Original Research Article

BIG DATA & SOCIETY

Big Data & Society
January–June: 1–12

DOI: 10.1177/20539517231155061
journals.sagepub.com/home/bds
SAGE

Contextualizing realism: An analysis of acts of seeing and recording in Digital Twin datafication

Paulan Korenhof[1], Else Giesbers[2] and Janita Sanderse[3]

Abstract

Digital Twins are conceptualized as real-time digital representations of real-life physical entities or systems. They are explored for a wide array of societal implementations, and in particular to help address fundamental societal challenges. As accurate digital equivalents of their real-life twin, Digital Twins substitute their physical twin in knowledge production and decision-making processes. They raise high expectations: they are expected to produce new knowledge, expose issues early, predict future behavior, and help to optimize the physical twin. Data play a key role here because they form the building blocks from which the Digital Twin representation is created. However, data are not neutral phenomena but products of human-technology interaction. In this article, we therefore raise the question of how a Digital Twin data collection is created, and what implications does this have for Digital Twins? To answer this question, we explore the data collection process in three cases of Digital Twin development at a university. Connecting to Jasanoff's theoretical framework of *regimes of sight*, we approach the creation of a data collection as acts of seeing and recording that influence how reality is represented in data, as well as give a certain legitimacy and authority to the data collection. By examining the acts of seeing and recording and their respective roles in producing the data collection, we provide insight into the struggles of representation in Digital Twins and their implications.

Keywords

Digital twins, datafication, data production, non-neutrality, regimes of sight

Introduction

The ongoing development of digital technologies enables the collection, storage, and processing of increased amounts of data. This in turn allows for the advancement of new data-driven technologies like Machine Learning and Big Data. The new kid on the block of data-driven technologies that is finding its way from niche use cases to overall societal application is the "Digital Twin." While there is no uniform definition of what a "Digital Twin" is, a Digital Twin is commonly conceptualized as a real-time and highly accurate digital representation of a particular real-life physical entity or system as it changes over time and under diverse conditions (Liu et al., 2021; Van der Valk et al., 2020). Digital Twins make use of "evolving data" to mirror their physical twin one-to-one as it changes in the real world, which differentiates them from "plain models" that only provide snapshot representations (Wright and Davidson, 2020). Examples of entities that are being "twinned" are airplanes, farms, oceans, and cities. As accurate digital equivalents of their real-life twin, Digital Twins substitute their physical twin in knowledge production and decision-making processes (Korenhof et al., 2021). Experimentation with the Digital Twin is expected to reveal previously unknown dependency relations of the real-life physical entity, expose issues early, predict future behavior, and offer directions for optimization (e.g., European Commission, 2022; Grieves and Vickers, 2017; Schleich et al., 2017).

Currently, Digital Twins are explored for a wide array of societal implementations, and in particular to help address

[1]Philosophy Group, Wageningen University, Wageningen, Netherlands
[2]Innovation - and Risk Management and Information Governance, Wageningen Economic Research, Wageningen, Netherlands
[3]Data Science, Information Management & Projectmanagement Organisation, Wageningen Economic Research, Wageningen, Netherlands

Corresponding author:
Paulan Korenhof, Philosophy Group, Wageningen University, Wageningen, Netherlands.
Email: research@korenhof.eu

fundamental societal challenges. For instance, Digital Twins are explored as a tool to help improve health by twinning the human body, reduce nitrogen emission by twinning farms, and improve sustainability by twinning greenhouses. The European Commission is even explicitly looking at Digital Twins to play a pivotal role in the European Union's Green Deal policy that aims to battle climate change (European Commission, 2022).

While there is a high societal request for Digital Twins, not many are implemented yet. This is because building a Digital Twin is complex, especially in the case of natural entities (Pylianidis et al., 2021). At the heart of this complexity lies the fact that a Digital Twin is inherently different from its physical twin: as Floridi points out, "[a] computational fluid dynamics simulation, used to model and simulate the behaviour of flowing air, is neither windy nor wet in itself" (2013: 321). Digital Twins represent their physical twin by means of data and models (Wright and Davidson, 2020). Data play a crucial role here because they form the building blocks from which the mirror representation is constructed by modeling algorithms. At the base of Digital Twin development therefore lies a process of "datafication": reality needs to be put "in a quantified form so that it can be tabulated and analyzed" (Mayer-Schönberger and Cukier, 2013: 78). However, research from philosophy of technology, science and technology studies, and critical data studies has shown that data are not neutral phenomena, but rather products of human choices made in interaction with technology (see e.g., Gitelman and Jackson, 2013; Kitchin, 2014; Mayer-Schönberger and Cukier, 2013). This raises the question of what shapes the making of a Digital Twin data collection, and what implications does this have for Digital Twins?

Due to the novelty of Digital Twins, there is still little known about their design process. An early valuable contribution here is by Solman et al. (2022), who studied the design process of Digital Twin windmills. They identify decisions on data as one of the key factors responsible for what is and is not represented by a Digital Twin. We provide a complementary contribution with this article by zooming in on these decisions and their interplay with technology in the creation of the data collection. Our reason for this is two-fold. First, given the societal expectations and expected future implementation of Digital Twins in key areas of society, like healthcare, food production, and climate management, we believe it is important to gain insight into the datafication process underpinning Digital Twins and identify potential issues. Second, on a more general level, with the increasingly important role of data for knowledge production, we think that Digital Twins with their high data requirements provide a forecast of the challenges that we may expect as the request for data-driven knowledge production keeps increasing.

In order to gain insight into the process of datafication and its implications, we examine three Digital Twin projects: one sees to tomato crops, one to a farm, and one to the glucose and triglyceride levels in the human body. The set-up of our article is as follows: we start by discussing the theoretical framework that underpins our research. Connecting to Jasanoff's theoretical framework of *regimes of sight* in particular, we approach the creation of a data collection as acts of seeing and recording that influence how reality is represented in data, as well as give a certain legitimacy and authority of the data collection. Next, we introduce the three cases and explain our methodology. This is followed by our case analysis in which we examine the acts of seeing and recording and their respective role in producing the data collection. Based on this case analysis, we then identify three main challenges in the Digital Twins' politics of representation arising from the datafication process. We conclude with some final remarks on these challenges.

Background: The production of data

As already pointed out in the introduction, data are not neutral representations of reality. They are not already lingering in the atmosphere merely waiting to be collected. On the contrary, data are *produced:* they are the product of human decisions and practices and rest on certain assumptions about reality and what the data can represent (Gitelman and Jackson, 2013; Kitchin, 2014). Data are produced "by abstracting the world into categories, measures and other representational forms—numbers, characters, symbols, images, sounds, electromagnetic waves, bits" (Kitchin, 2014: 1). Data show something that reality itself does not immediately show us: a growing plant does not display its growth per day in numbers. Relations between things or the existence of things of which we would be unaware without the data can be revealed. Jasanoff therefore describes the production of a data collection as "an act of seeing and recording something that was previously hidden and possibly nameless" (Jasanoff, 2017: 2). These insights are in line with a longer history in research and are not restricted to digital data. Research in philosophy of science and science and technology studies has since long asserted that theory and equipment affect assumingly objective representations of reality (see, e.g., Godfrey-Smith, 2021; Hacking, 1983; Latour and Woolgar, 1986). Different materializations of data involve different technologies, instruments, agents, information infrastructures, practices, methods of knowledge production, and required expertise, which in turn all influence how reality is represented in the data—in sum, they involve different acts of seeing and recording. To put it more concretely, producing a data collection requires the making of decisions on the level of recording: what kind of data to collect (e.g., data about the weather, citizens movements, growth speed of a plant), what software and hardware to employ for this (e.g., sensors, scraping tools, algorithms), what measuring systems to use (e.g., weight, length), and what form to

put it in (e.g., numbers, RGB values, text) (Kitchin, 2014; Winsberg, 2019). Such decisions entail the selection, inclusion, and exclusion of data (Brine and Poovey, 2013; Gitelman and Jackson, 2013). All these decisions take place in an interplay with acts of seeing: a certain view on what is datafied, why it matters, the purpose of the data collection, and its context (Dourish and Gómez Cruz, 2018; van Dijck, 2014).

Moreover, the acts of seeing and recording are involved in what Jasanoff calls *regimes of sight* (2017). Jasanoff explores the politics of representing the world in data. She argues: "if a data set is to elicit a social response, knowledge of something that matters and principles for understanding why it matters must be generated together, or coproduced" (2017: 2). She describes three modes of seeing the world that can underpin a data collection: a *view from nowhere*, a *view from everywhere*, and a *view from somewhere*. The *view from nowhere* reflects the view "of science as traditionally imagined: impartial, objective, seeing or making sense of facts as if from a neutral viewpoint that renders what is seen devoid of distortion and human bias" (Jasanoff, 2017: 3). The political claim of the *view from nowhere* is that this mode of seeing stands outside of politics. In contrast, the *view from everywhere* is that it is politically inclusive of all affected positions. It is the view typical for "expert advisory bodies in modern democracies, claiming to represent all relevant aspects of the problem put before them, and capable of aggregating them into a credible, reliable, whole" (Jasanoff, 2017: 3). Last, Jasanoff introduces the *view from somewhere*. This is "the subjective gaze of the eyewitness, and even more the experienced knower, who has not only seen but possibly endured in person the consequences of the facts attested to" (Jasanoff, 2017: 3). The political claim of the *view from somewhere* is one of personal injury. The view underpinning a data collection provides it with a certain legitimacy and authority on which people can act (Jasanoff, 2017).

The creation of a data collection is thus shaped by acts of seeing and recording. These acts not only influence how reality is represented in data, but also are involved in a regime of sight that gives a certain legitimacy and authority to the data collection.

Cases and methods

In order to gain insight into the acts of seeing and recording that shape a Digital Twin data collection, we explored three cases of Digital Twin development in the life sciences. The studied projects are: i) Virtual Tomato Crop, ii) Me, My Diet and I, and iii) Digital Future Farm. All three projects were executed at Wageningen University and Research (WUR) and had a runtime of approximately three years. While all three projects originate from the life sciences, they had a focus on different physical entities and had different research teams working on them. The goal of these projects was to get a clear idea about what is needed to develop and run a Digital Twin, and all three projects aimed to develop a working prototype. The projects wanted to realize "greater reproducibility and transparency about what is and isn't working [in a Digital Twin design]". We performed an additional and overarching study connected to these projects to provide insight into the influencing factors in the data collection process and their implications. Before discussing our methodology, we first briefly describe the three projects.

Virtual tomato crops

The Virtual Tomato Crops (VTC) project developed a Digital Twin for a tomato crop in a greenhouse. The Digital Twin is fed with real-time data deriving from sensors in a real greenhouse. Every week, measurements were taken of inter alia the plant height, leaf area and leaf weight. Thermal cameras created images of the plant. Other sensors in the greenhouse measured environmental factors like temperature, humidity, and light intensity. The goal of this Digital Twin is to offer refined predictions about plant development and serve as a decision-support-tool for tomato growers as well as a resource for researchers. It can be used to make both short-term decisions (e.g., temperature in the greenhouse) as well as long-term decisions (e.g., the type of greenhouse or type of tomato that should be used) to optimize the cultivation of a tomato crop.

Me, My Diet and I

The Me, My Diet, and I (MMDI) project focused on building a personalized Digital Twin that can predict the rise of glucose and triglyceride after eating a meal. Both glucose and triglyceride can indicate that there is a risk for cardiovascular diseases. Several studies in which personalized data is gathered were used in this research. Four of these studies were already executed in the past as part of other projects in which the developers participated. Additionally, a new study was performed in which phenotypes are made of 220 people. Participants are required to follow a specific diet and undergo multiple tests, like an MRI scan and biopsies, so a curve can be seen. While the developers would have liked to use real-time data, this was not possible. The final goal of this Digital Twin was to improve people's health by providing dietary decision support.

Digital Future Farms

The Digital Future Farm (DFF) project aimed to display the existing nitrogen cycle of arable or dairy farms. The goal of this Digital Twin was not to strive for an illustration of a "perfect nitrogen cycle" (one of the interviewees said: "the story that you can create a perfect cycle is nonsense"),

but to answer the question "whether you can come as far with less [nitrogen]". This project did not actively gather data on farms but made use of existing data and external data sets. Examples of used data sets are data about the weather, the soil type, pesticide use, and irrigation. This Digital Twin was developed with two kinds of end-users in mind: first, the farmers. Farmers can use the Digital Twin as a decision support tool. By working with scenarios, the farmer can make trade-offs and determine what is best for their farm. Second, the scenarios can also be used by researchers to measure strategic alternatives and their consequences. This can help to support innovations in agriculture.

Exploration of the three cases

Our departure point for the case study is Jasanoff's analysis that the creation of a data collection involves acts of seeing and recording (2017). These acts are expressed in the decisions of the developers. We therefore focused our research on these decisions and their products. In order to explore these decisions and their context, we took a blended ethnographic approach, consisting of interviews, document research, and participatory observation. Our reason for the blended approach is that we wanted to obtain insight into the conscious as well as unconscious decisions regarding the data collection and the factors that influenced these decisions. The blending of diverse methods let us cast a wide net and thereby increased the likelihood of getting a larger picture.

The document research consisted of close reading of internal documents regarding the aims and progress of the projects, publications by the project teams, as well as documents used for external communication to stakeholders or peers. The document research provided insight into the view on the physical twin that underpinned the data collection, the collected data, the state of affairs of the Digital Twin, the stakeholders, the used technology, and the goals of the projects. The document research informed our choices for who to interview and what questions to ask.

The goal of the interviews was to get insight into the decisions regarding the data collection and the factors that influenced these decisions. The core teams of the projects consisted of roughly five to ten researchers (what was considered to be the core team differed a bit depending on whom we asked and what document we consulted). Next to the core team, other researchers on occasion participated in the projects. These worked on specific smaller parts of the project. The size of the occasional researchers differed per project and ranged from seven to about thirty-five people. The project leaders had the main oversight and control over the project. They indicated that they were either the ones making all the decisions with regard to what data to collect, or that it was a collective process in which they were involved in all the decisions made. Based on the previous, we decided to hold semi-structured interviews with seven people: three project leaders, three other core team members working in diverse roles in the projects, and with a Digital Twin methodology researcher who works in an overarching platform that supports the projects. The interviews were semi-structured and an interview guide was developed providing the main structure for the conversation. This approach allows us to deviate from the interview guide, ask follow-up questions, and/or discuss affiliated topics that come up during the interview (Kallio et al., 2016). Interviewees were questioned about: (1) the goal(s) of the Digital Twin, (2) decisions that were made regarding data collection, (3) choice-makers that steered toward these decisions, (4) the way data was gathered, (5) additional influences in the data gathering process, and (6) do's and don'ts in data gathering. The interviews were held online between June and October 2021. All but two interviews were held in Dutch. The quotes that are used in this article are translated by the authors. The other two interviews were held in English.

The participatory observation served to uncover influences on the data collection that did not surface in the interviews or document research, as well as to check and elaborate our findings. It consisted of passive to moderate participation. The first author is part of a platform that aims to provide support to the project teams on the level of ethical and social aspects of the Digital Twins, but was not directly involved in the Digital Twin data collection or development. In this capacity, she attended meetings, presentations and workshops related to the projects over the course of two years and took field notes. Such events occurred about every other month. In January 2022, the authors gave a presentation to the projects teams about our preliminary results, which was followed by a discussion about our findings. The discussions during the diverse meetings, presentations, and workshops allowed us to confirm our analyses and clarify remaining questions.

Last, we want to acknowledge that gaining more detailed insight into the data collection would have benefited from being physically between the developers and other researchers on the work floor. Unfortunately, the COVID-19 restrictions that were in place during most of the execution of this research did not allow us to do so.

Seeing and recording in the datafication process

In this section, we discuss our analysis of the acts of seeing and recording that shaped the data collection in the three Digital Twin projects. We start with discussing the main modes of seeing underpinning the projects. The primal institutional locus of all three projects was science. However, as we move to the acts of recording, we show that these acts, and in particular the available technologies,

had a far-reaching influence on the materialization of the data collection as well as on the view the developers had on what the datafied physical twin should and could be. Last, we show how difficulties with regard to available resources and technologies led to practices in which views with different institutional loci, purposes, and motives seeped into the data collection.

Seeing the physical twin

The three projects we studied responded to a call for proposals set out by WUR to develop a Digital Twin. In the call for proposals, Digital Twins are described as a "synergetic, holistic digital copy of reality" and in further documentation they are conceptualized as "computer models of individual objects or processes that are updated on the basis of real-time information"[1]. Correspondingly, respondents indicate that they were asked to "come up with something that reflects reality" and offers a "bigger picture" with regard to a particular entity or process. The project teams understood a Digital Twin as a real-time and multi-faceted representation of a specific real-life physical twin that can help users to improve decision-making with regard to the physical twin. Moreover, they wanted to develop a reliable Digital Twin that can be distinguished from "all the commercial apps and promises that are made."

In order to datafy a physical entity for Digital Twin purposes, the developers needed to have a certain understanding of what needs to be put into data—the physical entity—as well as a view on how to do this in a justified manner. In all three projects, this driving view underpinning the datafication of the physical entity has its institutional locus in science. The main work environment of the three projects is a university. Here, the Digital Twin projects were embedded in a public research-oriented work culture that steered the projects toward scientific goals. Jasanoff calls this type of view underpinning the data collection a *view from nowhere*: a view "of science as traditionally imagined: impartial, objective, seeing or making sense of facts as if from a neutral viewpoint" (Jasanoff, 2017: 3).

For the datafication of the physical entity, the projects build on previous scientific research done by members of the project team. In the teams, academic researchers with different scientific backgrounds were involved. The scientific backgrounds can be roughly categorized into three general disciplines: life sciences (e.g., animal science, molecular biology, plant science), computer sciences (e.g., artificial intelligence, machine learning, data science), and social sciences (e.g., economics, political science). While the projects employed people with various backgrounds, we found that the main guidance for the datafication of the physical entity, and in turn the way in which the physical entity was brought into focus, was predominated by one of a few scientific approaches. It was for the majority the background of the project leader in combination with the university's main areas of research interest, that shaped the approach of the physical entity. For example, the DFF chose to focus on "nitrogen emission because a) there is a lot attention for and b) there are many models at the university." Other respondents explained that they chose to focus on a physical twin about which they had a significant amount of previous knowledge, because developing a Digital Twin was already a big challenge in itself. Previous knowledge shaped the developers view on what data to collect data because "you know that a plant likes CO_2, water, light, so you are not going to measure the mood of humans [and the influence it has on a plant]. You are just going to measure the things that you have prior knowledge of". By building on previous research, the insights already produced there are reiterated in the data set: data which research has shown to be important is gathered and thus in turn, more likely to be confirmed as relevant for the Digital Twin representation.

The dominant scientific approaches that underpinned the datafication of the physical entities originate from the life sciences. More specifically, the approaches to the physical twin entail a naturalist view on reality. In these approaches, reality is regarded as something that can be understood by empirical observation, experimentation, and the application of natural laws. The main mode of representing reality in naturalist approaches is by means of data and models, which resonates with the conceptualization of Digital Twins as realistic data-driven models. The naturalist view on reality had significant implications for the manner in which a physical entity was framed in terms of data. For example, the focus of MMDI project was on "[p]eople with a predisposition for diabetes". However, the MMDI Digital Twin does not represent all possible data about a human being on macro- and micro-level, but focused on metabolism, and in particular on glucose, triglyceride and digestion related data. Other elements of the reality of a human being, like psychological state, or the costs of a certain diet, are not represented in the data set. Scientific approaches from the social sciences seemed to play only a secondary role in the datafication process by for example advising on how the employ the Digital Twin to promote behavior-change. One of the projects did want to take economic data into account, but this was planned for a later stage. Overall, the scientific views underpinning the Digital Twin development thus reflected a naturalist view of the physical twin and called for the use of certain kinds of data.

Normative influences. Corresponding to the contemporary understanding of science as something that should not only be objective but also needs to adhere to certain ethical and legal norms, the data collection was at some points restricted. The project teams took account of ethical concerns raised by stakeholders and ethicists and adjusted their data collection accordingly. For example, in the case of DFF, the project seeks to use data generated

by farms. As one of the interviewees explained, the farms are the life of the farmers. Some farmers are concerned that their company data may be used against them to restrict their autonomy by for example leading to more regulation by the government, or may be used by organizations to profit from them at the expense of the farmer. Concerns about privacy, loss of autonomy, the freedom to conduct a business, or the risk of data misuse, were in some cases a hurdle for people to get involved in a study and share their data, as well as a reason for the developers to be careful with regard to what data they collect. In the case of the MMDI project, the team had to comply with strict medical-ethical norms because they made use of medical data. According to a respondent, these "are 10 times stricter than the ones [the general guidelines for the gathering of data] they have at the university (…). Whenever you work with people, that is just very important."

Additionally, we saw some influence of legal restrictions. These only appeared on the radar in relation to the processing of personal data, which meant that the developers had to comply with the General Data Protection Regulation (GDPR). Overall, this seemed to have less influence on what data the developers collected than the ethical norms they took into account.

We understand these restrictions to the data collection as being inherent to the *view from nowhere*.

The unintended somewhere. In assembling the data collection, the implications of the geographical situatedness of the *view from nowhere* became evident; the developers were *somewhere* and from *somewhere.* This locality of the project teams influenced their data collection on two levels. First, the project teams needed to have sufficient access to the physical twin to perform their research and to gather data. The more data the projects needed to collect themselves, the more the location of the physical twin played a role in its datafication as the research team needed to have sufficient access to the physical entity. Given the fact that the projects are located at WUR in the Netherlands, there was a preference for the physical twins —whether it concerns people, farms, or greenhouses with tomatoes—to be located in the Netherlands. This led to the collection of data with a Dutch geographical origin. Second, the geographical situatedness came with a focus on and mastery over certain languages, which affected the recruitment of test subjects as well as the re-use of data. One of the respondents stated: "[a]t a certain moment you have a model, but if you want to have your model up and running in China or Brazil, you have to check where that data is. And that is hard. For example, in Belgium they are bilingual. You are searching in Wallonia at what platform you can find the data, but I cannot read French so I cannot find it. Then I call a colleague, but that gets harder when you are further away from the Netherlands. We have not mastered that yet."

Recording the physical twin

Next to acts of seeing, the production of a data collection requires acts of recording: the production of a data collection requires the developers to translate their views on the physical entity into concrete data. Here, the view underpinning a data collection not only matters for the choices the developers make with regard to what data to collect, but it also *matters* in the sense that it is involved in mattering practices where diverse elements of the material world and the views of the developers mutually affect each other in producing a data collection (Dourish and Mazmanian, 2011; Dourish, 2017; Thomer and Wickett, 2020). This is a process of materialization: people need to choose and employ software, hardware, measuring tools, perform maintenance, choose what formats to use and set up the infrastructure (Kitchin, 2014; Winsberg, 2019). Technology plays a fundamental role here: it is constitutive for the data by being the means and medium of the data production. Technology thereby embodies the mold in which the Digital Twin is cast. Without technology, there is no digital data. By being the instruments for data collection as well as the medium in which the data are stored and processed, technology influences the datafication of the physical twin. In this section, we discuss how the material conditions of the Digital Twin development coproduced the data collection with the views of the developers.

Expansive and promoting influence. A crucial role in the data collection is played by the availability and capability of measuring technology like sensors planted in the physical entity or its surroundings, cameras on satellites and drones, and hand-used sensor equipment: these technologies produce the data about the physical twin. These technologies opened up all sorts of options for the projects to produce data. For example, the developers had access to a high-tech greenhouse which provided them with the opportunity to work with state of the art technologies like climate sensors, multi-spectral 3D laser scanners, and chlorophyll fluorescence cameras. With this, the developers were able to produce new kinds of data about the physical twin, data that they could not generate before. Moreover, the availability of measuring technologies and supporting infrastructure easily allows for gathering more data than may be needed. This was especially the case when the developers were uncertain what data and how much data were needed for the Digital Twin. In this case, the developers decided to gather a lot of data which can probably not all be used in the end, to make sure they did not gather too little data: "so, the bottom-line is that we probably measure more than we will need."

In one of the projects the interplay between the opportunities offered by technology clearly rose to the foreground in their Digital Twin design. One of the interviewees described their decision-making process with regard to

data-gathering as a "chicken-or-egg" issue: "we do not know what we should measure because we do not know for sure what the model looks like. We only know what the model will look like if we know what we can measure at all." This reveals Digital Twin design as an iterative process in which designers shift back and forth between technology (what can be measured) and scientific views (what does the model look like). This also shows that technology not only directly influenced the data set, but also influenced how developers thought about the datafication of the physical twin. The availability of certain measuring tools inspired developers to think about and look into data that may not have been on their radar before. When asked about to what extent technology influenced the data collection, one of the respondents stated: "[o]h yes everything! Because you know, when it is technically possible, we will do it". As many of the new technologies that the developers explored and used are measurement technologies, the technology in turn solidifies the role of empirical observations in the data collection. The technology thereby promoted the materialization of a naturalist view on reality in the data collection.

Restrictive and conflicting influence. At the same time, technology, including access to it, posed certain restrictions to the data collection and had some adverse effects. As one respondent stated: "If it cannot be measured, then we do not have it." The data that could be produced was restricted by the technology available to the Digital Twin developers. In all cases, the developers experienced the limits of technology: they would have liked to have certain data but did not have the technology to gather it. One of the reasons for this was that the corresponding measuring technology simply does not exist. One of the respondents stated: "there are a lot of other things we want to have, but that just does not exist at the moment". However, another prominent reason was that the technology was too expensive. Access to some databases, like certain satellite data, as well as equipment like sensors, need to be bought. In some cases, the sensors for gathering certain data are so expensive that the project team could only use a few of these sensors or none at all. This is illustrated by one of the respondents: "What you want to measure has to be financially possible. That is an important prerequisite (…) Modelers can indicate what they want to measure but it has to be financially possible. We will discuss that and if it does not work, we discuss if another variable that is clear enough would be needed. Or could we do it another way, but with less quality?". The limits imposed by the budget for the projects thus forced the developers to explore other ways to datafy the physical twin that are more economical on a large scale. However, this produces somewhat or fully different data and may come at the price of accuracy of the Digital Twin. The financial challenges thus instigated creative thinking, but also led to a data collection that might not be the most perfect version of a Digital Twin could possibly be, but to a version with fewer or sub-optimal features. The availability and capacities of technologies thus placed certain limits on how the physical twin could be datafied, and in turn, limited what the Digital Twin could show.

Also, the developers needed sufficient time to technically produce the data. In all three Digital Twin projects, "time" was mentioned as a significant restricting factor of influence on the datafication of the physical twin: the type and amount of data in the data collection depends on the time it costs to produce and process it. "[The available time] will mean that you reduce your data collection to a manageable amount." Certain methods of data production (e.g., producing soil samples) are so time-consuming that using them was not a realistic option. This meant that although the developers would have liked to have certain data for the Digital Twin and the technology was available, they still had to do without them. Additionally, some of the projects experienced extra time-pressure due to delays caused by the COVID-19 outbreak and its corresponding counter-measures which restricted access to the technology located at the university buildings.

Additionally, we found that the developers expressed a certain hesitation with regard to the use of particular technologies. One of the respondents stated: "You can create an incredible model that needs fancy data, but if nobody can properly measure that in their greenhouse, it will never work. The model has to have a certain level of realism, otherwise it won't work, but it also needs to be practical so the data should be measurable." Here, we see the technology gives rise to a tension between the view on the data collection from a scientific perspective (science as goal) and the view from a market perspective (user product as goal).

Moreover, technology inadvertently affected the data collection. When errors or bugs occur, or if equipment or sensors break, data may be erased, or simply not recorded. This occurred on one of the projects where a part of the data was lost due to a technical error.

In all cases, the projects could get their hands on less data than they would have liked due to direct technical limitations (the technology did not exist or broke down), or indirect technical limitations (the technology was too expensive or time-consuming to use). As one of the respondents stated: "We will never get to [a Digital Twin where] N = 1 due to finances and time." The consequence was that some of the initial ideas of the project teams turned out to be unfeasible and they had to change their ideas about the Digital Twin they were developing: "When I look back, I see that we reconsidered the level of complexity. Initially, we tried to do too much." Due to the limitations, the Digital Twin was reduced to a particular subset of the physical twin. Respondents described similar considerations made in all three projects. The restrictive implications of

the technology, its availability, and use, thus spilled over into the representative ambitions of the Digital Twin.

Last, the data that is produced by technology can conflict with the views of developers on the physical entity. Exemplary for such conflicts was the "cleaning" of data. "Data-cleaning" entails the removal of incorrect or corrupt data from a data set. This is needed when the measuring equipment used to acquire a certain desired type of data may be considered unreliable or prone to performance issues. One of our respondents argued that it is standard for them to "take insecurities for measurements into account. It comes down to making an estimation of how insecure your observations are. I would take a large insecurity on satellite observations. Or maybe even filter them; if I do not believe certain observations, I throw them out". They gave the example of a satellite image that showed a green field in a time when the developers are certain that there are no crops growing on that field. In this case, the green image was caused by weeds instead of crops. What is interesting here is that the data in itself is not a corrupt representation of the field, since the field was indeed green, but that it did not reflect what the developers considered the "right" cause of "green" for the physical entity. The view provided by the technology conflicted with the developers view. While the view from the developers underpins the data collection and choices for the to-be used technologies, this shows that the technologies thus have a certain influence independent of the developers in the view they materialize. Ensuring that the data collection reflects the developers' *view from nowhere* thus requires them to interfere with the data.

A view from where?

Since a lot of data is needed for Digital Twin development, the projects collaborated with other actors and used data gathered by other projects, internal and external to the university, in order to save time, money, and not overburden the physical twin. Examples of such actors are public and private research institutes, other universities, governmental data institutes, commercial parties like suppliers of hardware elements, and owners of the physical entity like farmers. This data was commonly (co)produced by these actors for their own purposes (be it the public good, product development, or for data-trading purposes) and according to their own rationale of what was important about the physical entity, and in line with their own (scientific) discipline. The project teams re-used this data or collaborated in the data production for the development of their Digital Twin. Data re-use and collaboration turned out to be a crucial part of the datafication process in Digital Twin development. One of the respondents states: "One of the prerequisites of developing a Digital Twin is that you know what data is available and what data isn't. (…) You have to have people [in the research team] who are already in that world for a few years and know what data is out there."

The re-use of data involved many data sets that were produced earlier by members of the project teams in other research collaborations. These data sets where then brought together in the Digital Twin, which required an interdisciplinary group of people to work with the data. This re-use of data by different scientific disciplines raises challenges of "science friction" (Edwards et al., 2011): the interoperability of data requires clarity for all involved what the data means. The project teams aimed to address this challenge by means of adding meta-data to the data collection. Meta-data is data about data, like information about the time and place of the creation of the data, the creator, the size, the source, etc. Its main purpose is to help others to better understand the data. One of respondents indicated that good meta-data management was pivotal because the project teams included people with different scientific backgrounds. This led to a situation were "everyone has their own kind of language. So, you don't understand each other even if you are all speaking English. (…) So, when we store the information (…) everyone needs to know what it means."

Additionally, the data re-use involved asking for consent, and depending on the data owner, having to pay for the data set. For some data sets this turned out to be costly, and sometimes complex because it was not always clear who owned the data is or how access could be provided. In some cases, the costs for data access turned out to be too expensive to access the database on a real-time basis. Also, owners of the physical entity that is twinned were sometimes looking to make money with the data that can be generated from their entity. The same goes for mediating companies that store data, like cloud services, and are looking to get certain costs funded by the Digital Twin developers. As one of the respondents stated: "What you see is that data becomes valuable." Questions of data ownership and the monetization of data and data infrastructures pose a challenge for Digital Twin developers because collecting the right data can easily become too expensive.

In case of collaboration in the data-production, the difference between the motivations and goals of these actors for producing data and the project teams can lead to conflicts: "[The team has to make] short- or long-term considerations regarding financial means; business [is more for the] short-term, fundamental research [strives for the] long-term. Both have advantages and disadvantages. You just cannot do it alone. We are not Elon Musk, Bill Gates, or Jeff Bezos. We just do not have the means." The consequence of the involvement of other actors is that they can try to exert influence on the datafication of the physical twin. When we asked our respondents if the stakeholders influenced the project, one of them mentioned that "when you work in science, that is not the intention. Of course, you have companies who are involved and interested in the subject and they can say for that matter 'can you take this along in the questionnaire', but they are not determining what we are going to

do. We [the project team] will determine that". Discussions and power struggles between data coproducers with diverse interests about what data to collect and what not, thus underlie the materialization of these data sets.

With data re-use and the collaboration with diverse actors in the co-production of data, different views and motives can be involved various sub-sets of the data collection. This results in a large and multi-faceted assembled data collection that is produced by combined, merged, and sometimes conflicting views and motives, in sum, an assembled view that can come from *anywhere*.

The Digital Twin's struggle for realism

The value of Digital Twins lies in their ability to represent the reality of their physical twin. However, as we have shown in this article, the creation of this representation requires developers to datafy a physical entity, and this is a process riddled with interpretations and choices that affect what the Digital Twin can and will show of the physical twin. In this section, we discuss three main challenges for the regimes of sight underpinning Digital Twins that we identified in the datafication process: a materialized perspective, legitimacy, and data production and power.

A materialized perspective

The aim of a Digital Twin is to provide an objective real-time representation of reality. All three projects have their roots in science, performed according to the standards of a university and aiming for objectivity. This is the *view from nowhere* in Jasanoff's regimes of sight. However, as Jasanoff already pointed out, the questions scientists aim to address already entail a certain framing and selection (Jasanoff, 2017: 11). In the case of the Digital Twin development, the Digital Twin conceptualization resonates with a naturalistic approach to the physical entity. We found this confirmed by the scientific views underpinning the data collection. While the project teams consisted of researchers with diverse backgrounds, it was apparent that the datafication process was rooted in only one or a few scientific disciplines from the life sciences that approach reality through empirical observation that can be expressed in quantified forms. This view was then materialized in a data collection by acts of recording. Technology played a coshaping and ambivalent role here. On the one hand, it had an expansive influence by enabling the various ways of producing data, while promoting the naturalist view in the datafication process. On the other hand, it had a restricting and conflicting influence by itself being the limit of what was possible in the data production, or producing data that did not match the views of the developers. The role of technology in the data production process shows that in practice the views from Jasanoff's *regimes of sight* are not "pure" views: the view that underpins a data collection is itself influenced by the options and limitations offered by technology. Despite the ambitions for Digital Twins to be holistic and realistic representations based on an objective scientific outlook, the datafication process thus already shows that their building blocks reflect a particular and only partial materialized *perspective* on reality that is co-shaped by access to technology.

Meanwhile, a Digital Twin itself is intended to be something different from a scientific study; it is intended to be a tool for improved decisions-making about its physical twin. Its goals go beyond purely analytic purposes as a Digital Twin is expected to help users optimize its physical twin. Its anticipated users are people who do not necessarily have a background in the sciences underpinning the Digital Twin, like farmers or potential diabetics. Moreover, the materialization of the partial perspective in a data collection has far-reaching consequences: only that which is datafied and can be processed in the Digital Twin's models is represented. In turn, what is not represented is not taken into account in the Digital Twin's predictions and optimization advice. Meijas and Couldry therefore argue that "[i]ssues of power permeate these apparently neutral forms of datafication" (2019: 4). This raises challenges with regard to transparency and trust for Digital Twin users. It is important that the Digital Twin data collection is transparent for laypersons so that they understand what is and is not represented by a Digital Twin. Moreover, users will need to trust the Digital Twin without being able to thoroughly assess or reproduce it. This brings us to the second issue: the legitimacy of the data collection.

Legitimacy

The view underpinning a data collection matters not only for its influence on what data is collected, but also because it provides the data collection with a certain legitimacy and authority on which people can act (Jasanoff, 2017). As we found in the three cases, the project teams needed to work together with other actors—like farmers, commercial companies, and other research institutes—to achieve the high data demands for a Digital Twin and re-use and coproduce data. With limited resources and access to only certain types of technology, the developers made to a more or lesser degree use of sensors and data sets offered by external parties. Getting closer to the Digital Twins ambitions of realism often means getting different parties and data owners involved. We expect that this will be the case for the majority of Digital Twins that are developed by public research institutes like universities. While the Digital Twins developed by project teams from a university thus seems to be underpinned by a solid *view from nowhere*, it is in practice generally underpinned by multiple views. As such, multiple views influenced what data did and did not become part of the data collection. The data produced and co-produced by the diverse actors are combined and merged into the

representation of the physical entity provided by the Digital Twin. Given the combined influence of the views of diverse actors on the Digital Twin data collection, we suggest characterizing the mode of seeing underpinning such an assembled and co-produced representation as a *view from anywhere*.

Looking at such a coproduced data collection through the lens of Jasanoff's regimes of sight shows us that the coproduction can have implications for the legitimacy and authority of the data collection. While the *view from anywhere* Digital Twin data collection appeals to the same legitimating discourse (objectivity) as the *view from nowhere*, it does so without the same institutional locus and legitimating practices for all collected data. The different data producers involved can have a different view on the to-be datafied entity. Moreover, they are not necessarily interested in the same data, nor do they necessarily have the same practices and standards for producing the data. Meanwhile, the data collection is used for the production of a decision-making tool produced by a university and thereby propagates a *view from nowhere*. This can easily mask other potential influences on the data collection. When users engage with the tools for which a *view from anywhere* data collection is used, we expect that it will be difficult for them, if not impossible, to discern if and which parts of the data collection are derived from other producers, and what interests, institutional loci and legitimizing claims underpin the data collection. This may lead to a misplaced trust of users in the authority of the data collection. They may expect the Digital Twin to be based solely on traditionally scientific legitimating practices like peer reviews, reproducibility, transparency, and the adherence to certain ethical concerns and the goal to serve the public interest, while in fact a part of the data collection lacks such validating practices and was produced with different interests in mind.

Data production and power

Interests, legitimating practices, and trust become even bigger issues if we reflect on who can build a Digital Twin. The Digital Twin cases show the challenges public research institutes are faced with in contemporary knowledge production methods. The building of Digital Twins is tangled up in a struggle with regard to who has sufficient access to data and data producing technologies. Given the high societal expectations of Digital Twins, this may become a pressing issue with potentially far-reaching social implications: with limited financial resources, universities and the like may have a harder time to produce or get access to data than large commercial enterprises that have a bigger budget. In a world where decisions are increasingly supported by data-driven knowledge, the studied cases reveal a struggle that may plague more, if not all, upcoming data-driven analytical tools: a power struggle over data production and control. Publicly funded scientific institutes may have trouble getting enough resources to produce sufficient data for these forms of knowledge production on their own. If an important part of our knowledge production moves from public institutes to commercial enterprises, this shifts a significant amount of power in the direction of the latter. Given our dependence on knowledge for engaging with the world and addressing problems, this is a serious reason for concern as these companies commonly have their own profit-driven interests and may not be very committed to the public interest. Meanwhile, the legitimating practices underpinning the *view from nowhere,* like peer review and reproducibility, may not be made possible due to the wish of companies to protect their business secrets and sources of income.

Conclusion

In this article, we explored the acts of seeing and recording, as well as their implications, in three cases of Digital Twin development in the life sciences. We used Jasanoff's theoretical framework of *regimes of sight* as a lens to draw out the different influences on and implications of the production of the data collection. By examining the acts of seeing and recording and their respective role in producing the data collection, we provided insight into the struggles of representing reality in Digital Twins.

In the cases we explored, the Digital Twin developers worked on the base of a scientific view of the physical entity, a *view from nowhere* that aims to be objective and impartial, providing a realistic understanding. The *view from nowhere* is a good fit with the conceptualization of Digital Twins as realistic real-time representations. However, we identified three main challenges that the Digital Twin datafication process poses for a Digital Twins' claims to reality:

1. The data collection embodies a *naturalist* and *partial* perspective on reality in which technology plays a strong coshaping role. Other perspectives on reality play no or a secondary role.
2. Data re-use and coproduction give rise to a *view from anywhere:* a combination of multiple views that together give shape to a data collection. This can undermine the legitimacy and authority of the data collection by lacking the institutional locus and legitimating practices of the *view from nowhere*.
3. The high data requirements of data-driven knowledge production and decision- making tools place significant power in the hands of those that produce or control data sets or data-producing technologies, including power over the manner in which reality is reflected by a data collection.

These challenges are not restricted to Digital Twins. Our study reveals the struggles of power and claims to reality that take place at the very foundation of the development of data-driven knowledge production methods that aim to objectively reflect reality. The outcome of these struggles

affects the potential and limits of the tools that are build, the perspectives they offer on reality, and the trust we can place in them. Due to limited resources, universities and the like are likely to face difficulty in having sufficient access to data and data producing technologies in order to build such tools. While data re-use and coproduction are effective ways to solve this, this does raise the challenge of potentially very different views, practices, and purposes of other parties underpinning (parts of) the data collection. To account for this impact of data re-use and coproduction, we introduced an addition to Jasanoff's regimes of sight: the *view from anywhere*. Further research is needed to explore how to best address the issues that revolve around a *view from anywhere* data collection. A promising starting point is to look back at Jasanoff's *regimes of sight*: the legitimating practices and transparency demands that underpin traditional scientific data production can be important assets to avert some of the risks that come with the *view from anywhere* datafication process.

Acknowledgments

We would like to thank Sanneke Kloppenburg for her feedback on an early version of this article. Additionally, we would like to thank our two anonymous reviewers for their valuable feedback and for giving us the means and opportunity to significantly improve this article. Last, we would like to thank the editors of Big Data & Society, in particular Sachil Singh, for a smooth and pleasant publishing process.

Declaration of conflicting interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Wageningen University, (grant number Exploratory project funding - Digital Twins, 22823).

ORCID iDs

Paulan Korenhof https://orcid.org/0000-0003-3450-1404
Else Giesbers https://orcid.org/0000-0002-2906-6850
Janita Sanderse https://orcid.org/0000-0002-4503-5237

Note

1. https://www.wur.nl/en/newsarticle/WUR-is-working-on-Digital-Twins-for-tomatoes-food-and-farming.htm, last accessed 24-08-2022.

References

Brine KR and Poovey M (2013) From measuring desire to quantifying expectations: A late nineteenth-century effort to marry economic theory and data. In: Gitelman L (ed) *'Raw Data' is an Oxymoron*. Cambridge, Massachusetts: MIT press, pp. 61–76.

Dourish P (2017) *The stuff of bits: An essay on the materialities of information*. Cambridge, Massachusetts: MIT Press.

Dourish P and Gómez Cruz E (2018) Datafication and data fiction: Narrating data and narrating with data. *Big Data & Society* 5(2): 2053951718784083.

Dourish P and Mazmanian M (2011) Media as material: Information representations as material foundations for organizational practice. In Third international symposium on process organization studies (Vol. 92).

Edwards PN, Mayernik MS, Batcheller AL, et al. (2011) Science friction: Data, metadata, and collaboration. *Social Studies of Science* 41(5): 667–690.

European Commission (2022) Shaping Europe's digital future: Destination Earth. Available at: https://digital-strategy.ec.europa.eu/en/policies/destination-earth (accessed 24-08-2022).

Floridi L (2013) *The philosophy of information*. Oxford: Oxford University Press.

Gitelman L and Jackson V (2013) Introduction. In: Gitelman L (ed) *'Raw Data' is an Oxymoron*. Cambridge, Massachusetts: MIT press, pp. 1–14.

Godfrey-Smith P (2021) *Theory and Reality: an introduction to the philosophy of science*. Chicago: University of Chicago Press.

Grieves M and Vickers J (2017) Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. In: Kahlen J, Flumerfelt S and Alves A (eds) *Transdisciplinary Perspectives on Complex Systems*. Cham: Springer, pp. 85–113.

Hacking I (1983) *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge: Cambridge University Press.

Jasanoff S (2017) Virtual, visible, and actionable: Data assemblages and the sightlines of justice. *Big Data & Society* 4(2): 2053951717724477.

Kallio H, Pietilä AM, Johnson M, et al. (2016) Systematic methodological review: Developing a framework for a qualitative semi-structured interview guide. *Journal of Advanced Nursing* 72(12): 2954–2965.

Kitchin R (2014) *The data revolution: Big data, open data, data infrastructures and their consequences*. London: SAGE Publications Ltd.

Korenhof P, Blok V and Kloppenburg S (2021) Steering representations—towards a critical understanding of digital twins. *Philosophy and Technology* 34: 1751–1773.

Latour B and Woolgar S (1986) *Laboratory life: The construction of scientific facts*. Princeton, New Jersey: Princeton University Press.

Liu M, Fang S, Dong H, et al. (2021) Review of digital twin about concepts, technologies, and industrial applications. *Journal of Manufacturing Systems* 58: 346–361.

Mayer-Schönberger V and Cukier K (2013) *Big data: A revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt.

Mejias UA and Couldry N (2019) Datafication. *Internet Policy Review* 8(4): 1–10. DOI: 10.14763/2019.4.1428

Pylianidis C, Osinga S and Athanasiadis IN (2021) Introducing digital twins to agriculture. *Computers and Electronics in Agriculture* 184: 105942.

Schleich B, Anwer N, Mathieu L, et al. (2017) Shaping the digital twin for design and production engineering. *CIRP Annals* 66(1): 141–144.

Solman H, Kirkegaard JK, Smits M, et al. (2022) Digital twinning as an act of governance in the wind energy sector. *Environmental Science & Policy* 127: 272–279.

Thomer AK and Wickett KM (2020) Relational data paradigms: What do we learn by taking the materiality of databases seriously?. *Big Data & Society* 7(1): 2053951720934838.

Van der Valk H, Haße H, Möller F, et al. (2020) A taxonomy of digital twins. In: Americas Conference on Information Systems.

Van Dijck J (2014) Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance & Society* 12(2): 197–208.

Winsberg E (2019) Computer Simulations in Science. Available at: https://plato.stanford.edu/entries/simulations-science/ (accessed October 2022).

Wright L and Davidson S (2020) How to tell the difference between a model and a digital twin. *Advanced Modeling and Simulation in Engineering Sciences* 7(1): 1–13.