (2019) *The Augmented Radiologist: Challenges and Opportunities for Widescale Implementation of AI-based Applications in Dutch Radiology Departments*. Master's Thesis, Utrecht University, NL.

Swift AJ, Lu H, Uthoff J et al. (2020) A Machine Learning Cardiac Magnetic Resonance Approach to Extract Disease Features and Automate Pulmonary Arterial Hypertension Diagnosis. *European Heart Journal-Cardiovascular Imaging*. 0:1-10.

Sundberg M (2006) Credulous Modellers and Suspicious Experimentalists? Comparison of Model Output and Data in Meteorological Simulation Modelling. *Science Studies* 19(1): 52-68.

Thoreau F (2016) 'A mechanistic interpretation, if possible': How does predictive modelling causality affect the regulation of chemicals? *Big Data & Society* July-December: 1-11.

Tonekaboni S, Joshi S, McCradden MD and Goldenberg A (2019) What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. *Proceedings of Machine Learning Research 1-21*. ArXiv 2019; published online May 13. https://arxiv.org/abs/1905.05134 (preprint)

Topol EJ (2019) High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nature Medicine* 25(1):44-56.

Van Baalen S, Carusi A, Sabroe I and Kiely DG (2017) A social-technological epistemology of clinical decision-making as mediated by imaging. *Journal of Evaluation in Clinical Practice* 23(5): 949–958.

Van Baalen S and Carusi A (2019) Implicit trust in clinical decision-making by multidisciplinary teams. *Synthese* 196: 4469–4492.

White DD, Wutich A, Larson KL, Gober P, Lant T and Senneville C (2010) Credibility, salience, and legitimacy of boundary objects: water managers' assessment of a simulation model in an immersive decision theatre. *Science and Public Policy* 37(3): 219-232.

Winsberg E (2010) *Science in the Age of Computer Simulations*. Chicago: Chicago University Press.

Winter P and Carusi A (*Forthcoming*) (De)Troubling Transparency: Artificial Intelligence for Clinical Applications. *Medical Humanities.*

Airoldi Massimo (2022) *Machine Habitus: Toward a Sociology of Algorithms*. Cambridge, UK: Polity Press. 200 pages, ISBN: 9781509543281

Malte Rödl
malte.rodl@slu.se

Sociology originates in theorising social relationships and social interaction between humans. While especially STS research has moved to include ordinary machines as co-constitutive for sociality and everyday life, by now machines have started to 'learn' and internet platforms have given rise to such learning machines interacting with humans on an everyday basis. In his book *Machine Habitus: Toward a Sociology of Algorithms*, Massimo Airoldi (currently Assistant Professor of Sociology at the University of Milan) investigates the consequences of this development for (cultural) sociology. Concretely, the author provides a theorisation of inductive machine learning (ML) algorithms using the Bordieusian concept of 'habitus,' thereby proposing to understand ML algorithms as social agents and formulating a 'techno-social' account of "the good old circle of socio-cultural reproduction" (p. 143). The underlying rationale is that, similar to humans, ML is based on learning through experience.

At first, in chapter two, 'Culture in the Code,' Airoldi tackles the question of how machines learn. Drawing on analogies to the sociological concept of socialisation, the author develops a notion of 'machine socialisation,' for which in case of supervised ML algorithms there are three distinct steps: Firstly, similar to genes for a human, ML systems are programmed and set up for a specific purpose (which Airoldi calls *'deus in machina*'). Secondly, analogous to primary socialisation, a supervised ML algorithm is trained with existing data using a global data context, i.e., they "acquire a sort of 'practical reason'" (p.59) about, for example, general relevance and irrelevance. Finally, similar to secondary socialisation, ML algorithms are applied to (and learn within) a local data context, whereby they adapt to interacting with specific individuals and their preferences, i.e., they become personalised. Airoldi suggests that due to the lack of a global data context, unsupervised ML only undergoes step one and three of this analogy.

Next, in the third chapter, 'Code in the Culture,' Airoldi reverses the starting point of the previous chapter, wondering instead "how do socialized machines participate in society — and, by doing so, reproduce it" (p. 23)? Airoldi suggests to tackle this question along dimensions of, firstly, cultural alignment of both algorithmic outputs and an individual's or society's understanding, and, secondly, information asymmetry, such as how much the algorithm knows about the user's preferences and how much a user knows about the origin and manifestation of a ML algorithm's outputs. Based on this, Airoldi suggests four ideal types of interactional configuration between humans and algorithms: On the one hand, when there is high information asymmetry—i.e., when the algorithm knows a lot about the user but the user may not be aware of the aims of the algorithm—algorithms can reinforce (assist when there is cultural alignment) or transform (nudge in case this is not given) users. On the other hand, when the user is highly aware of the algorithm or the algorithm does not know much about

the user, algorithms and users can co-produce (collaborate) or alternatively, there may be misunderstandings (disillusionment of the user in case there is no cultural alignment).

In the fourth chapter, 'A Theory of Machine Habitus,' Airoldi outlines his main contribution, and only here provides a definition: "Machine habitus can be defined as the *set of cultural dispositions and propensities encoded in a machine learning system through data-driven socialization processes*" (p. 113, italics in original). Providing this definition, Airoldi adapts Bordieu's writing to the more-than-human, suggesting that even though machines "have no consciousness or meaningful understanding of reality, they contribute practically to the reproduction of society, with its arbitrary discourses, invisible boundaries, and structures" (p. 112). Concretely, the author suggests a framework on how symbolic boundaries ("how people and content are ranked and associated in both algorithmic outputs and people's minds," p. 137) and as a result social boundaries (both as a direct consequence or because of implicated changes in economic, cultural, or symbolic capitals) are shaped by both habitus and machine habitus on user-level and platform-level. The result is an analytical toolkit of four prototypical scenarios of 'techno-social reproduction' which can be distinguished along two axes based on the previous chapters: global data contexts (platform-level, through algorithmic setup and training data) v local data contexts (user-level, through personalised suggestions) on one axis, and reinforcement (cultural alignment, i.e. algorithmic outputs and user/societal predispositions are aligned) v transformation (lack of cultural alignment) of existing understandings on another axis. In order of least to most implications on the "power configuration of the field" (p. 139), these four combinations are: boundary differentiation (alignment with individual preferences), boundary fragmentation (nudging of individual users beyond their preferences), boundary normalisation (alignment on platform-level, reinforcing societal predispositions), and boundary reconfiguration (nudging on a platform level, e.g. when goals or assumptions underlying algorithmic infrastructure are updated). Airoldi notices, however, that there will likely be additional social dynamics at play, and "that the temporal oscillations and multiplicity characterizing user-machine dispositional trajectories make these scenarios no more than static approximations of ever-flowing bundles of practice" (p. 142).

In the final chapter, 'Techno-Social Reproduction,' the author summarises the main points of the book and outlines a research agenda that builds on understanding algorithms as social agents. Here, the author reminds the reader that

> the locus of the power piloting our digital lives is ultimately not the algorithmic code, but rather the hierarchical culture sedimented within it and elsewhere: a socially fabricated matter made, on the one hand, of platform owners' and machine creators' arbitrary goals and interested assumptions and, on the other, of machine trainers' habitual practices, tacit rules, prejudices and implicit assumptions. (p. 146)

Arguing on the basis of the ever-growing relevance of ML, Airoldi calls for an inclusion of the study of algorithms and mechanisms of techno-social reproduction in a novel sociological research agenda, and ultimately suggests a "(cultural) sociology of algorithms" (p. 150). He suggests four research directions, for all of which existing literature is provided, namely: following the machine creators, following the users, following the medium, and following the algorithm. In an ideal world, the author concludes, ML should be designed to be 'ignorant,' that is to include horizontal and exchange-oriented relationships — just like humans —, instead of being an opaque, all-knowing system ridden with information asymmetry.

Airoldi is aware that the book's propositions might not be surprising or novel to the reader, especially those "familiar with STS or ANT literature" (p. 118). He emphasises that "The purpose of this book is to restate the obvious in a sociologically less obvious fashion, deliberately designed to 'transgress' disciplinary borders, as suggested by Bordieu himself" (p. 31). Thus, it is unavoidable that some readers may find some parts of the book redundant. Nevertheless, Airoldi's sociologically grounded theorisation of ML algorithms as social agents may be intriguing for STS and ANT scholars, and an "epistemological rupture" (p. 149)

for sociologists with a suggestion of socialised machines as "a source and factor of social order" (p. 147).

In opposition to much recent critical scholarships on algorithms, Airoldi suggests that research should go beyond the study of algorithmic biases and instead focus on concepts from cultural sociology "like *culture, socialization, practice*, and *habitus* [which] open a whole new set of questions" (p. 48, italics in original) when applied to the study of ML systems. Accordingly, the examples in the later stages of the book are increasingly focused on everyday life, including taste and cultural consumption, which requires some efforts from readers to transfer insights to other research areas, but which (so I thought) is highly rewarding. Given the focus of the book, Airoldi does not include much discussion on the configuration of everyday life through affordances (a term not even mentioned in the otherwise useful index) of ML system's inputs or platform design.

The consistent and thorough focus on the influence of algorithmic systems on everyday life, consumer society, and culture is an important contribution to research on algorithms. Airoldi's "mechanisms of techno-social reproduction" (p. 149) open up possibilities to account for "second-order consequences [of ML algorithms] on society and culture" (p. 85), thereby affirming the constitutive impacts of ML on societal meaning-making and enabling research to, for example, interrogate ML's involvement in and contribution to the multiple crises of late capitalist consumer culture. Overall, then, despite some redundancy for STS and ANT scholars, the book fruitfully links various literatures including Bordieusian cultural sociology, STS, and critical algorithm studies. It provides an eclectic introduction into the social scientific study of algorithms paired with intriguing concept development, providing the reader with the necessary analytical tools to understand and theorise ML algorithms as social agents participating in techno-social reproduction.

Timcke Scott (2021) Algorithms and the End of Politics: How Technology Shapes 21st Century American Life. Bristol: Bristol University Press. 198 pages. ISBN 978-1529215328

Katrina Nicole Matheson
York University

In *Algorithms and the End of Politics* (2021), Scott Timcke offers a Marxian analysis of digital technologies and politics in U.S. society. Although the title of the book suggests an interrogation at the intersection of critical data studies and politics, Timcke's wheelhouse - as a comparative historical sociologist studying race, class and technology – is squarely in political analysis. The result is a blistering deconstruction of American democracy to demonstrate his overarching point that advancements in artificial intelligence and other data-driven technologies do not reconceptualise politics but are instead "a new kind of communication that preserves an old kind of polity" (p. 148). To be sure, the old kind of polity is one rooted firmly in capitalism.

Timcke begins with a scathing assessment of an American political system overrun by capitalism. Using examples across the political spectrum, Timcke argues that American democracy "has only been acceptable as a management style for capitalism" (p. 3). It is against this socio-political backdrop that Timcke extends the framing of "unfreedom and class rule" to the "digital realm" (p. 3). Moreover, he argues that the capitalist ruling class has captured computational resources and is using them to drive their self-serving global agenda.

Timcke argues that this asymmetrical application of increasingly complex digital technology has led paradoxically to a simplification of the social world, with datafication a prime example. Using a rendering of the term similar to Van Dijck's (2018) notion of 'dataism,' Timcke defines datafication as an ideology that advocates for the "implementation of computational reason to oversee human life" (p. 4). Drawing on the work of Fuchs (2021) and Srnicek (2017), Timcke concludes that datafication has weakened U.S. democracy.

Although Timcke's class-conscious approach to datafication is an important contribution to existing debates in STS, he frequently loses threads salient to critical data studies in his dense and discursive political analysis. For example, in tracing the conditions of growing inequality and voter disaffection that gave rise to the Trump presidency, Timcke rebukes the Democratic Party's commitment to a neoliberal economic system that elevates Facebook/Meta, despite CEO Marc Zuckerberg's dubious mantra 'move fast and break things.' He calls the Democratic Party's emphasis on performative respectability to mask its commitment to the socially ordered status quo a 'sterile' ideology unable to foster human flourishing (p. 10). Though Timcke's conclusion has merit, he misses here an opportunity to link the notion of sterile governance back to applied datafication/dataism. Given the book title's invocation of 'algorithms,' a more impactful example might have been the Democratic embrace - at a minimum through persistent regulatory inaction - of predictive and surveillant algorithmic tools and the business opportunities that are built around them. Predictive policing and judicial sentencing,

algorithmic screening for social services, and predatory applications of data-driven marketing are all potentially more illustrative examples of an applied ideological sterility that systematically obstructs human flourishing, than the abstraction of Zuckerberg's mantra. Nevertheless, Timcke's critique of the larger political system is well-taken: the Democratic Party can't address the forces that gave rise to the Trump presidency because it is a wholesale subscriber to those same forces which exist to serve the capitalist elites. His assessment that the American public has, at present, no developed mechanism of resistance to counteract rapidly intensifying datafication regimes is central to Timcke's arguments on how to move forward.

Chapter One builds on the illusion of an American two-party system by examining the role of algorithms in either reifying or threatening public conceptualizations of political legitimacy. Invoking the work of Beer (2017) as well as Ruppert et al. (2017), Timcke echoes the need for a thick description of algorithmic encodings to understand how authority is expressed algorithmically, but adds that an analysis of the mode of production (i.e. who is creating the value vs. who is accumulating the capital) is needed in scholarly considerations of algorithmic regulation. The goal of this, in Timcke's view, is to encourage scholarship that offers a pathway for the data subject to consider participation in data politics as an avenue for revolutionary social change. In other words, rather than prioritizing research that ensures algorithms can recognise and potentially exploit Black female faces as accurately as white male faces, researchers should instead strive to achieve technologies of liberation for the data subject.

Following a Chapter Two that describes Timcke's notion of datafication as mentioned above, Chapter Three explores communication technology in Gramscian theory, especially its role in winning the active consent of subordinate classes. Billionaires not only rationalise their self-interest in the media but also "demand veneration as exemplars of moral virtue" (p. 64). Leaning into a portrayal of benevolence, billionaires have invested heavily in the news sector and are often lauded for what is perceived as a nearly philanthropic pursuit. Timcke makes the point that such investments are not philanthropic but instead allow "digital men of power" access to levers that effectively control "the means of mental production" (p. 71) – that is, targeting criticisms of their accumulating wealth, no matter their political origins.

Chapter Four builds on the notion of billionaires in media to unpack the neoliberal response to the challenge of credible, socialist-leaning U.S. presidential candidate, Bernard Sanders. In Timcke's view, the rise of Sanders reflected a populous fatigued by financial and other crises, who saw a Sanders presidency as a plausible path to winning power. As a result of the threat his movement posed to entrenched capitalist interests, Sanders was met with cultural mechanisms enforced by a "willing and compliant media" to smear him as sexist (p. 78). Ultimately, Timcke concludes that the nomination of Biden over Sanders in the 2020 Democratic primary demonstrates that "the party decided" against inclusive political economic reform espoused by Sanders and employed communication technology under control of threatened billionaires to facilitate its preferences (p. 95).

In Chapter Five, Timcke draws heavily on the work of Reed (2002), Roediger (1999) and others to conclude that markets depend on racism and sexism to reproduce themselves. He suggests that notions of race arise from modernity to embody a relationship to authority and, by extension, to capital. Timcke concludes that capitalist polity is deeply committed to perpetuating both sexism and racism because each acts as a compelling externalization used to justify political failures and contradictions (e.g. the explanation that Trump was elected in 2016 because of sexism against Hilary Clinton, as opposed to the failure of her policy platform).

In Chapter Six, Timcke expands the role of Marxian contradictions as applied to misinformation. He argues that although modern technology may spread misinformation more readily, misinformation itself is a longstanding tool relied upon by capitalists to mystify and deflect inevitable contradictions (e.g. between labour/capital, commodity/value, etc.). Timcke says:

"Put simply, American political parties must distract citizens from the primary causes of oppression and alienation…Misinformation is not

an engineering problem or a social problem, but the active avoidance of a social question" (p. 126).

Chapter Seven closes out Timcke's argument with an analysis of algorithmic processes (e.g. artificial intelligence) in U.S. state security initiatives. Here, he summarises his approach to the entire book: to explore how surveillance cultures combine elements of hegemony (consent) and domination (coercion) to shape digital society. This encapsulates Timcke's call for a shift in scholarly mindset towards a macro view, while simultaneously eschewing typical lines of argument about ethics and equity found in critical data studies pieces.

Timcke's tone and Marxist analysis resemble that of Srnicek's (2017) work on platform capitalism. Whereas Srnicek has focused on deconstructing the socio-cultural personas of Big Data enterprise to reveal its vile profit-seeking core, Timcke similarly pulls back the curtain on the digital ecosystem of modern political enterprise (which is now a two-way street between corporations and politicians operating in broad daylight). Timcke's conclusion that class dynamics within our social hierarchies have remained static over the long arc of capitalism echoes Couldry and Mejias's (2019) assessment of 'data colonialism' as the manifestation of an unchanging social structure of domination and exploitation which emerged during historical colonialism and continues through to this day. Moreover, in the way that Benjamin (2019) has drawn academic attention to the persistent harm of entrenched racism in the co-production of digital spaces while inspiring us to readjust the default paradigm, so too is Timcke attempting to do here for class subordination. Too bad, he doesn't quite meet the bar. While his political analysis is revealing, this book lacks the empiricism and ethnographic detail that we have come now to expect from prominent scholars of algorithmic overreach, for example Zuboff (2019) or Eubanks (2017). Accordingly, *Algorithms at the End of Politics* reads much more like a political manifesto than the average STS scholar might prefer. Nevertheless, Timcke's message for researchers in the field is important: class dynamics cannot be omitted from sociological interrogations of algorithmic technology and political economy. This is a particularly timely message given the resurgence – after two to three generations of dormancy – of American labour unions, which in some ways is being led by employees of Big Data companies (Bose, 2021). In the time since the book's publication, a fledgling workers' union has sprung from the grassroots at an Amazon facility in New York, with additional organizing efforts ongoing. The distribution warehouse, known for its inhumane conditions and high worker turnover, has become the material site of resistance against a digital capitalist giant – who spent more than US$4 million to convey misinformation about the perils of unionization to captive worker-audiences during their organizing campaign. For as advanced as Amazon's technology is, the current strife between the company and its workers feels distinctly twentieth century. Here, Timcke's primary argument plays out: rather than change conceptualizations of work and social class relationships, technological advancements appear only to provide a more powerful vehicle for entrenched capitalism to do what it has always done – exploit labour.

As STS scholars, it's tempting to frame technoscientific research in ways that reify the existing legal, financial and social hierarchies, even as we openly confront ethical matters of race and gender equity. Unfortunately, supremacy of the economic ruling class is just as invisible, pervasive and consequential as white and cis-gendered male supremacy in academic spaces. The dawning of a class-conscious labour movement in the United States reaffirms Timcke's concluding optimism and it is from here that the subfield of critical data studies might also take a cue: "There is no sociological law that stipulates that algorithmic life must be inherently discriminatory [to members of subordinate social classes] … I think there is much heart to be taken from resurgent broad-based socialist politics in the U.S. When [digital] democratization does come, it will emerge from this venue" (p. 155).

References

Beer D (2017) The social power of algorithms. *Information, Communication & Society* 20(1): 1–13. https://doi.org/10.1080/1369118x.2016.1216147

Benjamin R (2019) *Race After Technology: Abolitionist Tools for the New Jim Code*. Newark: Polity Press.

Bose N (2021) *U.S. Labor Movement's next Frontier Is the Tech Industry, AFL-CIO's Shuler Says*. Reuters, December 6th. Available at: https://www.reuters.com/technology/us-labor-movements-next-frontier-is-tech-industry-afl-cios-shuler-says-2021-12-04/ (accessed 6 May 2022).

Couldry N and Mejias UA (2019) Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject. *Television & New Media* 20(4): 336–349. https://doi.org/10.1177/1527476418796632

Eubanks V (2017) *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press.

Fuchs C (2021) The Digital Commons and the Digital Public Sphere: How to Advance Digital Democracy Today. *Publishing, the Internet and the Commons* 16(1). https://doi.org/10.16997/wpcc.917

Reed A (2002) Unraveling the Relation of Race and Class in American Politics. *Political Power and Social Theory* 15: 265–274. https://doi.org/10.1016/s0198-8719(02)80026-6

Roediger D R (1999) *Wages of Whiteness: Race and the Making of the American Working Class*. London: Verso.

Ruppert E, Isin E and Bigo D (2017) Data politics. *Big Data & Society* 4(2): 205395171771774. https://doi.org/10.1177/2053951717717749

Srnicek N (2017) *Platform Capitalism*. Cambridge, UK: Polity.

Van Dijck J (2014) Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance & Society* 12(2): 197–208. https://doi.org/10.24908/ss.v12i2.4776

Zuboff S (2019) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. First edition. New York: Public Affairs.

Science & Technology Studies

Volume 35, Issue 4, 2022