Valuation Studies

Valuation Studies vol. 9(1) 2022

Contents:

© 2022 The author CC BY *This work is licensed under a Creative Commons Attribution 4.0 International License.*
Hosted by Linköping University Electronic press
http://valuationstudies.liu.se
Valuation Studies logo by Rebecca Elfast

Valuation Studies is a peer reviewed open access journal connecting several vibrant research fields working on the study of valuation as a social practice.

Editors-in-Chief: *José Ossandón*, Department of Organization, Copenhagen Business School, Denmark
Trine Pallesen, Department of Organization, Copenhagen Business School, Denmark

Board of editors:
Liliana Doganova, Mines ParisTech, FR
Martin Giraudeau, Sciences Po, FR
Claes-Fredrik Helgesson, Uppsala University, SE
Monika Krause, London School of Economics and Political Science, UK
Francis Lee, Chalmers University of Technology, SE
Andrea Mennicken, London School of Economics and Political Science, UK
Fabian Muniesa, Mines ParisTech, FR
José Ossandón, Copenhagen Business School, DK
Trine Pallesen, Copenhagen Business School, DK
Teun Zuiderent-Jerak, Vrije Universiteit Amsterdam, NL

Advisory board:
Kristin Asdal, University of Oslo, NO
Patrik Aspers, Uppsala University, SE
Diane-Laure Arjaliès, HEC Paris, FR
Frank Azimont, EMLYON Business School, FR
Alexandra Bidet, Centre National de la Recherche Scientifique, FR
Michel Callon, Mines ParisTech, FR
Alberto Corsin, Spanish National Research Council, ES
Barbara Czarniawska, University of Gothenburg, SE
Emmanuel Didier, Centre National de la Recherche Scientifique, FR

Advisory board (continued):
Marion Fourcade, UC Berkeley, US
Susi Geiger, University College Dublin
lIsabelle Huault, Université de Paris Dauphine, FR
Klaus Høyer, University of Copenhagen, DK
Peter Karnøe, Aalborg University, DK
Paul Kockelman, Columbia University, US
Michèle Lamont, Harvard University, US
Javier Lezaun, University of Oxford, UK
Celia Lury, University of Warwick, UK
Donald MacKenzie, University of Edinburgh, UK
Bill Maurer, Department of Anthropology, University of California, Irvine, USA
Liz McFall, The Open University, UK
Peter Miller, LSE, UK
Yuval Millo, University of Warwick, UK
Liz Moor, Goldsmiths, UK
Jan Mouritsen, Copenhagen Business School, Denmark
David Stark, Columbia University, US
Jesper Strandgaard Pedersen, Copenhagen Business School, Denmark
Kjell Tryggestad, Copenhagen Business School, Denmark
François Vatin, Université Paris Ouest, FR
Steve Woolgar, University of Oxford, UK and Linköping University, SE

Editorial office:
José Ossandón, co-Editor-in-Chief
Trine Pallesen, co-Editor-in-Chief
Anne Mosegaard Gravholt, editorial assistant
Department of Organization, Copenhagen Business School, Denmark

The journal IT-infrastructure is provided and maintained by the *Linköping University Electronic Press* (LiU e-press).

Web: http://valuationstudies.liu.se
E-mail, editors: editors.valuationstudies@gmail.com
Twitter: @Val_Studies

Valuation Studies 9(1) 2022: 1-10

Theme issue editorial

Digitizing Valuation

Francis Lee, Andrea Mennicken, Jacob Reilley, and Malte Ziewitz[1]

There is hardly anything that has not been digitized these days. Healthcare, finance, insurance, science, warfare, work, and social life have all been subject to technoscientific practices that process data in the form of 1s and 0s (Negroponte 1995). This shift, which is commonly glossed as "digitization," is sometimes described as radical or recent, when in fact it has been going on for almost a century (Grier 2007). By now, we are confronted with an expansive ecology of smartphones, data centers, platforms, and algorithmic computation, which is unprecedented in terms of its scale and influence. Digitization has become inextricably woven into the social fabric and practices of valuation are no exception (Kornberger et al. 2017; Lee and Helgesson 2020; Mennicken and Kornberger 2021).

But what does it mean to study digitized valuation practices? On the one hand, valuation has been digitized through algorithmically generated ratings, metrics, scores, and rankings – all of which more or less visibly drive contemporary data economies. On the other hand, it

[1] Alphabetical order, authors contributed equally.

Francis Lee is Associate Professor of Technology and Social Change at Chalmers University of Technology.
Andrea Mennicken is Associate Professor of Accounting at the London School of Economics and Political Science.
Jacob Reilley is an Assistant Professor of Accounting at the University of Groningen.
Malte Ziewitz is Associate Professor at the Department of Science & Technology Studies at Cornell University.

© 2022 The authors This work is licensed under a *Creative Commons Attribution 4.0 International License*.
https://doi.org/10.3384/VS.2001-5992.2022.9.1.1-10
Hosted by Linköping University Electronic press
http://valuationstudies.liu.se

is not clear what exactly has been changing in the process. Digitizing does not simply mean that we translate analogue practices of valuation into code. It also involves delegating the task of generating moral judgments to increasingly sophisticated technical systems. Do seemingly new practices of valuation like predictive analytics, sentiment analysis, and emotion recognition require new concepts and methods for their study? What does it take to study a phenomenon that is so obvious at a distance and yet so murky when we have a closer look?

Our themed issue comprises a series of papers which attempt to answer these questions through a set of empirically grounded studies. All papers respond to our initial call and will be published in two parts: Part 1, which is featured here, includes the first four papers, whereas Part 2 will follow in the coming months.[2] However, before we introduce the contributions in more detail, we outline six themes that summarize how we as editors have come to think about "digitizing valuation" in the course of working on this themed issue. We believe that these themes are useful as a springboard for thinking about new directions in the study of digitized valuation, and we will revisit them in an afterword to Part 2 of the themed issue.

Digitization

The idea of digitizing valuation is often associated with increasing the speed, scale, or variability of how valuation occurs. In practice, however, the situation is more complex. How do other terms like quantification (translating things into numbers), computation (bringing mathematical operations to bear on quantified things),[3] datafication (rendering things in the world as data which can be saved, edited and circulated), or automation (delegating actions to machines) relate to notions of digitization? Do these distinctions matter when it comes to scrutinizing valuation, and if so, how? Are some things easier to digitize than others? What things are excluded from the databases and processes of valuation (e.g., Bowker 2000)? How are valuation practices and metrics digitized, and what becomes excluded as an overflow or externality (Lee 2022; cf. Callon 1998)? What things, objects, people, or contexts are lost, and with what consequences? The answers to these questions should help parse out the different facets of

---

[2] Unlike traditional *Special Issues,* Valuation Studies uses *Themed Issues* to refer to a series of papers responding to a particular theme. These papers can appear in one, two, or more issues of the journal. The original call for papers related to our themed issue is available here: https://valuationstudies.liu.se/Theme_Call_Digitizing_Valuation

[3] In general language use, the idea computation is distinguished from digitization in that it applies mathematical or statistical methods to numbers. Computation can be done not only by machines, but also by humans (Grier, 2007).

digitization, and provide insights about the ways value practices are organized.

Infrastructures

While digitization is an important concept to unpack, our understanding of its characteristics and consequences can be bolstered by acknowledging the infrastructures through which it is enacted. Digitized forms of valuation do not emerge out of nowhere, and they do not appear in isolation. Rather, they are supported and changed through various forms of infrastructure (Bowker et al. 2019). Dissecting digitized valuation in this manner would allow us to ask questions about the intertwining of valuation and different means of organizing knowledge, sorting things out (Star and Ruhleder 1996; Bowker and Star 1999; Star 1999), and governance that are prevalent in infrastructure studies (Ziewitz 2012; Kornberger et al. 2019). Where can we locate the infrastructures underpinning digitized valuation? How are valuation practices infrastructured? What changes with digitization? What remains unchanged? Who is doing the infrastructuring? Which actors or what valuations are assembled and made visible through these infrastructures (Star 1991; Star and Strauss 1999)? How do new digital infrastructures reshape the practices of valuation, or the very things being valued (Kornberger et al. 2017; Reilley and Scheytt 2019)?

Power and agency

Opening up our inquiry of digitized valuation to infrastructure points us toward questions of power and agency. Valuation is never a neutral or objective practice, but is always informed by judgments, norms, and habits, as well as competing attempts to appraise and evaluate (Dewey 1939). How are some valuations granted precedence over others, and does this occur differently in digitized environments than it does in analogue ones? How does digitization shape which valuations "matter"? How do we deal with technologies that (re)configure the power over valuation? Can we explore the reconfiguration of calculative agencies (Callon and Muniesa 2005; cf. Cochoy 2008), i.e., how the digitization of valuation re-forms spaces and collective agencies that give certain actors more power than others? We might also ask how actors value different configurations of agency (Lee and Helgesson 2020), or how the actors we engage with study, analyze, and think about what a good set-up of agency would be (Ziewitz 2019; Ziewitz and Singh 2021). What new modes of intervention are enabled by digitizing valuation? In this context, it might be fruitful to explore the power effects of "protocol" (Deleuze 1992; Galloway 2004; Galloway and Thacker 2004; Kornberger et al. 2017;

Mennicken and Kornberger 2021).[4] The notion of protocol, which is borrowed from computer science, can be useful, as it helps demarcate the contradictory nature of the power apparatus that underlies and is made up by digitized valuation (see here for instance the case of platform ratings and rankings, as in the case of Uber but see also the case of changes in hotel ratings as discussed by Balsinger and Jammet in this issue).

Automation and judgment

Attending to the power effects of "protocol" also draws our attention to the ways judgment and automation are (re)configured digitally. What is the relationship between human judgment and digital infrastructures? Where is judgment possible and for whom (cf. Cochoy 2008)? Posing such questions would allow us to examine how the space for human judgment is reconfigured by digitization, and the extent to which automated systems give certain actors more space for judgment then others. We encourage more in-depth investigations of the specific situations in which human and automated judgments are valued (cf. Lee and Helgesson 2020). For instance, the automatic ranking of call-center workers' call-rate might be performed as a valuable thing to automate in some situations, while in other situations this might be abhorred. Automated, digitized categorization has been shown to be less able to accommodate conflicting rationalities. As Alaimo and Kallinikos (2020: 1398) note, the objects stemming from algorithmic categorization have the potential to form "Babel Towers." Algorithmic categorization tends to displace, relocate, and conceal human inputs; yet, at the same time, human biases and stereotypes are injected into algorithmic work, including digitized valuation work (Bechmann and Bowker 2019).

Accountability, fairness, recourse

The topics of automation and judgment raise potent questions about accountability, fairness, and recourse. When power and agency are moved around by digital infrastructures (c.f. Lee 2021), and when the boundaries between human judgment and automation become blurred, how are accountability, fairness, and recourse factored into the digitized infrastructures of valuation (Citron and Pasquale 2014; Benjamin 2019)? This is an interesting question to ask while examining digital infrastructures that are in the making. How is fairness (re)configured and (re)valued in the nascent stages of digital infrastructure formation? Who and what do we measure, and how are

---

[4] According to Galloway and Thacker (2004: 8), "protocols are all the conventional rules and standards that govern relationships in networks." In this sense, a protocol is a technology that regulates flow, directs space, codes relationships, and connects life forms (Galloway and Thacker 2004: 10).

questions of fairness addressed within these practices? How are relations of accountability reconfigured, and who or what becomes accountable to whom (Ziewitz 2012)? Often it is those people or objects that are measured that are being implicated in accountability webs while the people who construct the measurements of valuation are not (Ziewitz and Singh 2021). How might agencies and infrastructures be reconfigured so that there are possibilities for recourse? For addressing what are perceived as improper valuations?

Generativity and performativity

Finally, it can be useful to draw specific attention to the emergent properties of quantification and measurement in digitized valuation. Digital infrastructures of valuation do more than assess or evaluate (e.g., a taxi ride, a trip). They help link up and connect (e.g., service providers and users on platform organizations). In so doing they provide not only an important interface for interactions and exchanges over distance. They are also at the heart of the creation of new markets and forms of organizing (Kornberger et al. 2017; Mennicken and Kornberger 2021). They provoke the creation of new worlds through the creation of objects that are not so much the outcomes of programmatic aspirations or models, but of a surplus of data and traces, which produce new possibilities for discovery and intervention (see also Alaimo and Kallinikos 2022).

Many of the contributions in this themed issue, including the four in this first part, allow us to compare and contrast new forms of automated algorithmic valuation with older forms and practices of valuation. They enable us to take a closer look at what is new or distinctive with digitized valuation.

We open with Krüger and Petersohn and their article entitled "From Research Evaluation to Research Analytics." This article explores from a historical perspective changes in the digitization of bibliometric measurement and their effects on academic performance evaluation. In so doing it helps us reflect on what is specific about new digital forms of research evaluation. Whereas the bibliometric measurement of academic performance has been digital since the computer-assisted invention of the Science Citation Index, more recently we have been witnessing some key shifts. Citation databases are not only indexing an increasing variety of publication types, as exemplified by the proliferation of altmetric data aggregators. New ways of digital bibliometric data production and assessment have also contributed to an extension of indicator-based research evaluation towards data-driven research analytics. Focusing on interoperability, scalability, and flexibility as core material specificities of the new digital infrastructures of bibliometric evaluation, Krüger and Petersohn trace their emergence and examine their consequences for our understanding of academic performance and practices of academic performance (e)valuation.

Next, Balsinger and Jammet investigate the intertwinement of automation and judgment in the context of hotel ratings in the Swiss hospitality industry. They explore how new platform-generated valuations intersect with older forms of professional valuation. Going beyond describing the opposition between online consumer reviews and traditional judgment devices, their analysis shows that valuation on the platform is based upon a permissive hierarchical integration of a plurality of valuation poles with algorithmic valuation at its center. This shift destabilizes the evaluative landscape with regards to three issues: lack of transparency of the algorithmic ranking, the weakening and even undermining of formulaic valuation, and the issue of singularization of the online offer.

Arnelid, Johnson, and Harrison scrutinize implications of emotion recognition in digitized valuation, zooming in on the specific case of a care robot that was introduced at a Toronto hospital. The article unpacks not only how emotion detection works in this context. It also queries whose emotions are being measured, and what the use of care robots can say about the norms and values shaping care practices today. The authors show how a fragmentation and associated commercialization of care work is exemplified by the introduction of care robots. In doing so, the article explores the generative nature of valuation (e.g. in provoking certain emotional responses and new relations of accountability).

Finally, Cevolini and Esposito take us to the field of car insurance. In the insurance industry, algorithmic predictions are increasingly being used to assess the risk exposure of potential customers. The article examines the impact of digital tools in the field of motor insurance, where telematics devices produce data about policyholders' driving style. Cevolini and Esposito argue that current experimentation with such new digital tools is moving in the direction of proactivity: instead of waiting for a claim to occur, insurance companies intervene in people's behavior to mitigate risks. The authors go on to explore potential consequences of such practices on the social function of insurance, which makes risks bearable by socializing them over a pool of insured individuals. They query how such a shift can lead to an isolation of individuals in their exposure to risk, affecting in turn their attitudes toward the future, as well as broader societal understandings of fairness, accountability and power.

Moving forward, we believe that it is fruitful for the social and human sciences to attend to the dynamics between digitization and valuation for years to come. In an age of machine learning, algorithms, and big data, we need to keep exploring the themes and questions outlined here in order to "stay with the trouble" (Haraway 2016) represented by digitized valuation. By attending to and asking critical questions about the themes we outline above—digitization; infrastructures; power and agency; automation and judgment; accountability, fairness, and recourse; as well as generativity and performativity—we can start a much-needed critical inquiry into what digitization means for valuation and its study. After all, to digitize is to value.

References

Alaimo, Cristina, and Jannis Kallinikos. 2020. "Managing by Data: Algorithmic Categories and Organizing." *Organization Studies* 42(9) 1385–1407.

Alaimo, Cristina, and Jannis Kallinikos. 2022. "Organizations Decentered: Data Objects, Technology and Knowledge." *Organization Science* 33(1): 19–37.

Bechmann, Anja, and Geoffrey C. Bowker. 2019. "Unsupervised By Any Other Name: Hidden Layers of Knowledge Production in Artificial Intelligence on Social Media." B*ig Data & Society* 6(1): 2053951718819569.

Benjamin, Ruha. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Cambridge: Polity.

Bowker, Geoffrey C. 2000. "Biodiversity Datadiversity." *Social Studies of Science* 30(5): 643–683.

Bowker, Geoffrey C., and Susan L. Star. 1999. *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

Bowker, Geoffrey C., Julia Elyachar, Martin Kornberger, Andrea Mennicken, Peter Miller, Joanne R. Nucho, and Neil Pollock. 2019. "Introduction to Thinking Infrastructures." In *Thinking Infrastructures,* edited by Martin Kornberger, Geoffrey C. Bowker, Julia Elyachar, Andrea Mennicken, Peter Miller, Joanne R. Nucho, and Neil Pollock, 1–13. Bingley: Emerald.

Callon, Michel. 1998. "An Essay on Framing and Overflowing: Economic Externalities Revisited by Sociology." *The Sociological Review* 46: 244–269.

Callon, Michel, and Fabian Muniesa. 2005. "Peripheral Vision: Economic Markets as Calculative Collective Devices." *Organization Studies* 26(8): 1229–1250.

Citron, Danielle K. and Frank A. Pasquale. 2014. "The Scored Society: Due Process for Automated Predictions." *Washington Law Review* 89(1): 1–33.

Cochoy, Franck. 2008. "Calculation, Qualculation, Calqulation: Shopping Cart Arithmetic, Equipped Cognition and the Clustered Consumer." *Marketing Theory* 8(1): 15–44.

Deleuze, Gilles. 1992. "Postscript on Societies of Control." *October* 59: 3–7.

Dewey, John. 1939. "Theory of Valuation" in *International Encyclopedia of Unified Science,* Vol. 2, 4.

Galloway, Alexander. 2004. *Protocol: How Control Exists after Decentralization*. Cambridge, MA: MIT Press.
Galloway, Alexander, and Eugene Thacker. 2004. "Protocol, Control, and Networks." *Grey Room* 17(Fall): 6–29.
Grier, David A. 2007. *When Computers Were Human*. Princeton, NJ: Princeton University Press.
Haraway, Donna. 2016. *Staying with the Trouble: Making Kin in the Chthulucene*. Durham, NC: Duke University Press.
Kornberger, Martin, Dane Pflueger, and Jan Mouritsen. 2017. "Evaluative Infrastructures: Accounting for Platform Organization." *Accounting, Organizations and Society* 60(60): 79–95.
Kornberger, Martin, Geoffrey C. Bowker, Julia Elyachar, Andrea Mennicken, Peter Miller, Joanne R. Nucho, and Neil Pollock. 2019. *Thinking Infrastructures*. Bingley: Emerald.
Lee, Francis. 2022. "Ontological Overflows and the Politics of Absence: Zika, Disease Surveillance, and Mosquitos at the ECDC." *SocArXiv*.
Lee, Francis. 2021. "Enacting the Pandemic: Analyzing Agency, Opacity, and Power in Algorithmic Assemblages." *Science & Technology Studies,* 34(1): 65-90.
Lee, Francis, and Claes-Fredrik Helgesson. 2020. "Styles of Valuation: Algorithms and Agency in High-throughput Bioscience." *Science, Technology, & Human Values* 45(4): 659–685.
Mennicken, Andrea, and Martin Kornberger. 2021. "Von Performativität zu Generativität: Bewertung und ihre Folgen im Kontext der Digitalisierung." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 73(1): 451–478.
Negroponte, Nicholas. 1995. *Being Digital*. New York: Knopf.
Reilley, Jacob, and Tobias Scheytt. 2019. "A Calculative Infrastructure in the Making: The Emergence of a Multi-Layered Complex for Governing Healthcare." In *Thinking Infrastructures,* edited by Martin Kornberger, Geoffrey C. Bowker, Julia Elyachar, Aandrea Mennicken, Peter Miller, Joanne R. Nucho and Neil Pollock Vol. 62, 43-68. Bingley: Emerald.
Star, Susan L. 1991. "The Sociology of the Invisible: The Primacy of Work in the Writings of Anselm Strauss." In *Social Organization and Social Process: Essays in Honor of Anselm Strauss,* 265–283. New York: Aldine de Gruyter Hawthorne.
Star, Susan L. 1999. "The Ethnography of Infrastructure." *American Behavioral Scientist* 43(3): 377–391.
Star, Susan L., and Karen Ruhleder. 1996. "Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces." *Information Systems Research* 7(1): 111–134.
Star, Susan L., and Anselm Strauss. 1999. "Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work." *Computer Supported Cooperative Work* (CSCW) 8(1): 9–30.

Ziewitz, Malte. 2012. *Evaluation as Governance: The Practical Politics of Reviewing, Rating and Ranking on the Web*. PhD Thesis. University of Oxford, Oxford.

Ziewitz, Malte. 2019. "Rethinking Gaming: The Ethical Work of Optimization in Web Search Engines." *Social Studies of Science* 49(5): 707–731.

Ziewitz, Malte and Ranjit Singh. 2021. "Critical Companionship: Some Sensibilities for Studying the Lived Experience of Data Subjects." *Big Data & Society* 8(2): 1–13.

Francis Lee is Associate Professor of Technology and Social Change and senior researcher at the unit for Science, Technology and Society at Chalmers University of Technology in Sweden. His research focus lies on the politics of information infrastructures and knowledge production. Of particular concern for Lee is how information infrastructures—such as AI, algorithms, or big data—shape science and society. For instance he is interested in how disease outbreaks are constructed with algorithms, how AI is reshaping research practices of the biosciences, or how risky people are constructed and tracked using different types of information infrastructures. His webpage is available at https://francislee.org.

Andrea Mennicken is Associate Professor of Accounting at the London School of Economics and Political Science, and Co-Director of the Centre for Analysis of Risk and Regulation. Her work has been published in *Accounting, Organizations and Society*, *Financial Accountability and Management*, *Foucault Studies, Organization Studies, Sociologie du Travail,* amongst others. Her research interests include social studies of quantification and accounting, auditing, governance and regulation, processes of marketization, standardization, and public sector reforms. Lately, she completed an international research project exploring the changing relationships between quantification, administrative capacity and democracy in healthcare, correctional services and higher education. Recent publications include "What's New with Numbers: Sociological Approaches to the Study of Quantification" co-authored with Wendy Espeland, *Annual Review of Sociology* 2019, and *The New Politics of Numbers: Utopia, Evidence and Democracy*, co-edited with Robert Salais, 2022, Palgrave Macmillan.

Jacob Reilley is an Assistant Professor of Accounting at the University of Groningen. His research focuses on processes of quantification and digitization in non-profit organizations. He is currently researching how risk and quality are digitized in hospital management systems; more recently, he has also examined how impact is accounted for in NGOs. His work has been published in *Research in the Sociogy of Organizations*, *Critical Perspectives on Accounting*, *Financial Accountability and Management*, and *Accounting, Auditing and Accountability Journal*.

Malte Ziewitz is Associate Professor at the Department of Science & Technology Studies at Cornell University. An ethnographer and sociologist, he studies the changing role of governance and regulation in, of, and through digitally networked environments. His recent work has looked at the lived experience of credit scoring subjects, the search engine optimization (SEO) industry, and attempts at algorithmic regulation. At Cornell, he directs the Digital Due Process Clinic, a clinical research program that helps ordinary people cope with, understand, and challenge automated decision systems.

Valuation Studies 9(1) 2022: 11-46

Theme issue contribution

From Research Evaluation to Research Analytics. The digitization of academic performance measurement

Anne K. Krüger and Sabrina Petersohn

Abstract

One could think that bibliometric measurement of academic performance has always been digital since the computer-assisted invention of the Science Citation Index. Yet, since the 2000s, the digitization of bibliometric infrastructure has accelerated at a rapid pace. Citation databases are indexing an increasing variety of publication types. Altmetric data aggregators are producing data on the reception of research outcomes. Machine-readable persistent identifiers are created to unambiguously identify researchers, research organizations, and research objects; and evaluative software tools and current research information systems are constantly enlarging their functionalities to make use of these data and extract meaning from them. In this article, we analyse how these developments in evaluative bibliometrics have contributed to an extension of indicator-based research evaluation towards data-driven research analytics. Drawing on empirical material from blogs and websites as well as from research and policy papers, we discuss how interoperability, scalability, and flexibility as material specificities of digital infrastructures generate new ways of data production and their assessment, which affect the possibilities of how academic performance can be understood and (e)valuated.

Keywords: Infrastructure studies; data analytics; evaluative bibliometrics; academic performance measurement

Anne K. Krüger, Leader of the research group "New Data, New Knowledge? Infrastructures of performance measurement in science", Berlin-Brandenburg Academy of Sciences and Humanities.
Sabrina Petersohn is Postdoc at the German Centre for Higher Education Research and Science Studies.

© 2022 The authors *This work is licensed under a Creative Commons Attribution 4.0 International License.*
https://doi.org/10.3384/VS.2001-5992.2022.9.1.11-46
Hosted by Linköping University Electronic press
http://valuationstudies.liu.se

Introduction

With the promise of "using advanced data analytics and super-computer technology" for processing "large amounts of data to generate powerful analyses and visualizations on demand" (Elsevier 2021a) by Elsevier's software program SciVal, academic performance measurement seems to have entered a "brave new world" of research evaluation. Yet, one could say that bibliometric measurement of academic performance has always been digital. Since the launch of the first citation database, the Science Citation Index (SCI), in 1964, bibliometrics has been tied to computer-based citation databases.[1] The SCI was based – though still with punch cards – on the use of newly developed IBM computers to compile scientific literature based on indexing citations (Wouters 1999). So what is new about digitized valuation in academic performance measurement?

Since the 1970s, citation databases have hugely broadened their range of functionality, content, and coverage and developed into an expansive digital infrastructure. What had been conceived of initially as a new method of information retrieval has evolved into a predominant tool for research evaluation (Garfield 1964; de Rijcke and Rushforth 2015; Petersohn and Heinze 2018). The digitization of academic performance measurement has since then accelerated at a rapid pace. Technological developments such as greatly increased storage and computing capacities as well as advanced data harvesting and assessment techniques have opened up an abundance of new data sources such as books and funding acknowledgements but also downloads, twitter mentions and likes, the latter coined "altmetrics" (Franzen 2015; Haustein et al. 2016). This datafication (Boyd and Crawford 2012; Mayer-Schönberger and Cukier 2013) of academic publishing and evaluation has triggered an unforeseen dynamic of expansion and diversification in bibliometric infrastructure for academic performance measurement.

In this article, we analyse the development of digital infrastructure in evaluative bibliometrics which has contributed to an extension of indicator-based research evaluation towards data-driven research analytics. With our case of bibliometric infrastructure, we aim to contribute to the study of digitized valuation by highlighting how the material specificities of digital infrastructures influence the production and assessment of data in valuation processes. In a first step, drawing on comprehensive empirical material from blogs and websites from data providers, funders, and companies as well as on research and policy papers, we demonstrate how bibliometric infrastructure, including not only citation databases but also persistent identifiers,

[1] Godin has shown that there have also been analogue "forerunners to bibliometrics" (Godin 2006: 109) at the beginning of the 19th century when psychologists started collecting information about their disciplinary output of publications.

altmetric data aggregators, software tools, and information systems enable new modes of data production and assessment based on their distinct features of interoperability, scalability, and flexibility (Tilson et al. 2010; Büchner 2018). In a second step, we discuss how this development unfolds a "generative potential" (Mennicken and Kornberger 2021: 464) extending indicator-based research evaluation towards data-driven research analytics, influencing the possibilities of how academic performance can be understood and (e)valuated.

A short introduction to evaluative bibliometrics

Bibliometrics, the scientific discipline at the intersection of library and information science, sociology, history of science, and science policy, revolves around the application of mathematics and statistical methods to measure scholarly communication and to generate insights into the growth, structure, and development of scientific fields (Pritchard 1969; Debackere et al. 2019). Already in its formative years and fostered by the growing interest of the nascent science policy community in the 1960s and 1970s, the sub-field of evaluative bibliometrics (Narin 1976) branched out, providing methods, tools, and techniques for the quantitative measurement of academic performance in terms of its impact and output (Furner 2014; Debackere et al. 2019). After an experimental phase in the 1970s and 1980s, the use of bibliometrics in science policy and research management became a consolidated, yet continuously disputed practice in the 1990s. It nevertheless proliferated in performance-based funding schemes in national research assessment (Hicks 2012) and institutional resource allocation models (Hammarfelt et al. 2016) down to the use of individual-level metrics for getting hired or tenured, showcasing achievements and self-monitoring impact as well as obtaining funding (Nicholas et al. 2020). Furthermore, the 2000s saw the advent of university rankings as global benchmarking tools that were and still are building on citation databases (Hazelkorn 2011; van Raan 2019). Additionally, by this time, the use of evaluative bibliometrics had become institutionalized and among some stakeholders, such as research administrators, an (even too) popular practice (Gingras 2016).

These practices of quantified research evaluation rely heavily on bibliometric indicators such as the h-index or highly cited publications which have become an integral part of researchers' CVs (Nicholas et al. 2020) or on the use of publication counts and other aggregate output measures as witnessed, for instance, in the Australian performance-based research funding formula (Butler 2003). This indicator-based research evaluation generates insights into academic performance that is supposed to complement (Moed 2007; Derrick and Pavone 2013) or is feared to supplant or override judgement by academic peers, thereby moving evaluation "from a skilled operation to an automated, mechanical one" (Gingras 2016: 57).

This concern of automated analysis, however, was exacerbated by developments at the turn of the century. With the rise of the internet and subsequent digitization of academic publishing the idea of measuring scholarly impact online through webometric methods such as content analyses of web pages or hyperlink counts started to gain traction (Thelwall et al. 2006). Growing criticism regarding the narrow conception of scholarly impact as well as the growth of social media platforms spurred the development of altmetrics in 2010 (Björneborn and Ingwersen 2001; Priem 2014; Nuredini et al. 2021).[2] With the development of altmetrics and respective tools to generate them, a conceptual shift took place from the closed universe of citation databases towards a myriad of different data types such as clicks, downloads, views, tweets, mentions, or likes that were attributed relevance for indicating research performance. Yet, these data did not only extend the database for bibliometric analyses of research performance; they furthermore turned the idea of indicator-based research evaluation as measuring scientific merit *within* academia towards including research impact upon society at large.

The rise of altmetrics demonstrates two things: First, it shows that the digitization of academic publishing and communication has enabled new modes of data production for evaluative purposes. Second, it highlights how technical developments in bibliometric infrastructure can influence our understanding of what academic performance is about. While research on evaluative bibliometrics has been strongly centred on methodological questions of database coverage and quality, indicator construction, their usage, and consequences (de Rijcke et al. 2016; Moed 2017),[3] we therefore contend that the story of evaluative bibliometrics should not be told with a focus on common and alternative indicators for academic performance alone. Instead, research on academic performance measurement should also take into account the constantly progressing development of digital infrastructure that provides unprecedented data sources and respective tools to produce, process, and assess data. Focusing on the digitized bibliometric infrastructure behind academic evaluation, we ask how its constant growth affects the possibilities for academic performance measurement. We suggest that the ongoing

---

[2] Altmetrics comprises many different types of "online metrics that measure scholarly impact" (Haustein 2016: 415) which are generated on social networking platforms such as Facebook and ResearchGate, reference managers like Zotero and Mendeley, microblogging sites such as Twitter and social data sharing on Figshare or Github (Haustein 2016: 415).

[3] See for an exception Aström (2016) who has started theorizing on the relation between digital infrastructure, indicators, and evaluation practices in bibliometrics.

development of digital infrastructure gradually extends evaluative practices from indicator-based evaluation towards data-driven research analytics. The evaluation of scientific practice no longer depends on predefined indicators alone. Instead, constantly expanding possibilities in data production and assessment are becoming the drivers for how academic evaluation can be practised raising questions about the influence of data-driven analytics on the understanding and valuation of scientific practice as such.

Digital infrastructures in valuation processes

By 1995, Theodore Porter had already concluded that processes such as the production of seemingly objective performance measurement through quantification are influenced by technology: “Once the numbers are in hand, results can often be generated by mechanical methods. Nowadays this is usually done by computers” (Porter 1995: 6). This quote shows that Porter still thought of computers as assistance to quantitative evaluations. However, today, they have come to play a crucial role not only in data assessment, but furthermore in data production through automated processes extending the amounts of assessable data to unprecedented quantities. The assessment of data through automated tools (Amoore and Piotukh 2015) and, moreover, the digitized production and processing of data about social practices and individual characteristics have become a crucial feature in current valuation processes (Lupton 2016; Fourcade and Healy 2017; Kiviat 2019). The availability of technologies is constantly generating more ways of how data can be easily produced and assessed for various kinds of evaluative practices.

The production and assessment of data through digital infrastructures have already been discussed in a considerable number of studies. Already in the 1990s, Bowker, Star and Ruhleder had addressed the question of how computer-based information systems were set up to produce data to support working routines based on predefined classificatory systems (Star and Ruhleder 1996; Bowker and Star 1999). Studying the introduction of computer-assisted administration to nursing care in hospitals, Bowker and Star demonstrated the performative effect of such infrastructure, intending to categorize the full range of nursing practices. While this infrastructure makes visible and acknowledges the multiple requirements of patient care that nurses constantly accomplish, it also defines how such work has to be done, allows for controlling employees, and makes other practices that are not captured within these categories become invisible (see also Star and Strauss 1999). Digital infrastructures thus perform a specific understanding of the processes they are supposed to support. Their performativity is based on inscribed meanings and accounts of worth that their developers and providers and often even their users take for granted.

The socio-material performativity of digital infrastructures has become a crucial aspect for studying the development and implementation of software systems for administrating and evaluating work processes in organizations. Yet, we also contend that the research perspective on performativity might need to be extended due to the increasing digitization of various processes and practices of everyday life jointly with the enormous growth of computing capabilities. While much research has focused on the socio-material assemblage of technology and social practice, particularly focusing on the *social practices* of either the providers or the users of technology and technology's performative effects on their practices (Pollock and Williams 2007, Orlikowski and Scott 2008; Wagner et al. 2011; Bowker et al. 2019), it has become necessary to ask for the "generative potential" (Mennicken and Kornberger 2021: 464) of digital infrastructures. They do not only *assist* social practices of data production and assessment, but instead *generate* themselves new objects and structures in terms of unprecedented quantities of datasets and linkages between them triggering new ways of assessing them.

In his study on the introduction of the stock ticker in financial markets, Preda addresses the stock ticker technology as a "generator" of new temporal structures in financial market practices (Preda 2006: 754). The stock ticker was able to constantly present data on prices making any variation in prices immediately visible. Preda finds that this material specificity of immediate price data visualization led to a restructuring of representational language, cognitive tools and categories, and group boundaries. He moreover argues that this new technology of data production and presentation dramatically changed how financial markets were enacted. Stock ticker technology made time become a crucial factor in "playing the investing game" (Preda 2006: 768). Alaimo and Kallinikos also argue for the generative capabilities of technology. Studying the recommender system of the audio streaming platform Last.fm, they discuss how automated technologies "blur the distinction between humans and machines" (Alaimo and Kallinikos 2021: 18) within organizations. Key operations in organizations are becoming performed by technology instead of human experts. Contrary to Bowker, Ruhleder and Star and their studies on the inscription of predefined classificatory systems into technology, they highlight that recommender systems do not build on predefined music genres, but instead construct new music categories by producing data about songs and their listeners focusing on relations between them (see also Unternährer 2021).

These studies altogether highlight the generative potential of technology due to its *material specificities* that do not only perform an effect on practices through inscribed and predefined classificatory systems. They also generate new ways of creating, categorizing, and thus structuring objects based on their technical capabilities of producing linked data. Studying the development of bibliometric infrastructure in academic evaluation makes it necessary to ask not only for normative classifications and accounts of worth about academic "performance", "output", and "impact" that are inscribed into technology. But also, to understand how the constantly progressing digitization of bibliometric infrastructure is changing how academic evaluation can be realized, we need to focus on the material specificities that are created through digitization and respective technological developments.

The material specificities of digital infrastructures have already been addressed in information systems research. Tilson et al. define digital infrastructures – particularly in contrast to physical infrastructures – "as shared, unbounded, heterogeneous, open, and evolving sociotechnical systems comprising an installed base of diverse information technology capabilities and their user, operations, and design communities". They emphasize that such infrastructures "cannot be defined through a distinct set of functions [...], or strict boundaries [...]. In contrast, they are characterized by dynamism and longevity and are relational in nature" (Tilson et al. 2010: 1–2). Digital infrastructures per se are not static and predetermined in their usages and meanings. Instead, their material specificities can be characterized through three distinct features (see also Büchner 2018): Digital infrastructures can be made *interoperable* with other tools and devices depending on their application programming interface (API). This interoperability enables more and diverse uses of the same data through connecting new devices and allowing for mutual data exchange. Yet, it also allows for interconnecting multiple sets of different data and analysing the relations between them. Digital infrastructures are also *scalable*. They can be easily reduced or enlarged and new modules with new functionalities can be constantly added to an already existing system. This scalability leads to the capability to constantly produce and process various kinds of data and metadata which makes digital infrastructures highly *flexible* in their application because the meaning of these data is not predefined. Instead, data is made meaningful through the inscribed functionalities of the infrastructures and how they are put into use. Digital infrastructures thus do not produce data that can only be used in a particular context.

The interoperability, scalability, and flexibility of digital infrastructures enable the aggregation and linkage of masses of data and allow for a constant search for new ways to extract meaning from them. These material specificities make the production and assessment of data an intrinsic characteristic of digital infrastructures. Which kinds of data can be produced, how these data can be linked, and, finally, how they can be processed and meaningfully assessed is not structured through the social inscription of categories and classifications alone. It also depends on material constraints and affordances provided by technology. Yet, we do not claim that such specificities necessarily lead to specific practices.[4] Instead, we argue that we should take the role of technology seriously in its generative potential to determine what counts as meaningful and valuable.

In our case study on current developments in evaluative bibliometrics, we trace the increasing interoperability, scalability, and flexibility of bibliometric infrastructure to shed light on how these features have enabled new ways of producing, aggregating, and linking data about scientific practice generating new possibilities for academic evaluation. We focus on specific parts of this digital infrastructure namely citation databases, altmetric data aggregators, and persistent identifiers as well as software tools and current research information systems. We discuss how they enable and promote a constantly progressing extension from indicator-based research evaluation towards data-driven research analytics that might change the understanding and valuation of scientific practice as such.

From citation indices to linked data: Charting the development of bibliometric infrastructure

Our analysis of the development and material specificities of central citation databases – the most influential altmetric data aggregators, mature persistent identifiers as well as widely spread software tools and current research information systems as crucial parts of bibliometric infrastructure – rests on a broad range of empirical material. We have searched websites from companies, foundations, and other organizations dealing with bibliometrics, research policy, and data analytics. Additionally, we have studied blogs,[5] the GitHub repository, research publications on bibliometric methods, indicators, tools, and databases, as well as grey literature such as policy documents or white papers to chart the growing landscape of bibliometric infrastructure. In our data collection and analysis, we have focused on developments from the year 2000 onwards when the databases Scopus and Google Scholar emerged as first competitors to the long-lasting monopoly of Web of Science as the only provider of citation data transforming bibliometric infrastructure into “a crowded marketplace” (de Rijcke and Rushforth 2015).

---

[4] See for arguments against technological determinism MacKenzie and Wajcman (1999).

[5] These blogs comprise, in particular, the Bibliomagician and Leiden Madtrics as well as blogs from Crossref, ORCiD, and ROR.

Citation databases

The fundamental backbone of bibliometric infrastructure consists of citation databases that collect and store data and metadata about publications, their authors, and citations (see Figure 1). In 1992, the original Science Citation Index was acquired from its inventor Eugene Garfield and his Institute of Scientific Information by the information company Thomson Reuters and renamed Web of Science. Web of Science has broadened its coverage and selection policy beyond the initial focus on journal articles and also added conference proceedings, books, and data over the years (Birkle et al. 2020), thereby extending its content towards applied sciences, arts and humanities as well as social sciences. The ownership of the citation database changed again in 2016 when the company Clarivate acquired Web of Science. Web of Science is accessible to subscribers by a web interface for basic searches and for referring to the Journal Impact Factor, h-index and other citation metrics. APIs allow "power users" in research management to apply more advanced searches and analyses (Birkle et al. 2020). It currently covers up to 155 million records of publications (Martín-Martín et al. 2021).

For more than 40 years, Web of Science has remained the one and only citation index available. Its monopolistic position was challenged in 2004 by the international publisher Elsevier which launched its curated, selective citation database Scopus. Scopus contains at least 76 million records (Martín-Martín et al. 2021) and has become an equally important resource for bibliometric large-scale analyses and policy purposes such as national and institutional research assessments, governmental policy analyses and reports as well as university rankings (Baas et al. 2020). Like Web of Science, it incorporates citation and journal metrics, some of them especially developed based on Scopus data, like the CiteScore (Teixeira da Silva and Memon 2017). APIs provide limited or full access to citation records, search functionalities, and download options depending on the subscription model chosen (Baas et al. 2020).

Shortly after the introduction of Scopus, Google Scholar was launched by big tech giant Google. Google Scholar differs significantly from the traditional citation databases. Contrary to Web of Science and Scopus, it does not only provide free access to its database with a simple, easily accessible and usable web interface. Google Scholar also indexes online available research documents of any kind of quality, form, and type, regardless of whether the content is peer-reviewed or not or even published in a journal. It represents the first academic web engine of its kind (Orduña-Malea et al. 2014). Instead of offering curated content following the principle of selectivity, Google Scholar applies an unsupervised indexing process based on automated bots crawling the web. Citation counts can only be provided based on the extraction of cited references from retrieved full texts which impacts not only data quality but also the computation of respective citation metrics (López-Cózar et al. 2019). Although no official figures exist, it is estimated that it contains more than 300 million records (Martín-Martín et al. 2021).

Another free citation database that also functions like an academic search engine based on Bing's web crawling infrastructure was Microsoft Academic Search, which was developed in 2006 by Microsoft in response to Google Scholar. The database developed at a rapid pace: being limited at first to computer science and technology fields, it expanded to more subject categories based on agreements with different source providers and improved technical features such as browsing capabilities. Similar to the aforementioned databases, Microsoft Academic Search also contained bibliometric performance indicators as well as visualizations of publication, citation and authorship networks (Orduña-Malea et al. 2014). While this version was silently obsoleted in 2012, Microsoft opted for a relaunch in 2016 with a new design and motivation. Cloud-computing and artificial intelligence technologies formed the technological backbone of Microsoft Academic and of its core component, the Microsoft Academic Graph. The graph was a network-like structure comprising bibliographic metadata and the relationships among them (Wang et al. 2020). By means of machine reading and artificial intelligence all Bing-indexed webpages, metadata feeds, and publishers were text-mined and organized into the graph (Microsoft 2021b). The Microsoft Academic Graph covered around 255 million records from all stages of research publications, ranging from preprints to reprints (Orduña-Malea et al. 2014; Wang et al. 2020). Via an API data could be retrieved either as raw data or as pre-processed data (Hug et al. 2017). However, Microsoft Academic was discontinued in 2021 (Microsoft 2021a). Non-profit initiatives such as the database OpenAlex have stepped into this void, using data from the Microsoft Academic Graph and combining it with more data gathered from other sources and web crawls. It was launched in spring 2022 by OurResearch (OpenAlex 2022).

Besides OpenAlex the most recent addition to the database backbone of bibliometric infrastructure is Dimensions by Digital Science, which was launched in 2018. This database differs significantly from the others because it is not a strictly bibliographical database but also contains a wider set of document types such as awarded grants, policy papers, clinical trials and patents next to scholarly publications and their citations. It sources data from a variety of organizations, indices, and initiatives. The proclaimed ambition of Dimensions is to broaden the narrow frame of publication and citation analyses (Herzog et al. 2020). The database now amounts to over 105 million records (Martín-Martín et al. 2021). Dimensions aims for "not only aggregating millions of previously siloed records but also creating links between these records based on increasing occurrence of persistent identifiers, as well as AI-based techniques, and by mining relationships referred to in full text" (Herzog et al. 2020: 390). It thus enables the linkage of different datasets providing encompassing metadata about publications, their authors, their funding, and resulting "output" such as patents, clinical trials, or policy references. The developers refrain from creating their own metrics and indicators as do Web of Science, Scopus and Google Scholar, or from providing data for university rankings. However, Dimensions actively encourages large-scale bibliometric analyses including indicator development by the scientometric research community via a dedicated API for data retrieval and analysis (Herzog et al. 2020: 390). Dimensions is accessible in its free version from online interfaces that allow for contextual search and data visualizations for research purposes. Institutional reporting and analyses or consulting are, however, only possible based on subscriptions (Herzog et al. 2020: 390).

Although differing in their coverage and selection policies, these citation databases have grown by millions of records since their inception. The increasing data volume allows for flexibly deriving citations from increasingly varied sources and publication types as well as other forms of output such as patents or policy papers. While most of these databases incorporate and effectively disseminate their own set of metrics and indicators (Jappe 2020), developments such as the Microsoft Academic Graph, OpenAlex or Dimensions' approach of "linked research data from idea to impact" (Dimensions 2021a) demonstrate that networked graphs become increasingly important extending indicator-based research evaluation towards research analytics. Instead of predefined indicators, such ways of producing, linking, and presenting masses of data display correlations that provide the ground for "discovery and analytics" (Dimensions 2021a) without any pre-given operationalization of what academic performance is about. While the providers of OpenAlex and Dimensions promote the scientific use of their linked data, fee-based licence models also exist for commercial and large-scale purposes. With these licence models, database providers also foster research analytics' entrance into the market for data-driven research intelligence.

Besides the linkage of different datasets, all of these citation databases display a high amount of interoperability and scalability. As we highlight in the following sections, they incorporate persistent identifiers to enhance data quality and also offer APIs for data retrieval and analysis. These interfaces allow for their integration in software tools and current research information systems that provide meaning to these masses of linked data on research and researchers.

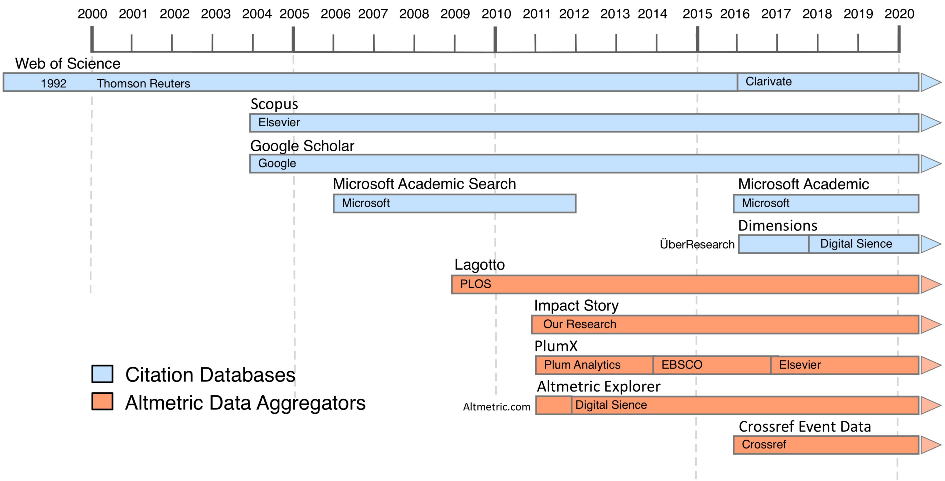

Figure 1. Development of citation databases and altmetric data aggregators.
Source: Authors' own illustration.

Altmetric Data Aggregators

With the digitization of academic publishing, it has not only become easier to collect data about publications, authorship, and citations. Also, the communication behaviour of researchers including "publishing, posting, blogging, scanning, reading, downloading, glossing, linking, citing, recommending, acknowledging" (Cronin 2005: 196; cited in Haustein et al. 2015) has turned into a new source for tracking the usage of research publications. Since then, new ways of producing data on research usage have been established. So called altmetric data aggregators[6] collect data resulting from views, downloads, blog posts, tweets and other digitally visible forms of usage based on sources such as bibliographic reference managers, social media platforms or even policy documents and make them publicly available (see Figure 1). While it takes some time for publications to become cited, these data are propagated as measuring research impact in real-time by the scientific community and even society at large (Priem et al. 2010).

The Public Library of Science (PLOS) became the first database to produce data about the online usage of research articles. In 2009, they started the open source application Lagotto to provide data based not only on their own counts of views and downloads (Lagotto n.d. a) but also on other external sources such as the bibliographic reference manager Mendeley, the social media platform Twitter, and Crossref.

[6] See for a comprehensive comparison of different aggregators Zahedi and Costas (2018) and Ortega (2020).

They classified these data into the categories "viewed, saved, discussed, recommended, and cited" (Lagotto n.d. b) as these categories were intended to represent different forms of user engagement and thus of impact (Lin and Fenner 2013).

These first developments were followed by the "altmetrics manifesto" from Jason Priem and colleagues in 2010 which called for enlargement of the focus of measuring research impact beyond the often problematic citation counts for better giving credit to researchers and their impact (Priem et al. 2010). In 2011, jointly with Heather Piwowar, Priem founded the non-profit organization ImpactStory.[7] ImpactStory is an online platform where researchers can create a profile including different kinds of their research output like publications and pre-prints as well as datasets and software. Based on a software tool called "total-impact" Priem and Piwowar claim to be able to "capture unprecedented amounts of data showing all sorts of uses of all sorts of products by all sorts of people" (Priem and Piwowar 2012). ImpactStory provides researchers but also other users such as funders with information about the usage of these research items including various sorts of mentions in academic contexts, the geographical reach of their research, or their open access activities. Priem and Piwowar highlight that ImpactStory does not only provide its users with numbers but also puts these numbers in context by comparing them with achievements of other researchers (Priem and Piwowar 2012). In addition, ImpactStory allows for the reuse of its data providing a free API.

ImpactStory, however, is not the only online platform that aggregates different sorts of data on the usage of research. In the same year, two further altmetric data aggregators were launched based on a similar idea. Altmetric.com displays article-level metrics as a colourful "altmetric donut" with each colour highlighting a different kind of source where an article has been mentioned. It furthermore provides the *Altmetric Attention Score*, which is "an automatically calculated, weighted count of all of the attention a research output has received" (Altmetric n.d. a). Like ImpactStory, Altmetric.com equally addresses not only researchers but also publishers, research organizations, and funders. In 2012, the start-up became part of the Digital Science portfolio (Wikipedia 2021a). Altmetric.com provides a free API for scientometric research, but access can also be purchased by commercial users (Altmetric n.d. b).

Another altmetric data aggregator has been developed by the company Plum Analytics. With their tool PlumX, they provide altmetrics for a broad variety of research objects (Herb 2019). PlumX is thus neither focused on the individual researcher nor on research

[7] Today, the organization has been renamed OurResearch, with ImpactStory as one of their products.

publications. Nonetheless, it can be used to track the altmetrics of individual research output. Since 2015, with the PlumX suite, it has also included a benchmarking tool for research organizations (Wikipedia 2021b). Unlike ImpactStory, it has no freely accessible API. Instead, since it was acquired by Elsevier in 2017, which now uses PlumX across their journals to display the metrics of research articles, it is open only to Scopus subscribers (Scopus 2019). However, its metrics are publicly accessible.

A more recent development is Crossref Event Data, which started in 2016. The term “event data” refers to a similar kind of data deployed by other altmetric data aggregators such as a mention in a blog post or a comment on a social media platform. Contrary to Altmetric.com and PlumX, Crossref Event Data neither offers website plugins nor provides metrics or any other sort of data analysis. Instead, Crossref allows access through an open API highlighting that they only “provide the unprocessed data – you decide how to use it” (Crossref 2020a). They explicitly refrain from presupposing distinct uses of their metadata. Yet, they regard “data intelligence and analysis organisations” (Crossref 2020a) among their potential users.

These different altmetric data aggregators have in common that they seek to complement or even to outstrip traditional citation indices as the primary source of information about academic performance fostering data exchange and interoperability with citation databases and other software tools for research assessment by providing APIs. They flexibly build on the rapidly changing digital traces of research items and any kind of online interaction with them to generate meaning about impact within and beyond academia. To this end, they create new metrics and indicators such as the *Altmetric Attention Score* or *ImpactStory achievements*. However, these efforts to establish digital traces as meaningful indicators have also become ends in themselves[8] leading to a “lack of a theoretical foundation coupled with (…) pure data-drivenness” (Haustein 2016: 418). Altmetrics are thus not based on methodologically sound operationalization defining what they can or cannot indicate. Instead, their production is determined by technical affordances and commercial interests leaving open the question as to how these data can actually be interpreted and used for evaluation purposes. Indicators are thus not only the basis for research evaluation but themselves a product of data-driven analytics.

Persistent identifiers

Persistent identifiers are digital markers that were developed to unambiguously identify researchers, research organizations, and

[8] Additionally, academic social networking sites like ResearchGate and Academia.edu use data analytics to predict and foster social interactions among members attempting to identify future research trends (Delfanti 2021).

research objects (see Figure 2). The idea followed from the constant digitization of academic publishing and the growing digital storage of research output such as publications but also datasets, software, and other research objects. Persistent identifiers were invented from the 1990s onwards to address challenges resulting from the problem, also known as "link rot" (Klump et al. 2017: 1), where internet references did not permanently link an object to a persistent URL; URLs could change making the linked object inaccessible and irretrievable (Dellavalle et al. 2003). It was feared that research output might get lost if there was not a reference system for online publications and other digital research objects reaching beyond the unstable web links.

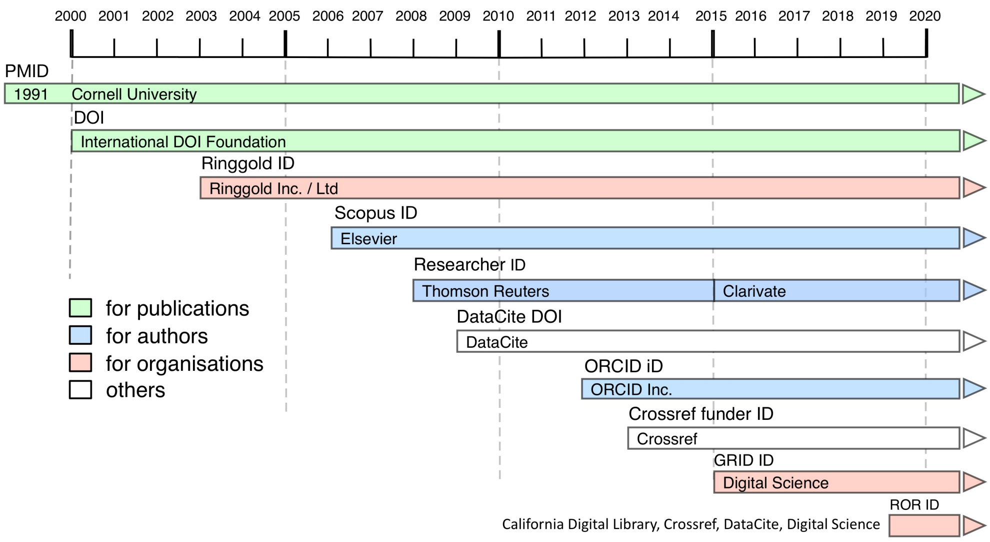

Figure 2. PID development.
Source: Authors' own illustration.

The PubMed ID (PMID) and the Digital Object Identifier (DOI) became the most prominent ones for research publications with the DOI as the only persistent identifier that is currently used by academic journals across publishers and disciplines. The DOI consists of letters and numbers that provide a unique and unambiguously identifiable signature for a particular research publication. It also provides metadata such as the author and the place where a research article is published that become inextricably linked to the article. The DOI has thus become the core technology for the digital academic publishing system allowing for the unambiguous identification and correct referencing of a publication and the constant monitoring of its output (Paskin 2010). In 1997, the International DOI Foundation (IDF) was established, which to date manages the assignment of DOIs to research publications (International DOI Foundation 2015).

The DOI gained traction particularly through the foundation of Crossref (before CrossRef) in 2000 (Crossref 2020b). Crossref was founded as an initiative of influential publishers who saw a need to adapt to the age of digital publishing. They started to provide links between articles and their references across journals of different publishers that were only possible based on the accurate identification of publications using the DOI as their key signature (Meadows et al. 2019: 3). The unambiguous identification of research publications facilitated the tracking of citations in other publications. To date, the IDF has assigned approximately 257 million DOIs to digitally as well as physically available objects through several registration agencies (International DOI Foundation 2021).

Besides institutionalization of the DOI as a standard marker for research publications, there are constant new attempts to broaden its scope or even to establish further persistent identifiers. While the DOI was originally designed for identifying research publications, organizations such as DataCite, which was founded in 2009, are attempting to enlarge the scope of the DOI towards further research objects such as research datasets. Crossref has moreover started to build a persistent identifier for funding bodies. They promote these efforts with the idea of having "transparency into research funding and its outcomes" (Crossref 2020c). Currently, there are, in particular, two further persistent identifiers which are pushed to the fore: the Open Researcher and Contributer ID (ORCID) for researchers and the Research Organization Registry (ROR) for research organizations.

The ORCID iD is designed as a persistent identifier for researchers. It was launched in 2012 and is operated by the non-profit organization ORCID Inc. It was founded by major publishers like Elsevier and the Nature Publishing Group but also by research organizations such as the European Molecular Biology Organization (EMBO) and the European Organization for Nuclear Research (CERN). The ORCID iD was developed as an overarching identifier based on software adapted from Thomson Reuters' ResearcherID system[9] and is now open source. By allowing to unambiguously identify authors of research publications, it responds to the problem that names of researchers are not unique, can be spelled differently and can change over time making it difficult to relate research publications to their authors (Wikipedia 2021c). To date approximately 11 million IDs have been assigned to authors (Wikipedia 2021c).

To get an ORCID iD researchers need to register themselves. On registering, researchers are provided with a profile to which they can

[9] Like the Web of Science ResearcherID, Elsevier has also set up a proprietary Scopus ID.

add their publications. ORCID has furthermore started to encourage researchers to integrate into their profile additional information about their CV, their funding, and their entire research output. Registered researchers can also agree to make their information available to organizations such as publishers, funders, or research organizations that have obtained ORCID membership. Through a member API ORCID members can have access to the profiles of individual researchers and their data. With the users' permission they can also include additional information to researchers' profiles. The ORCID iD is thus attempting to become an inclusive record of research careers encompassing a variety of information about individual academic careers and achievements. It is furthermore already "routinely used by academic-facing platforms as an authentication tool (such as the data repository Zenodo and some journal peer review systems), by publishers and journals to track article progress with authors, by institutions to build researcher performance profiles and also by research funders" (Klump et al. 2017: 3) that have started to include this information in their application process. In addition, research organizations have started "to update ORCID records and to register their employees and students for ORCID identifiers" (Klump et al. 2017: 8) making this information usable for internal monitoring and external reporting.

Persistent identifiers are developed not only for researchers but also for research organizations. The first identifier was the Ringgold ID which was established in 2003 at the request of the publishing industry to make institutional subscribers to publishers unambiguously identifiable because, like authors, organization names too can be spelled differently and change over time (Ringgold Inc. 2021). While Ringgold is designed to serve the needs of publishers helping them to connect different sets of information about the same customer (Ringgold Inc. 2021), the Global Research Identifier Database (GRID) was implemented by Digital Science in 2015. The GRID ID identifies organizations through information extracted from research funding grants and research paper affiliations and adds metadata such as "established dates, name aliases, acronyms and geolocation" as well as "links to external webpages such as Wikipedia and official websites" (GRID 2021a) and further persistent identifiers to them. It is exclusively linked to the database Dimensions from which it obtains data for creating GRID IDs. Simultaneously, the GRID ID can be used to draw data on organizational affiliations from the Dimensions database and to attribute it to a particular organization enabling the creation of an organization's record (GRID 2021b) which can be used for institutional reporting. The newcomer among the persistent identifiers for organizations is the Research Organization Registry (ROR), which was only established in 2019 based on an initiative by 17 organizations, among them Crossref, DataCite, and ORCID

(Ferguson et al. 2019: 14). To get started, the ROR relied on data from GRID, but was designed as an "an open, sustainable, usable, and unique identifier for every research organization in the world" (ROR n.d. a) aiming for the inclusion of comprehensive metadata. Its proclaimed goal is to provide a "proper description of relationships between contributors, contributions, research sponsors, publishers, and employers" (ROR n.d. b).

This short overview highlights that persistent identifiers simultaneously result from and contribute to the enormous growth of data and metadata about research and research practice. They serve to improve data quality by unambiguously identifying people, organizations, and objects, rendering data usable for different purposes. In addition, their registries provide APIs that allow for their interoperability with other systems and devices in various contexts. They are furthermore designed to be machine-readable[10] facilitating data processing and assessment through other devices. Persistent identifier registries also constantly produce new data through encouraging new entries within existing registries and the provision of additional metadata. Moreover, new persistent identifier registries are set up for further classes of objects such as data, software, grants, or conferences[11] contributing to the constant scalability of bibliometric infrastructure.

Persistent identifiers therefore play a decisive role in the linkage of data about research, researchers, and research organizations. Research practice can now be mapped from research funding to research results assigning output to individual researchers and research organizations.[12] They are announced as "an essential tool for resource management […] to ensure that the benefits of investment in research can be distributed and harvested over the long-term" (Dappert et al. 2017: 2). They are furthermore projected as contributing to the development of new "metrics around usage, reusage and other sorts of relationships between research objects" (Klump et al. 2017: 2). Persistent identifiers have become an indispensable means not only for attributing credit to researchers and research organizations for their scientific achievements, but also for making them much more accountable for the money they have spent.

---

[10] See Meadows et al. (2021).

[11] See for an overview about ongoing initiatives Ferguson et al. (2019).

[12] Check for initiatives such as the PID graph (Fenner and Aryani n.d.) or the Research graph (Research Graph Foundation n.d.).

Evaluative software and current research information systems

Bibliographic, citation, and altmetric data provided by citation indices and altmetric data aggregators are processed, linked, and analysed by a variety of software tools and current research information systems (CRIS) (see Figure 3). Software development sets in as early as the 1980s, since when it has accelerated and diversified. Already in 1997, Sylvan Katz and Diana Hicks observed the emergence of so called "desktop bibliometrics"[13] where "[a]dvanced scientometric tools are moving from the realm of the privileged few with access to mainframe and minicomputers to the desktop of researchers equipped with personal computers" (Katz and Hicks 1997: 141).

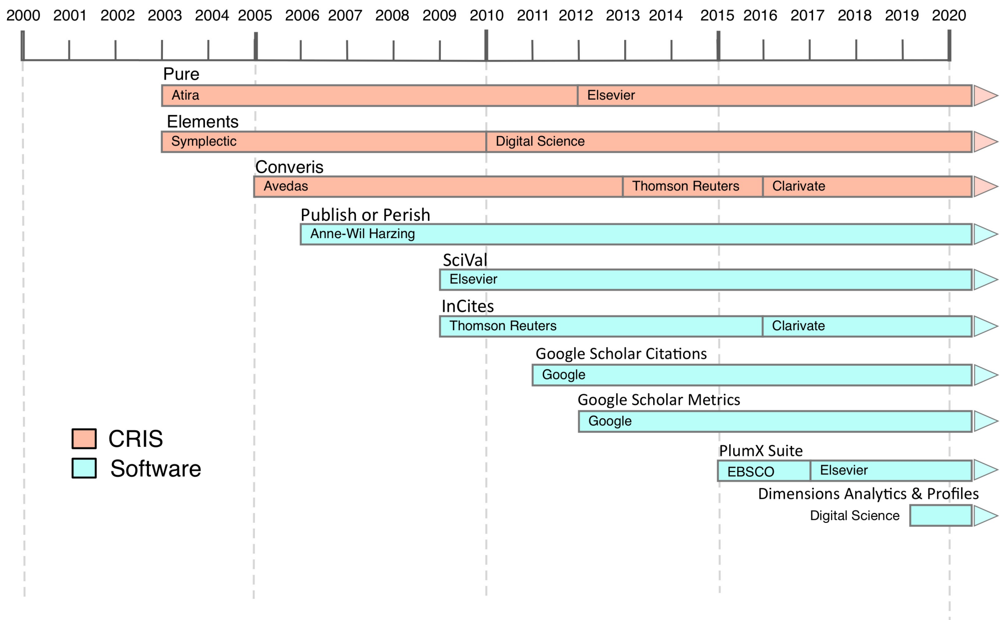

Figure 3. Development of software and current research information systems.
Source: Authors' own illustration.

The first major software product explicitly developed for evaluative purposes dates back to 2006, when management professor Anne-Wil Harzing developed the free software package Publish or Perish. It made citation analyses and a set of impact and output metrics based on bibliographic data from Google Scholar available to a wide audience ranging from individual researchers to librarians and research

[13] The term has recently been used in a different manner, denoting the application of bibliometrics by research managers and policy analysts, often including uninformed or even misuse of indicators (Bornmann 2020).

administrators (Harzing.com 2016a). The software had been initially designed to include research from disciplines that were not covered adequately in Web of Science and, in the meantime, has been extended to include Microsoft Academic, Scopus, and also Web of Science as databases for calculating individual-level impact metrics (Harzing.com 2016b). Invented one year after the introduction of the h-index, *Publish or Perish* has been and still is prominently used as an h-index calculator to support decisions for promotion, tenure, or funding applications (Harzing.com 2021).

Three years later, evaluative tools were rolled out on a larger scale through introduction of the commercial web-based software suites *SciVal* by Elsevier (Relx 2009) and *InCites* by Thomson Reuters (ISI Web of Knowledge 2009) in 2009. They permit a wide range of analytic functionalities packaged into several modules depending on the chosen subscription model of their users. They offer individual and organizational performance profiles, global comparisons with other research organizations, or expert searches (The Scholarly Kitchen 2014). They are designed to evaluate research productivity, research collaborations and impact as well as to offer benchmarking and reporting functionalities (Clarivate Analytics 2019a). In 2019, Digital Science also introduced a tool for research evaluation with *Dimensions Analytics* and *Dimensions Profiles*. The web applications build on Dimensions' data and can be used for complex analyses or for finding experts for reviews and collaborations and showcasing institutional research. *Dimensions Analytics* supports data exports to bibliometric mapping software and has integrated features from Altmetric.com (Dimensions 2021b, 2021c).

Drawing on data from the respective citation index of their providers these software tools enable the computation of research output and impact analyses as well as benchmarking functions (Clarivate Analytics n.d.; Elsevier 2021b). Research managers and administrators, academic librarians, and researchers themselves are the main targeted user groups for these products (Leydesdorff et al. 2016; Petersohn 2016). Their dashboards provide "enhanced visual data analysis" (Dimensions 2021b) with tables and multiple visual components based on graphs, maps, profiles, and plots, making them easily applicable.

Yet, the market for evaluative bibliometrics software is not dominated by proprietary products with restricted access alone. In 2011 and 2012, Google launched its free citation service *Google Citations* connected to *Google Scholar* profiles and the journal ranking *Google Metrics,* both delivering h-type and more citation metrics for authors and journals (Goldenfein et al. 2019). Compared to the three other software tools its functionalities are, however, limited. A major development in the market for analytical software tools has been triggered by artificial intelligence technologies such as machine learning and natural language processing that have not only become central technologies in the underlying databases but also in software programs. *SciVal* and *InCites* both proclaim to be "next generation analytics platforms" by incorporating these technologies (Elsevier 2021c; InCites 2021).

Current research information systems represent a distinct category within the bibliometric infrastructure because they consist of an integrated database and information system with a user interface for different applications. They integrate several different external and internal data sources such as bibliographic databases as well as internal human resources and financial systems for providing reports and producing outputs such as CV exports or content for organizational websites showcasing research (Sivertsen 2019). The most prominent current research information systems are Pure developed by the Danish company Atira (Relx 2012), Elements as a product of the British start-up Symplectic (Research Information 2015), both dating from 2003, and Converis, which was developed in 2005 by the German company Avedas (Information Today 2013).

These systems assume an increasingly important role not only in research reporting, assessment, and information management at the organizational level but also in supporting national research evaluation exercises as well as tenure programmes (Fondermann and van der Togt 2017; Kaltenbrunner and de Rijcke 2017; Lim 2021). Their potential has been recognized by Elsevier, Digital Science, and Thomson Reuters/Clarivate which acquired the three current research information systems, respectively in the years 2010 (Digital Science and Elements), 2012 (Elsevier and Pure), and 2013 (Thomson Reuters and Converis). As opposed to the administratively often less visible usage of evaluative software tools like *SciVal* and *InCites* in research organizations, current research information systems increasingly come with openly communicated, incentivized compliance policies in universities to foster digital collection and registration of research information. They furthermore increasingly represent a passage point for academics in research organizations that are required to register metadata about their research activities to be eligible for tenure programmes, promotion, or related assessment frameworks (Fondermann and van der Togt 2017; Kaltenbrunner and de Rijcke 2017; Piromalli 2019), or for having their research presented on the research organization's website.

Evaluative software tools and current research information systems thus play a key role in producing and assessing bibliometric data and providing meaning to it. They are tightly linked to and highly interoperable with other components of the bibliometric infrastructure. They provide APIs for connecting with other tools and draw on persistent identifiers to flexibly incorporate and link new (meta)data. Being structured in modules to perform distinct functions such as benchmarking or collaboration analysis, they are easily scalable. Whereas analytics software is confined to analysing data at the organizational level, current research information systems can even be scaled up from organizational to national level by means of integrating multiple data sources, enabled by APIs and persistent identifiers.

How evaluation is exactly done, however, is not a matter of only data content or indicators. While traditional indicators still form an integral part of these tools, claims for "predictive research analytics" have been raised. Providers of evaluative software contend that their advanced technologies not only allow for retrospective performance measurement. They also enable predictive analyses of future developments ranging from discovering trending topics to detecting potential high impact research. *Dimensions Analytics* is advertised as providing "enhanced discovery tools […] to deliver a full picture of past, current, and future research" (Dimensions 2021b). *SciVal* even claims to "enable […] users to envision alternate research groups by 'dragging and dropping' any researcher across the globe into hypothetical teams and gauge expected changes in performance by benchmarking 'fantasy' groups against existing groups" (EurekAlert 2011). Data scientists discuss "intelligent bibliometrics" as a promising (and profitable) new field arguing that "[t]raditional bibliometrics profile key topics and players using citation/co-citation and co-word statistics, but fail to identify complicated relationships to explain 'why' and 'how'" (Zhang et al. 2020: 1259). They claim that "[n]ovel bibliometric approaches, with the aid of advanced information technologies (e.g., machine learning and streaming data analytics), create new opportunities to uncover such relationships" (Zhang et al. 2020: 1259) enabling new kinds of complex analyses of research trends and future performance. Currently, publishers especially explore the potential of predictive analyses based on data from both citation databases and software tools for improving the performance and impact of their journals (Clarivate Analytics 2019b; Aspesi and Brand 2020). Yet, these predictive analyses might slowly be extended to the realm of national and organizational research assessment (Aspesi et al. 2019). The generative potential of software tools and current research information systems to constantly produce and link data on research and research practice has thus become a playground for testing new ways to channel research collaboration and to predict academic performance.

From indicator-based research evaluation to data-driven research analytics

Charting the development of bibliometric infrastructure within the last two decades, we have shown that citation databases, altmetric data aggregators, persistent identifiers, evaluative software tools, and current research information systems have experienced enormous growth in their content and functionalities. Citation databases are indexing an increasing variety of publication types, extending their coverage of subject fields and broadening their scope towards additional research products and outcomes such as books, patents, and more. Altmetric data aggregators are producing data on the reception of research outcomes in academia and society at large that are supposed to trace the "impact" of research beyond citations. Machine-readable persistent identifiers are created to unambiguously identify researchers, research organizations, and research objects linking them to additional metadata. And evaluative software tools and current research information systems are constantly enlarging their range of functionality to make use of these data and extract meaning from them.

Yet, it is not simply the sheer growth of these technologies for producing and assessing data about scientific practice and outcomes that has contributed to an ongoing proliferation of performance measurement in academia. It is in addition the increasing interoperability, scalability, and flexibility of these technologies and the datasets they produce that has moreover augmented the possibilities for academic evaluation. These material specificities of bibliometric infrastructure have generated a significant shift in the possibilities for practising evaluation of researchers and research organizations, giving way to data-driven research analytics based on what is digitally accessible and assessable.

The *interoperability* of different datasets and software tools through APIs and persistent identifiers allows for the linkage of various and constantly growing datasets through which data on researchers, research conditions, and research outcomes become related to one another. These linkages provide information not only about publication practices and their reception. They also strengthen belief in "return on investment". Relating particular grants and other sources of funding to researchers and research organizations allows for questions about the adequate allocation of resources rendering not only researchers but also funders accountable for how they spend their money.

Interoperability is also made available between databases and different software tools which provide the functionalities to draw meaning from these data. This interoperability of databases and software allows for the *scalability* of bibliometric infrastructure making it possible to constantly attach new data and functionalities to existing infrastructure. The scope of academic performance measurement can thus be permanently extended not only from the micro-level of individual researchers to comparisons between entire research organizations worldwide, but also in terms of new ideas for evaluation criteria and the evaluated subjects. The availability of new data such as mentions in policy documents (Overton n.d.) generates the potential to make these data usable for "policy impact" as a new evaluative criterion (Tattersall and Carroll 2018).

The scalability of the software and its functionalities also depend on the *flexibility* of data production and usage. Evaluative software tools and current research information systems exhibit generative potential by constantly extending their functionalities to integrate and assess new kinds of data and to extract meaning from them, making practices of research assessment increasingly data-driven. Moreover, the interoperability of databases and software through APIs and the machine-readability of data facilitated through persistent identifiers, their scalability, and their constant enlargement through the flexible integration and construction of new data generate an urge for prediction and trend analyses rather than retrospective evaluation of past achievements operationalized along predefined indicators. Evaluation of the past is turned into predictive analytics of the future.

These features of interoperability, scalability, and flexibility hold the generative potential to change academic performance measurement from indicator-based evaluation towards data-driven research analytics. They provide the material means for generating new evaluation categories as well as belief in the possibility of calculating and predicting successful research. Research analytics therefore not only claim to evaluate past research but generate an understanding of research practice as a predictable enterprise.

Conclusion

With our study on the development of bibliometric infrastructure, we discussed how the interoperability, scalability, and flexibility of bibliometric infrastructure contribute to an extension of indicator-based research evaluation towards data-driven research analytics highlighting how the material specificities of digital infrastructure generate new possibilities for the production and assessment of data in valuation processes. We argued that technology does not only have a performative effect on *how* evaluation is practised through predefined indicators and their inscription into technology but can furthermore generate a new understanding of what is actually evaluated. This is fostered by digital possibilities of producing and linking unprecedented masses of digitized data and the advancement of automated assessment technologies. The advent of data-driven research analytics catalysed through material specificities of digital infrastructure therefore holds a different approach from extracting meaning from data than does indicator-based research evaluation. It claims not only to extrapolate the future from past performance but moreover to genuinely discover novel topics, trends, and future achievements.

This is not only empirically relevant for understanding recent developments in the (e)valuation of academic performance. The data-drivenness of digital infrastructure also opens up new avenues for theory development in research on quantification. Quantification is so far understood as the production and communication of numbers that turn qualitative characteristics into quantities based on predefined metrics and indicators (Espeland and Stevens 2008; Mennicken and Espeland 2019). Quantification in this regard follows from operationalizing qualitative differences in terms of quantitative output according to a common metric. Yet, the data-drivenness of bibliometric infrastructure appears to work the other way round. Instead of constructing a priori a quantitative indicator for qualitative characteristics and performances, it is the massive production of digitized data and new assessment technologies from which follows how measurement and evaluation can be done (Krüger 2020). Thus, data-driven analytics do not systematically collect data based on operationalized indicators. Instead, it is the availability of large amounts of interlinked digital data subjected to algorithmic analysis that have the generative power to create new understanding of academic performance and scientific practice as such.

Yet, how these data are used is neither set nor predetermined through technology alone. It not only depends on their users (McCoy and Rosenbaum 2019; Lim 2021), but also on their providers. While we have referred to non-commercial datasets and tools such as ORCID or *Publish or Perish*, a substantial part of bibliometric infrastructure in use is owned by big private companies such as Clarivate, Elsevier, and Digital Science. Each of their product portfolios includes a current research information system and evaluative software tools that draw on their own respective citation databases. Bibliometric infrastructure has thus become a commercial product through creating – as Mirowski (2018) has put it – an encompassing "Panopticon of Science" that allows for "near real-time surveillance of the research process" (Mirowski 2018: 195).

Which part of the research process is subject to research analytics and under which premises is however still contingent. It depends on the potential "use cases" that their commercial providers advertise to win different kinds of customers even beyond researchers and research administration such as funders or publishers. Providers of research analytics follow the "institutional data imperative" (Fourcade and Healy 2017: 9) of modern organizations. They collect as much data as possible without any specific use in mind (see also Sadowski 2019). Consequently, building on the idea of "assetization" (Birch and Muniesa 2020) data on scientific practice have become an asset because they can constantly be repurposed for various uses depending on "data activation regimes" (Beauvisage and Mellet 2020: 77) or the "techcraft" (Birch et al. 2021: 2) that provide data with meaning and thus with economic value.

For our case of bibliometric infrastructure this implies that the assetization of data on scientific practices becomes possible through the material specificities of bibliometric infrastructure. Its interoperability and scalability allow for an increase in the amount of data – no matter if the data are provided by commercially operating enterprises such as Elsevier, Clarivate, or Digital Science or freely produced by non-profit organizations and later included in commercial products. The flexibility of these data allows for them to be put to use in various ways and contexts depending on how the functionalities of the evaluative software draw meaning from them. It thus appears to be the economic valuation of data on scientific practices – either as revenues for commercial providers or as return on investment for research management, policy agents, and funding agencies – that drives the extension from indicator-based research evaluation towards data-driven research analytics.

Acknowledgment: We express our sincere gratitude to the two anonymous reviewers and the editors for their very helpful comments. We also warmly thank Christian Reimer for creating the figures for this study.

References

Alaimo, Cristina, and Jannis Kallinikos. 2021. "Managing by Data: Algorithmic Categories and Organizing." *Organization Studies* 42(9): 1385–1407.

Aspesi, Claudio, and Amy Brand. 2020 "In Pursuit of Open Science, Open Access is not Enough." *Science* 368(6491): 574–577.

Altmetric. n.d. a. "The Donut and Altmetric Attention Score." https://www.altmetric.com/about-our-data/the-donut-and-score/, accessed December 22, 2021.

Altmetric. n.d. b. "Altmetric Details Page API Reference & Getting Started." http://api.altmetric.com/, accessed December 22, 2021.

Amoore, Louise, and Volha Piotukh. 2015. "Life beyond Big Data: Governing with Little Analytics." *Economy and Society* 44(3): 341–366.

Aspesi, Claudio, Nicole S. Allen, Raym Crow, Shawn Daugherty, Heather Joseph, Joseph T. W. McArthur, and Nick Shockey. 2019. "SPARC Landscape Analysis", March 29.

Aström, Fredrik. 2016. "Infrastructures as Analytical Framework for Mapping Research Evaluation Landscapes and Practices," Paper presented at Materiality and Social Dynamics of Information Infrastructures and Learning, Hindås, Sweden, October.

Baas, Jeroen, Michiel Schotten, Andrew Plume, Grégoire Côté, and Reza Karimi. 2020. "Scopus as a Curated, High-Quality Bibliometric Data

Source for Academic Research in Quantitative Science Studies." *Quantitative Science Studies* 1(1): 377–386.

Beauvisage, Thomas, and Kevin Mellet. 2020. "Datassets: Assetizing and Martketizing Personal Data." In *Assetization: Turning Things into Assets in Technoscientific Capitalism. Inside Technology*, edited by Kean Birch and Fabian Muniesa, 75–95. Inside technology. Cambridge, MA: The MIT Press.

Birch, Kean, D. T. Cochrane, and Callum Ward. 2021. "Data as Asset? The Measurement, Governance, and Valuation of Digital Personal Data by Big Tech." *Big Data & Society* 8(1).

Birch, Kean, and Fabian Muniesa (Eds.) 2020. *Assetization: Turning Things into Assets in Technoscientific Capitalism. Inside Technology.* Cambridge, MA: The MIT Press.

Birkle, Caroline, David A. Pendlebury, Joshua Schnell, and Jonathan Adams. 2020. "Web of Science as a Data Source for Research on Scientific and Scholarly Activity." *Quantitative Science Studies* 1(1): 363–376.

Björneborn, Lennart, and Peter Ingwersen. 2001. "Perspective of Webometrics." *Scientometrics* 50(1): 65–82.

Bornmann, Lutz. 2020. "Bibliometrics-Based Decision Trees (BBDTs) Based on Bibliometrics-Based Heuristics (BBHs): Visualized Guidelines for the Use of Bibliometrics in Research Evaluation." *Quantitative Science Studies* 1(1): 171–182.

Bowker, Geoffrey C., and Susan Leigh Star. 1999. S*orting Things Out: Classification and Its Consequences.* Cambridge, MA: The MIT Press.

Bowker, Geoffrey C., Julia Elyachar, Andrea Mennicken, Peter Miller, and Joanne Randa Nucho. 2019. "Introduction to Thinking Infrastructures." In *Thinking Infrastructures*, edited by Martin Kornberger, Geoffrey C. Bowker, Julia Elyachar, Andrea Mennicken, Peter Miller, Joanne Randa Nucho, and Neil Pollock. Special issue, *Research in the Sociology of Organizations* (62): 1–13.

boyd, danah, and Kate Crawford. 2012. "Critical Questions for Big Data." *Information, Communication & Society* 15(5): 662–679.

Büchner, Stefanie. 2018. "Digitale Infrastrukturen – Spezifik, Relationalität und die Paradoxien von Wandel und Kontrolle." *Arbeits- und Industriesoziologische Studien* 11(2): 279–293.

Butler, Linda. 2003. "Explaining Australia's Increased Share of ISI Publications – The Effects of a Funding Formula Based on Publication Counts." *Research Policy* 32(1): 143–155.

Clarivate Analytics. n.d. "InCites Help. About the Indicators Handbook." http://help.prod-incites.com/inCites2Live/indicatorsGroup/aboutHandbook.html, accessed December 22, 2021.

Clarivate Analytics. 2019a. "Datasheet Incites." https://clarivate.com/webofsciencegroup/wp-content/uploads/sites/2/2019/09/My_Org-Fact_Sheet-Web-of-Science-Group-2019.pdf, accessed December 22, 2021.

Clarivate Analytics. 2019b. “The Web of Science Group Launches New Publisher Analytics Reports.” https://clarivate.com/webofsciencegroup/news/the-web-of-science-group-launches-new-publisher-analytics-reports/, accessed May 15, 2022.

Cronin, Blaise. 2005. *The Hand of Science: Academic Writing and Its Rewards*. Lanham, MD: Scarecrow Press.

Crossref. 2020a. “Event Data.” https://www.crossref.org/services/event-data/, accessed December 22, 2021.

Crossref. 2020b. “History.” https://www.crossref.org/about/history/, accessed December 22, 2021.

Crossref. 2020c. “Funder Registry.” https://www.crossref.org/services/funder-registry/, accessed December 22, 2021.

Dappert, Angela, Adam Farquhar, Rachael Kotarski, and Kirstie Hewlett. 2017. “Connecting the Persistent Identifier Ecosystem: Building the Technical and Human Infrastructure for Open Research.” *Data Science Journal 16*.

de Rijcke, Sarah, and Alexander Rushforth. 2015. “To Intervene or Not to Intervene; Is That the Question? On the Role of Scientometrics in Research Evaluation.” *Journal of the Association for Information Science and Technology* 66(9): 1954–1958.

de Rijcke, Sarah, Paul F. Wouters, Alex D. Rushforth, Thomas P. Franssen, and Björn Hammarfelt. 2016. "Evaluation Practices and Effects of Indicator Use – a Literature Review." *Research Evaluation* 25(2): 161–169.

Debackere, Koenraad, Wolfgang Glänzel, and Bart Thijs. 2019. “Scientometrics Shaping Science Policy and Vice Versa, the ECOOM Case.” In *Springer Handbook of Science and Technology Indicators*, edited by Wolfgang Gänzel, Henk F. Moed, Ulrich Schmoch, and Mike Thelwall, 447–464. Cham: Springer International.

Delfanti, Alessandro. 2021. “The Financial Market of Ideas: A Theory of Academic Social Media.” *Social Studies of Science* 51(2): 259–276.

Dellavalle, Robert, Eric Hester, Lauren Heilig, Amanda Drake, Jeff Kuntzman, Marla Graber, and Lisa Schilling. 2003. “Going, Going, Gone: Lost Internet References.” *Science* 302(5646): 787–788.

Derrick, G.E., and V. Pavone. 2013. “Democratising Research Evaluation: Achieving Greater Public Engagement with Bibliometrics-Informed Peer Review.” *Science and Public Policy* 40(5): 563–575.

Dimensions. 2021a. “Linked Research Data from Idea to Impact.” https://www.dimensions.ai/, accessed December 22, 2021.

Dimensions. 2021b. “Dimensions Analytics.” https://www.dimensions.ai/products/dimensions-analytics/, accessed December 22, 2021.

Dimensions. 2021c. “Dimensions Profiles.” https://www.dimensions.ai/products/dimensions-profiles/, accessed December 22, 2021.

Elsevier. 2021a. “SciVal.” https://www.elsevier.com/solutions/scival, accessed December 22, 2021.

Elsevier. 2021b. "Research Metrics Guidebook." https://www.elsevier.com/research-intelligence/resource-library/research-metrics-guidebook, accessed December 22, 2021.

Elsevier. 2021c."See What SciVal Can Do For You." https://www.elsevier.com/solutions/scival/features, accessed December 22, 2021.

Espeland, Wendy Nelson, and Mitchell L. Stevens. 2008. "A Sociology of Quantification." *Archives Europeennes de Sociologie* 49(3): 401–436.

EurekAlert. 2011. "Elsevier Launches SciVal Strata to Enhance Research Evaluation." https://www.eurekalert.org/news-releases/888677, accessed December 22, 2021.

Fenner, Martin, and Amir Aryani. n.d. "Introducing the PID Graph." https://www.project-freya.eu/en/blogs/blogs/the-pid-graph, accessed December 22, 2021.

Ferguson, Christine, Jo McEntrye, Vasily Bunakov, Simon Lambert, Stephanie van der Sandt, Rachael Kotarski, Sarah Stewart, Andrew MacEwan, Martin Fenner, Patricia Cruse, René van Horik, Tina Dohna, Ketil Koop-Jacobsen, Uwe Schindler, and Siobhan McCafferty. 2019. "D3.1 Survey of Current PID Services Landscape." *FREYA Report* (rev. version).

Fondermann, Philipp, and Peter L. van der Togt. 2017. "How Wageningen University and Research Centre Managed to Influence Researchers' Publishing Behaviour Towards More Quality, Impact and Visibility." *Procedia Computer Science* 106: 204–211.

Fourcade, Marion, and Kieran Healy. 2017. "Seeing Like a Market." *Socio-Economic Review* 15(1): 9–29.

Franzen, Martina. 2015. "Der Impact Faktor war gestern: Altmetrics und die Zukunft der Wissenschaft." *Soziale Welt* 66(2): 225–242.

Furner, Jonathan. 2014. "The Ethics of Evaluative Bibliometrics." In *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact,* edited by Blaise Cronin and Cassidy R. Sugimoto, 85–107. Cambridge, MA: The MIT Press.

Garfield, Eugene. 1964. "'Science Citation Index' – a New Dimension in Indexing." *Science* 144(3619): 649–654.

Gingras, Yves. 2016. *Bibliometrics and Research Evaluation. History and Foundations of Information Science.* Cambridge, MA: The MIT Press.

Godin, Benoît. 2006. "On the Origins of Bibliometrics." *Scientometrics* 68(1): 109–133.

Goldenfein, Jake, Sebastian Benthall, Daniel Griffin, and Eran Toch. 2019. "Private Companies and Scholarly Infrastructure – Google Scholar and Academic Autonomy." *Critical Reflections,* Cornell Tech.

GRID (Global Research Identifier Database). 2021a. "Policies." https://grid.ac/pages/policies, accessed December 22, 2021.

GRID (Global Research Identifier Database). 2021b. "Disambiguator." https://grid.ac/disambiguate, accessed December 22, 2021.

Hammarfelt, Björn, Gustaf Nelhans, Pieta Eklund, and Fredrik Åström. 2016. "The Heterogeneous Landscape of Bibliometric Indicators: Evaluating

Models for Allocating Resources at Swedish Universities." *Research Evaluation* 25(3): 292–305.

Harzing.com. 2016a. "Publish or Perish." http://www.harzing.com/pop.htm, accessed December 22, 2021.

Harzing.com. 2016b. "Data Sources." https://harzing.com/resources/publish-or-perish/manual/using/data-sources, accessed December 22, 2021.

Harzing.com. 2021. "The Changing Usage of Publish or Perish Over the Years: Where, Why, When, What & Who?" https://harzing.com/blog/2021/01/the-changing-usage-of-publish-or-perish-over-the-years-where-why-when-what-who, accessed December 22, 2021.

Haustein, Stefanie. 2016. "Grand Challenges in Altmetrics: Heterogeneity, Data Quality and Dependencies." *Scientometrics* 108(1): 413–423.

Haustein, Stefanie, Cassidy Sugimoto, and Vincent Larivière. 2015. "Guest Editorial: Social Media in Scholarly Communication." *Aslib Journal of Information Management* 67(3).

Haustein, Stefanie, Timothy D. Bowman, and Rodrigo Costas. 2016. "Interpreting 'Altmetrics': Viewing Acts on Social Media through the Lens of Citation and Social Theories." In *Theories of Informetrics and Scholarly Communication: A Festschrift in Honor of Blaise Cronin*, edited by Cassidy R. Sugimoto, 372–406. Berlin: De Gruyter Mouton.

Hazelkorn, Ellen. 2011. *Rankings and the Reshaping of Higher Education: The Battle for World-Class Excellence*. Houndmills, Basingstoke: Palgrave Macmillan.

Herb, Ulrich. 2019. "Überwachungskapitalismus und Wissenschaftssteuerung." https://www.heise.de/tp/features/Ueberwachungskapitalismus-und-Wissenschaftssteuerung-4480357.html?seite=all, accessed December 22, 2021.

Herzog, Christian, Daniel Hook, and Stacy Konkiel. 2020. "Dimensions: Bringing down Barriers between Scientometricians and Data." *Quantitative Science Studies* 1(1): 387–395.

Hicks, Diana. 2012. "Performance-Based University Research Funding Systems." *Research Policy* 41(2): 251–261.

Hug, Sven E., Michael Ochsner, and Martin P. Brändle. 2017. "Citation Analysis with Microsoft Academic." *Scientometrics* 111(1): 371–378.

InCites. 2021. "AI-powered Insights for Digital Business." https://www.incites.com/, accessed December 22, 2021.

Information Today. 2013. "Thomson Reuters Enhances Research and Analytics Tools." http://newsbreaks.infotoday.com/Digest/Thomson-Reuters-Enhances-Research-and-Analytics-Tools-93825.asp, accessed December 22, 2021.

International DOI Foundation. 2015. "The DOI Handbook." https://www.doi.org/doi_handbook/1_Introduction.html, accessed December 22, 2021.

International DOI Foundation. 2021. "Factsheet. Key Facts on Digital Object Identifier System." https://www.doi.org/factsheets/DOIKeyFacts.html, accessed December 22, 2021.

ISI Web of Knowledge. 2009. "Series of Webinars – Demonstration of InCites." https://wok.mimas.ac.uk/news/2009/20091007.html, accessed December 22, 2021.

Jappe, Arlette. 2020. "Professional Standards in Bibliometric Research Evaluation? A Meta-Evaluation of European Assessment Practice 2005–2019." *PloS one* 15(4): e0231735.

Kaltenbrunner, Wolfgang, and Sarah de Rijcke. 2017. "Quantifying 'Output' for Evaluation: Administrative Knowledge Politics and Changing Epistemic Cultures in Dutch Law Faculties." *Science and Public Policy* 44(2): 284–293.

Katz, J. Sylvan, and Diana Hicks. 1997. "Desktop Scientometrics." *Scientometrics* 38(1): 141–153.

Kiviat, Barbara. 2019. "The Moral Limits of Predictive Practices: The Case of Credit-Based Insurance Scores." *American Sociological Review* 84(6): 1134–1158.

Klump, Jens, Fiona Murphy, Tobias Weigel, and Mark Parsons. 2017. "Editorial: 20 Years of Persistent Identifiers – Applications and Future Directions." *Data Science Journal* 16.

Krüger, Anne K. 2020. "Quantification 2.0? Bibliometric Infrastructures in Academic Evaluation." *Politics and Governance* 8(2): 58–67.

Lagotto. n.d. a. "About Lagotto." https://www.lagotto.io/, accessed December 22, 2021.

Lagotto. n.d. b. "Public Library of Science (PLOS)." https://www.lagotto.io/plos/, accessed December 22, 2021.

Leydesdorff, Loet, Paul Wouters, and Lutz Bornmann. 2016. "Professional and Citizen Bibliometrics: Complementarities and Ambivalences in the Development and Use of Indicators—a State-of-the-Art Report". *Scientometrics* 109(3): 2129–2150.

Lim, Miguel Antonio. 2021. "Governing Higher Education: The PURE Data System and the Management of the Bibliometric Self." *Higher Education Policy* 34(1): 238–253.

Lin, Jennifer, and Martin Fenner. 2013. "Altmetrics in Evolution: Defining and Redefining the Ontology of Article-Level Metrics." I*nformation Standards Quarterly* 25(2): 20–26.

López-Cózar, Emilio Delgado, Enrique Orduna-Malea, and Alberto Martín-Martín. 2019. "Google Scholar as a Data Source for Research Assessment." In *Springer Handbook of Science and Technology Indicators*, edited by Wolfgang Gänzel, Henk F. Moed, Ulrich Schmoch, and Mike Thelwall, 95–127. Cham: Springer International.

Lupton, Deborah. 2016. *The Quantified Self: A Sociology of Self-Tracking*. Cambridge: Polity Press.

McCoy, Chase, and Howard Rosenbaum. 2019. "Uncovering Unintended and Shadow Practices of Users of Decision Support System Dashboards in Higher Education Institutions." *Journal of the Association for Information Science and Technology* 70(4): 370–384.

MacKenzie, Donald, and Judy Wajcman. 1999. "Introductory Essay: The Social Shaping of Technology." In *The Social Shaping of Technology*, edited by Donald A. MacKenzie and Judy Wajcman. Maidenhead: Open University Press. http://eprints.lse.ac.uk/28638/, accessed December 22, 2022.

Martín-Martín, Alberto, Mike Thelwall, Enrique Orduña-Malea, and Emilio Delgado López-Cózar. 2021. "Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A Multidisciplinary Comparison of Coverage via Citations." *Scientometrics* 126(1): 871–906.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: Murray.

Meadows, Alice, Laurel L. Haak, and Josh Brown. 2019. "Persistent Identifiers: The Building Blocks of the Research Information Infrastructure." *Insights* 32(1).

Meadows, Alice, Helena Cousijn, Maria Gould, Ginny Hendricks, Julie Petro, and Natasha Simons. 2021. "PIDs 101: A Beginners' Guide to Persistent Identifiers." Presentation.

Mennicken, Andrea, and Wendy Nelson Espeland. 2019. "What's New with Numbers? Sociological Approaches to the Study of Quantification." *Annual Review of Sociology* 45(1): 223–245.

Mennicken, Andrea, and Martin Kornberger. 2021. "From Performativity to Generativity: Valuation and Its Consequences in the Context of Digitization." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 73: 451–478.

Microsoft. 2021a. Next Steps for Microsoft Academic – Expanding into New Horizons." https://www.microsoft.com/en-us/research/project/academic/articles/microsoft-academic-to-expand-horizons-with-community-driven-approach/, accessed December 22, 2021.

Microsoft. 2021b. "Microsoft Academic." https://academic.microsoft.com/home, accessed December 22, 2021.

Mirowski, Philip. 2018. "The Future(s) of Open Science." *Social Studies of Science* 48(2): 171–203.

Moed, Henk F. 2007. "The Future of Research Evaluation Rests with an Intelligent Combination of Advanced Metrics and Transparent Peer Review". *Science and Public Policy* 34(8): 575–583.

Moed, Henk. F. 2017. *Applied Evaluative Informetrics. Qualitative and Quantitative Analysis of Scientific and Scholarly Communication*. Cham: Springer International.

Narin, Francis. 1976. *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity.* Report to the National Science Foundation. Cherry Hill, NJ: Computer Horizons.

Nicholas, David, Eti Herman, Hamid R. Jamali, Abdullah Abrizah, Cherifa Boukacem-Zeghmouri, Jie Xu, Blanca Rodríguez-Bravo, Anthony Watkinson, Tatiana Polezhaeva, and Marzena Świgon. 2020. "Millennial

Researchers in a Metric-Driven Scholarly World: An International Study." *Research Evaluation* 29(3): 263–274.

Nuredini, Kaltrina, Steffen Lemke, and Isabella Peters. 2021. "4.3 Social Media and Altmetrics." In *Handbook Bibliometrics*, edited by Rafael Ball, 201–214. Berlin: De Gruyter Saur.

OpenAlex. 2022. "About the Data." https://docs.openalex.org/about-the-data-1, accessed May 15, 2022.

Orduña-Malea, Enrique, Alberto Martín-Martín, Juan M. Ayllon, and Emilio Delgado López-Cózar. 2014. "The Silent Fading of an Academic Search Engine: The Case of Microsoft Academic Search." *Online Information Review* 38(7): 936–953.

Orlikowski, Wanda J., and Susan V. Scott. 2008. "Sociomateriality: Challenging the Separation of Technology, Work and Organization." The Academy of Management Annals 2(1): 433–474.

Ortega, José-Luis. 2020. "Altmetrics Data Providers: A Meta-Analysis Review of the Coverage of Metrics and Publication." *El Profesional de la Información* 29(1): 1–23.

Overton. n.d.. "Is Your Work Influencing Policy?" https://www.overton.io/, accessed December 22, 2021.

Paskin, Norman. 2010. "Digital Object Identifier (DOI®) System." *In Encyclopedia of Library and Information Sciences,* 3rd edn, edited by Marcia J. Bates and Mary N. Maack, 1586–5392. Boca Raton, FL: CRC Press.

Petersohn, Sabrina. 2016. "Professional Competencies and Jurisdictional Claims in Evaluative Bibliometrics: The Educational Mandate of Academic Librarians", *Education for Information* 32(2): 165–193.

Petersohn, Sabrina, and Thomas Heinze. 2018. "Professionalization of Bibliometric Research Assessment. Insights from the History of the Leiden Centre for Science and Technology Studies (CWTS)." *Science and Public Policy* 45(4): 565–578.

Piromalli, Leonardo. 2019. "Research Information Management in Italy. The IRIS Platformised Infrastructure as a Sociotechnical Device." S*cuola democratica* 4: 297–319.

Pollock, Neil, and Robin Williams. 2007. "Technology Choice and Its Performance: Towards a Sociology of Software Package Procurement." *Information and Organization* 17(3): 131–161.

Porter, Theodore. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.

Preda, Alex. 2006. "Socio-Technical Agency in Financial Markets." *Social Studies of Science* 36(5): 753–782.

Priem, Jason. 2014. "Altmetrics." *In Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact,* edited by Blaise Cronin and Cassidy R. Sugimoto, 263–287. Cambridge, MA: The MIT Press.

Priem, Jason, and Heather Piwowar. 2012. "The Launch of ImpactStory: Using Altmetrics to Tell Data-Driven Stories." https://blogs.lse.ac.uk/

impactofsocialsciences/2012/09/25/the-launch-of-impactstor/, accessed December 22, 2021.

Priem, Jason, Dario Taraborelli, Paul Groth, and Cameron Neylon. 2010. "Altmetrics: A Manifesto." http://altmetrics.org/manifesto, accessed December 22, 2021.

Pritchard, Alan. 1969. "A Statistical Bibliography or Bibliometrics." *Journal of Documentation* 25(4): 348–349.

Relx. 2009. "Elsevier Launches SciVal Spotlight." https://www.relx.com/media/press-releases/archive/29-06-2009, accessed December 22, 2021.

Relx. 2012. "Elsevir Acquires Atira, a Provider of Research Management Solutions. https://www.relx.com/media/press-releases/archive/15-08-2012, accessed December 22, 2021.

Research Graph Foundation. n.d. "Research Graph Schema." https://researchgraph.org/, accessed December 22, 2021.

Research Information. 2015. "'A Taste for Openness and Change'." https://www.researchinformation.info/issue/octobernovember-2015, accessed December 22, 2021.

Ringgold Inc. 2021. "Ringgold Identifier." https://www.ringgold.com/ringgold-identifier/, accessed December 22, 2021.

ROR. n.d. a. "About." https://ror.org/scope/, accessed December 22, 2021.

ROR. n.d. b. "Scope." https://ror.org/scope/, accessed December 22, 2021.

Sadowski, Jathan. 2019. "When Data Is Capital: Datafication, Accumulation, and Extraction." *Big Data & Society* 6(1).

The Scholarly Kitchen. 2014. "Interview with Thomson Reuters: InCites Platform Offers New Analytics and Transparency." https://scholarlykitchen.sspnet.org/2014/08/08/qa-with-thomson-reuters-incites-platform-offers-new-analytics-and-transparency/, accessed December 22, 2021.

Scopus. 2019. "PlumX Metrics API Now Available for Scopus Subscribers." online: https://blog.scopus.com/topics/plumx-metrics#:~:text=PlumX%20Metrics%20is%20Plum%20Analytics%E2%80%99%20comprehensive%2C%20item-level%20metrics,chapters%2C%20and%20many%20more%29%20in%20the%20online%20environment, accessed December 22, 2021.

Sivertsen, Gunnar. 2019. "Developing Current Research Information Systems (CRIS) As Data Sources for Studies of Research." In *Springer Handbook of Science and Technology Indicators*, edited by Wolfgang Glänzel, Henk F. Moed, Ulrich Schmoch, and Mike Thelwall, 667–683. Cham: Springer International.

Star, Susan Leigh, and Karen Ruhleder. 1996. "Steps Toward an Ecology of Infrastructure. Design and Access for Large Information Spaces." *Information Systems Research* 7(1): 111–134.

Star, Susan Leigh, and Anselm Strauss. 1999. "Layers of Silence, Arenas of Voice. The Ecology of Visible and Invisible Work." *Computer Supported Cooperative Work* 8(1–2): 9–30.

Tattersall, Andy and Christopher Carroll. 2018. "What Can Altmetric.com Tell Us about Policy Citations of Research? An Analysis of Altmetric.com Data for Research Articles from the University of Sheffield." *Frontiers in Research Metrics and Analytics* 2

Teixeira da Silva, Jaime A., and Aamir Raoof Memon. 2017. "CiteScore: A Cite for Sore Eyes, or a Valuable, Transparent Metric?" *Scientometrics* 111(1): 553–556.

Thelwall, Mike, Liwen Vaughan, and Lennart Björneborn. 2006. "Webometrics." *Annual Review of Information Science and Technology* 39(1): 81–135.

Tilson, David, Kalle Lyytinen, and Carsten Sørensen. 2010. "Research Commentary —Digital Infrastructures: The Missing IS Research Agenda." *Information Systems Research* 21(4): 748–759.

Unternährer, Markus. 2021. "The Order of Recommendations." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 73: 397–423.

van Raan, Anthony. 2019. "Measuring Science: Basic Principles and Application of Advanced Bibliometrics." In *Springer Handbook of Science and Technology Indicators,* edited by Wolfgang Glänzel, Henk F. Moed, Ulrich Schmoch, and Mike Thelwall, 237–280. Cham: Springer International.

Wagner, Erica L., Jodie Moll, and Sue Newell. 2011. "Accounting Logics, Reconfiguration of ERP Systems and the Emergence of New Accounting Practices: A Sociomaterial Perspective." *Management Accounting Research* 22(3): 181–197.

Wang, Kuansan, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. "Microsoft Academic Graph: When Experts Are Not Enough." *Quantitative Science Studies* 1(1): 396–413.

Wikipedia. 2021a. "Altmetric." https://en.wikipedia.org/wiki/Altmetric, accessed December 22, 2021.

Wikipedia. 2021b. "Plum Analytics." https://en.wikipedia.org/wiki/Plum_Analytics, accessed December 22, 2021.

Wikipedia. 2021c. "ORCID." https://en.wikipedia.org/wiki/ORCID#cite_ref-8, accessed December 22, 2021.

Wouters, Paul. 1999. "The Citation Culture." PhD diss. Universiteit van Amsterdam. http://garfield.library.upenn.edu/wouters/wouters.pdf.

Zahedi, Zohreh, and Rodrigo Costas. 2018. "General Discussion of Data Quality Challenges in Social Media Metrics: Extensive Comparison of Four Major Altmetric Data Aggregators." *PloS one* 13(5).

Zhang, Yi, Alan L. Porter, Scott Cunningham, Denise Chiavetta, and Nils Newman. 2020. "Parallel or Intersecting Lines? Intelligent Bibliometrics for Investigating the Involvement of Data Science in Policy Analysis." *IEEE Transactions on Engineering Management.*

Anne K. Krüger is a sociologist working at the Berlin-Brandenburg Academy of Sciences and Humanities. She earned her doctoral degree from Humboldt University Berlin and has worked as a postdoc at the universities of Göttingen, Potsdam and HU Berlin, and at the German Centre for Higher Education Research and Science Studies. Her research focuses on the relationship between digitization, organization and (e)valuation. In her current project "New Data, New Knowledge?", she investigates how digital infrastructure affects bibliometric performance measurement and the understanding of research practice. Most recently, she has published an introduction to the sociology of valuation and evaluation [*Soziologie des Wertens and Bewertens*] with utb/transcript publishers.

Sabrina Petersohn has been a postdoc researcher at the German Centre for Higher Education Research and Science Studies since 2017. A trained sociologist, she works at the intersection of science studies and science policy by conducting research into competencies and task profiles in IT-supported research reporting, digital infrastructure in research assessment, research governance, responsible metrics education and the professionalization of evaluative bibliometrics. She holds a PhD in sociology from the University of Wuppertal and a Master's degree in sociology, political science and public law from the University of Potsdam. Previously she worked as a doctoral researcher at the Department of Sociology of Organizations at the University of Wuppertal (2016–17) and at the Computational Social Science Department of GESIS Leibniz Institute of Social Sciences in Cologne (2012–16).

Valuation Studies 9(1) 2022: 47-77

Theme issue contribution

What is Digital Valuation Made of? The integration of valuation poles on a reservation platform and its effects on the hotel industry in Switzerland

Philip Balsiger and Thomas Jammet

Abstract

Digital platforms act as new powerful intermediaries challenging existing market orders in many sectors. Algorithmically producing ratings and rankings often built from online consumer reviews, platforms are important players in the digitizing of valuation. This article asks how these new platform-generated valuations relate to other forms of valuation. It presents a qualitative case study of valuation in the hotel sector in Switzerland, drawing on interviews with professionals and a description of valuation categories on the Booking.com website. Going beyond the description of the opposition between online consumer reviews and traditional judgment devices, the analysis shows that valuation on the platform is based upon a permissive hierarchical integration of a plurality of valuation poles, with algorithmic valuation at its center. This destabilizes the evaluative landscape with regard to three issues: lack of transparency of the algorithmic ranking; weakening and even undermining of formulaic valuation; and the issue of singularization of the online offer.

Keywords: algorithms; Booking.com; digital platform; markets; online consumer reviews; ranking

Philip Balsiger is professor of economic sociology at the University of Neuchâtel. Thomas Jammet is scientific assistant at the School of Social Work Fribourg (HES-SO) and research associate at the University of Neuchâtel.

© *2022 The authors* *This work is licensed under a Creative Commons Attribution 4.0 International License.*
https://doi.org/10.3384/VS.2001-5992.2022.9.1.47-77
Hosted by Linköping University Electronic press
http://valuationstudies.liu.se

Introduction

In many economic sectors, digital platforms act as new powerful intermediaries and challenge existing market orders (Kenney and Zysman 2016; Kirchner and Beyer 2016; Srnicek 2017; Kirchner and Schüssler 2019). One important process through which they remake markets are via new forms of digital valuation such as ratings, rankings and online consumer reviews (OCRs) (Mellet et al. 2014; Stark and Pais 2020). Online valuation devices rely on an algorithmic logic, based on the transformation of user-generated content into a "trusted recommendation" developed by the platforms themselves. They add themselves on top of other, already existing "judgement devices" (Karpik 2010) on markets, such as those based on categories developed by critics, professional associations, experts or marketing (Beckert and Aspers 2011; Orlikowski and Scott 2014). How do the new platform-generated valuations relate to these other forms of valuation? This article empirically analyses these relations using a case study of the hotel industry, where OCRs and rankings produced by reservation platforms such as Booking.com challenge valuation devices that have long structured the market.

Several studies have analyzed the typical forms of valuation produced by platforms – OCRs, and the ratings and rankings based on them resulting from algorithmic treatment processes. OCRs typically take the standardized form of a rating and a written review (Beauvisage et al. 2013) and hold the promise of democratizing product and service evaluation by valuing everyday consumers' points of view, rather than that of 'experts'; they also cover much broader ground than guide books (Mellet et al. 2014). Studies on OCRs in the restaurant and hotel sectors have shown how wide adoption of these has had strong effects on professionals, who are forced to be reactive to this increasingly dominant form of valuation (Beuscart et al. 2016; Cardon 2014; Kim and Velthuis 2021). There are also indications that this "algorithmic" apparatus increasingly competes with and possibly replaces "formulaic" apparatuses of valuation such as those controlled by experts (Orlikowski and Scott 2014).[1]

Kornberger et al. (2017) propose to conceptualize platforms as "evaluative infrastructures". Their analysis demonstrates how platforms tend to integrate a plurality of valuation devices and put in place "distributed" valuation processes (2017: 85), albeit with a "hidden cursor" (2017: 89), since platforms seek to maximize revenue. We bring together the perspective on evaluative infrastructure and

---

[1] A formulaic apparatus of valuation is one in which "we see a formula at work" (Orlikowski and Scott 2014: 883). In the hotel sector, the formula consists of "a standards-centered model for what constitutes hotel accommodation, enacting both a method of hotel evaluation and a plan for hotel improvement" (Orlikowski and Scott 2014: 883).

discussion on the competition and interplay between algorithmic and "traditional" forms of valuation to advance the study of digital valuation on platforms. Our analysis of a reservation platform in the hotel sector shows how the platform combines different valuation forms, which we conceptualize as belonging to different valuation "poles". The spatial metaphor of poles allows us to distinguish ensembles of devices responding to different principles and driven by different actor types, which occupy distinct positions in the evaluative space. Going beyond the description of the opposition between online consumer reviews and traditional judgment devices, the analysis therefore suggests that the evaluative innovation of platforms consists of their integration into a plurality of valuation poles. This integration is hierarchical and permissive, with algorithmic valuation at its center. We use in-depth interviews with hotel owners and managers to show how this destabilizes the evaluative landscape with regard to three issues: the lack of transparency in the algorithmic ranking; the weakening and even undermining of formulaic valuation (Orlikowski and Scott 2014); and the singularization of the online offer. Although the domination of algorithmic valuation tends to weaken the formulaic and commercial valuation poles, the plasticity of online evaluative infrastructures also presents opportunities for hotel owners and managers.

The article uses a qualitative case study on the Swiss hotel sector and the reservation platform Booking.com. The hotel industry was one of the first industries to become "platformized", with the rise in the early 2000s of so-called online travel agencies (OTAs), and of specialized review websites such as Yelp or TripAdvisor. We conducted interviews with hotel managers, representatives of professional associations and sectoral organizations and other field actors to study the workings of valuation in this market. Adopting an inductive perspective on the evaluative infrastructure, we were attentive in particular to the perspectives of members of the hotel profession in order to understand how digital valuation shapes this economic sector. The interview data was complemented by a close description of the categories of valuation on Booking.com.

We start by discussing the literature and the main concepts and distinctions that will be used in this article's analysis. This is followed by a presentation of the case study and the methodological approach. The empirical part starts with a description and analysis of the evaluative infrastructure on Booking.com. In a second analytical part, we discuss the relations between the different valuation poles present on the platform and the effects of the dominant forms of digital valuation (lay and algorithmic valuation) on other valuation poles.

Digital valuation on platforms

Over the past decades, digital platforms have emerged in an increasing number of sectors as diverse as retailing, transport, food delivery and accommodation. They are a new form of economic organization (Kenney and Zysman 2016) and can be defined as "online sites and services that (a) host, organize, and circulate users' shared content or social interactions for them, (b) without having produced or commissioned (the bulk of) that content, (c) built on an infrastructure, beneath that circulation of information, for processing data for customer service, advertising, and profit" (Gillespie 2018: 18). Platforms either organize new markets or insert themselves into existing ones to their own benefit (Ahrne et al. 2015; Balsiger et al. 2022). Many scholars describe this phenomenon as the advent of a "platform economy" (Kenney and Zysman 2016) or even "platform capitalism" (Srnicek 2017) characterized by a few dominant players that algorithmically dictate the relationships and conditions of transaction among market participants, be they consumers, workers or organizations.

As new digital market intermediaries (Bessy and Chauvin 2013), platforms become "gatekeepers" (Lynskey 2017; Gillespie 2018) whose digital infrastructure and algorithmic tools of matchmaking have powerful effects on valuation (Kirchner and Beyer 2016; Kornberger et al. 2017). Their most characteristic evaluative devices are ratings and rankings (Stark and Pais 2020), which impact sellers on platformized markets and complement, or possibly compete with, other forms of valuation. Ratings are evaluations given by users to sellers (and sometimes vice versa, too) that offer feedback about performance. They are made visible on platforms, often as a numerical score. While platforms do not control them, they nonetheless have broad leeway in how they make them visible, how they calculate them, and also how they use them (for instance to exclude providers) (Stark and Pais 2020). They are typically one of the elements that go into the generation of rankings. Rankings are ordered lists of providers or goods, generated by algorithms and constantly updated and recalculated with the goal of favoring matches (Stark and Pais 2020). From the point of view of providers, rankings generate a hierarchy of visibility (Gillespie 2014; Fradkin 2017), with potentially important effects on sales. Broadly, two interpretations have been given to digital valuation on platforms: a "competition/replacement perspective" focuses on the effect of OCRs which increasingly compete with or even replace "traditional" valuation forms, while a "plurality perspective" presents digital valuation as plural evaluative infrastructures.

The competition/replacement perspective

A range of studies has highlighted how the characteristic forms of digital valuation – and in particular the innovation of OCRs –

increasingly compete with older forms of valuation. OCRs are a novel type of valuation device built on customer reviews of products and services that offer guidance to consumers (Jeacle and Carter 2011; Mellet et al. 2014; Beuscart et al. 2016). Taking the canonical form of a rating and a written review, OCRs combine aspects of personal judgment devices and impersonal devices (the building of a score), using neither expertise nor objectivation procedures (Mellet et al. 2014: 8). Instead, they constitute a form of democratization of judgment devices: valuing everyday consumers' opinions, they build on an egalitarian logic, as opposed to valuation devices based on the categories and points of view of critics, or professional associations and experts (Karpik 2010; Beckert and Aspers 2011; Mellet et al. 2014). OCRs offer both commensuration (through the building of an average score across reviews) and singularization (through the display of individual consumers' voices). Yet although OCRs do give voice to consumers, their calculation, categories and display are controlled by platforms.

Because of their widespread adoption, OCRs have significant effects on markets. They provoke reactivity (Espeland and Sauder 2007) from the evaluated sellers and service providers, who adapt their services and practices accordingly (Curchod et al. 2020; Kim and Velthuis 2021). Several studies observed the "overflow" (Orlikowski and Scott 2014) of online reviews on management practices in the hotel or restaurant industry (Scott and Orlikowski 2012; Cardon 2014; Orlikowski and Scott 2014; Beuscart et al. 2016; Kim and Velthuis 2021). Often, service providers will react to consumer comments, using OCRs as a form of reputation management (Beuscart et al. 2016; Wang et al. 2016; Kim and Velthuis 2021; Balsiger et al. 2022).

Besides these effects on management practices, the rise of OCRs also has effects on the overall functioning of market valuation. Through OCRs, lay judgments come to the fore, and studies have indicated that the *algorithmic apparatus* which configures these lay judgments tends to become more important than the *formulaic apparatus* of valuation often controlled by professions and "based on standards, principles, or prescriptions for achieving particular ends" (Orlikowski and Scott 2014: 883). The hypothesis here is thus that platform-generated valuation might eventually replace more traditional forms of formulaic valuation. Platforms are essentially seen as pushing a new form of valuation, based on rankings and OCRs.

The plurality perspective

More recently, a number of authors have put forward a characterization of platforms that insists more on their distributed nature and on their openness to a variety of valuation forms. Stark and Pais (2020), for instance, see platforms as an organizational form based on co-optation, meaning that as intermediaries, they tend to integrate or co-opt "the energy and creativity of actors who are on the platform" (2020: 51). Vallas and Schor (2020) speak of platforms as "permissive potentates" which "exercise power over economic transactions by delegating control among the participants" (2020: 282). Most relevant with regard to the issue of valuation, Kornberger et al. (2017) point out the plurality of valuation on platforms. They conceptualize platforms as "evaluative infrastructures", defined as an "ecology of devices that disclose values of actions, events and objects in heterarchically organized systems (such as platforms) through the maintenance of protocol[2]" (Kornberger et al. 2017: 85). Evaluative infrastructures include typical rankings and ratings, but also many other market devices and evaluation mechanisms (Kornberger et al. 2017: 85) – a plurality of "judgment devices" (Karpik 2010). By pointing at evaluative *infrastructures*, these authors maintain that the distinguishing characteristic of platforms is the "distributed" nature of valuation processes they put in place. Their illustrative case study of eBay shows three major characteristics of platforms as evaluative infrastructures. First, the plurality of evaluative devices that build an infrastructure allows for a "complex set of possibilities for making connections" (Kornberger et al. 2017: 89). The nature of valuation processes on platforms is thus distributed. Second, evaluative infrastructures importantly build on user-generated information and are generative, insofar as they "do something other than verify and validate the world as it is. Rather, they disclose the world that the digital traces and extensive data mining provide" (Kornberger et al. 2017: 89). Third, in spite of the distributed nature of valuation, platforms also exert control. While infrastructural disclosure may be endogenous, it is also influenced by what might be called the "hidden cursor" of platform organizations: the commercial imperative for platform owners to maximize revenues via traffic to their platform (Kornberger et al. 2017: 89). Plurality, thus, does not rhyme with equality.

In this article, we seek to combine these perspectives to advance the analysis of platforms' evaluative infrastructures. Going beyond an opposition between platform/algorithmic valuation and non-platform (formulaic or other) forms of valuation, we look more closely at how a specific platform *combines* different forms of valuation which we conceptualize as belonging to different valuation *poles* (commercial, lay, algorithmic, formulaic), which are in tension with each other. Adopting a perspective attentive to the plurality of valuation forms, we are explicitly interested in exploring the tensions between these poles and the power relations that explain how such tensions are resolved

---

[2] The term "protocol" designates the form of decentralized control characteristic of platforms.

within platform infrastructure. How does the new configuration of digital valuation affect forms of valuation that have traditionally shaped markets? Do platforms tend to impose one form of valuation – lay valuation such as OCRs, or forms of algorithmic valuation – or do they rather seek to integrate as many valuation forms as possible? And how do producers/service providers deal with and react to the evaluative infrastructure generated by platforms? To address these questions, we look at the hotel sector and analyze the role of so-called online travel agencies (OTAs), in particular Booking.com, and the consequences of its rise in valuation of the hotel industry. In the next section, we present this case as well as the methodology used to analyze it.

Case and method

The hotel industry is one of the sectors most affected by the phenomenon of platformization. It was "disrupted" early on by OTAs, like Booking.com or Expedia, which have expanded their activity worldwide. These platforms can be categorized as e-commerce platforms and play the characteristic role of new digital intermediaries. OTAs incorporate typical OCRs, but as reservation portals, they are more than "just" review websites like Yelp or TripAdvisor. While OTAs also develop OCRs, their goal is not only to serve as a "judgment device" (Karpik 2010) for consumers, but to generate actual economic transactions in the form of hotel bookings. Indeed, the business model of OTAs is one of commissions – for each successful transaction they take a commission in the order of 12–15% of the price. The interest of the platform is thus to maximize the number of transactions, and the evaluative infrastructure is built to this effect. As we will see, the search results produced by the platform thus incorporate many more elements than just the average ratings of consumers.

The hotel industry is well suited for an analysis of the effects of digital valuation. Platforms have become significant players for hotel reservations; at the same time, the hotel industry is characterized by strongly established valuation forms controlled by professionals such as the stars rating system; finally, a few influent studies have analyzed the effects of algorithmic valuation on hotels and the proximate restaurant industry when OCRs were a relatively new feature (Scott and Orlikowski 2012; Mellet et al. 2014; Orlikowski and Scott 2014; Beuscart et al. 2016), which constitutes an opportunity to discuss and expand this literature.

Our study looks at the hotel industry in Switzerland, a country with close to 5,000 hotels[3] and where tourism is a traditional and important sector of the economy. In terms of platforms, the focus is on Booking.com, which is by far the most important reservation platform in Switzerland, with a market share among OTAs of around 75%, representing almost 30% of all hotel reservations (Schegg 2019). Founded in 1996 and headquartered in Amsterdam, Booking.com is also one of the biggest OTAs worldwide and is today part of the publicly traded BookingHoldings, which also owns other OTAs and reservation platforms in other sectors. The great majority of hotels in Switzerland are present on Booking.com.

Data and analyses presented here are part of a broader research project on the reactions of the main players in the hotel market to the rise of platforms. In the course of this project, two researchers conducted 24 interviews with representatives of professional hotel associations and sectoral tourism organizations, hotel owners and hotel managers. All interviews were conducted between April 2019 and January 2020 and focused on the professionals' perceptions and reactions to the rise of digital platforms, in particular OTAs. We also conducted interviews with four IT service providers, a representative of a customer review aggregator company (RealReview[4]) and an official of Booking.com in Switzerland, amounting to a total of 30 interviews.

The interviews were transcribed and analyzed inductively by two members of the research team, using thematic coding with the help of the software Atlas.ti. For this article, the authors further analyzed the themes related to online consumer reviews, ratings and rankings, and qualification/valuation, in a back-and-forth process between data and theory, based on the literature on OCRs and platforms as evaluative infrastructures. In addition to this interview data, the analysis presented here also draws on an in-depth description of the evaluative categories of Booking.com and its technological affordances.

The evaluative infrastructure of Booking.com: Permissive hierarchical integration of valuation poles

Booking.com is an online travel agency: a search engine that seeks to create matches between people trying to reserve a room and hotels or other accommodation types. The search engine will provide consumers with a list of available hotels/rooms, for the entered time period and

[3] According to a report by the Swiss Tourism Federation from 2020, 4,646 hotels. https://www.stv-fst.ch/sites/default/files/2021-06/STV_STIZ_2020_EN.pdf (accessed November 19, 2021).

[4] The name of this company has been changed for confidentiality reasons.

place. The main component of the evaluative infrastructure is this search result list, which is actually a ranking. It is generated by an algorithm which will rank the hotels available on the chosen dates. We will first present the functioning of this ranking, before discussing other elements of Booking.com's evaluative infrastructure. Finally, we will discuss the different sources of the data that goes into constructing the various forms of valuation and their integration.

Ranking of search results

As is generally the case in the platform economy, the precise components and calculations that go into Booking.com's algorithm are kept secret (Gillespie 2014; Pasquale 2015; Just and Latzer 2017). However, the company does reveal that the algorithm takes into account two main aspects: elements related to the seller, and elements related to the buyer. First, the ranking is generated according to hotel "characteristics":

> We look first at the hotel's performance – conversion and cancellation rates, also the reviews, etc. That kind of defines the performance of a single property [and] where we rank this property. Because [we are] commission-based, we want to have or give to the customer the best hotel where the chance is the highest that he/she will also book and have a good stay. (Booking.com official, January 13, 2020)

The ranking on Booking.com is thus very different from the rankings produced by review websites such as TripAdvisor or Yelp. The latter explicitly give ranking to hotels (e.g., #1 hotel in Paris, #5 pizzeria in Naples) which are exclusively based on OCRs and the same for each user. Booking.com does not produce such a numbered ranking. It generates search result lists – although those results are indeed always implicitly ranked since they appear as a list. As the interviewee quoted above stated, OCR is but one element that goes into the calculation of these results; hotel performance is another important element, as is data about the user. Contrary to review websites such as TripAdvisor, which function essentially as judgment devices (at least initially – some now offer the chance to make reservations), Booking.com is a transaction-based platform and the judgment devices it makes available and uses to generate search results are there in order to favor bookings.

Since searches on Booking.com for hotels in a given location will often yield dozens or even hundreds of results, the ranking greatly determines visibility. There are possibilities for hotels to "buy" a better ranking position, which will favorably affect their position in the ranking (through boosts, special deals, or by becoming a member of the "preferred partner program"). These are specific programs offered by Booking.com. To be part of such programs, hotels pay: for instance, to participate in the preferred partner program, hotels will accept to pay a higher commission.[5] However, these programs are not open to any hotel: a hotel needs to “perform” well to be able to participate in this. For instance, a hotel that does not get many bookings will not be accepted, because it will not help the platform to increase its revenue (interview with Booking.com official). In other words, a hotel that isn’t competitive on the platform will generally be ranked low and therefore be less visible (which is likely to make the hotel even less competitive). The possibility of participating in these programs is further limited by the fact that the platform restricts the number of hotels that can be part of them, in any given location.

Second, according to the Booking.com representative interviewed, the ranking is also personalized with regard to the customer, using data that is available to the platform. For users that have a Booking.com account, this data includes their detailed search and booking history, which provides information about their habits and preferences; for users without a profile, the algorithm uses more general data that are available (such as the country where the user is searching from, search terms entered in Google, etc.). This customization of results means that the ranking is technically never the same for two different customers.

Plurality of evaluative infrastructure

So far, the description gives the impression of an evaluative environment that is strongly directed by the platform. However, the algorithmic ranking is by no means the only component of the platform’s evaluative infrastructure. There are two important additional features that need to be added to this description: first are all the various forms of differentiation or criteria of valuation/evaluation that are integrated into the platform and made visible. A look at a search result list (see Figure 1) immediately reveals their diversity. Besides the names of hotels, one indeed finds many different categories that allow users to differentiate offers. User ratings are an important component of this. There is an overall score for a number of subcategories: staff, cleanliness, location, quality/price, comfort, facilities, wi-fi. Each hotel has such ratings; and depending on the hotel (and possibly also the users), different aspects will be made visible. In addition to that, there is a great variety of other more or less objective information: stars, price, location (distance to city center or other points of interest), conditions of reservation, type of establishment,

[5] The program and its conditions are explained on Booking.com’s Partner Hub website: https://partner.booking.com/en-us/help/growing-your-business/increase-revenue/all-you-need-know-about-preferred-partner-program (accessed April 15, 2021).

type of bed, facilities such as pool, parking, spa, popularity with different types of customers, etc. Second, the platform offers its users the possibility to customize their search results according to all of these criteria. Built into the platform infrastructure are numerous possibilities to do so, the most important being (a) the possibility of changing the ranking of search results by looking at hotels by price or by customer rating, other aspects or combinations thereof; (b) the possibility of filtering the results according to specific criteria; (c) the possibility of looking at a map view, scroll over it and select hotels this way. If they wish, users can completely change the search results, according to the evaluative criteria they prefer.

Overall, this in-depth look at the Booking.com interface reveals the centrality of evaluative criteria. Multiplying categories of qualification of all sorts, the platform is first and foremost an evaluative infrastructure. The algorithmic ranking, developed by the platform to favor potential matches, is an important tool therein. But it is integrated into a much broader infrastructure with a multiplication of possibilities for customers to navigate this space and evaluate the accommodation offers, according to a diversity of evaluative criteria and the corresponding categories. As in the eBay case analyzed by Kornberger et al. (2017), the evaluation is dynamic and open-ended; the platform does not want to provide a definitive rating or ranking "but rather a complex set of possibilities for making connections" (Kornberger et al. 2017: 87).

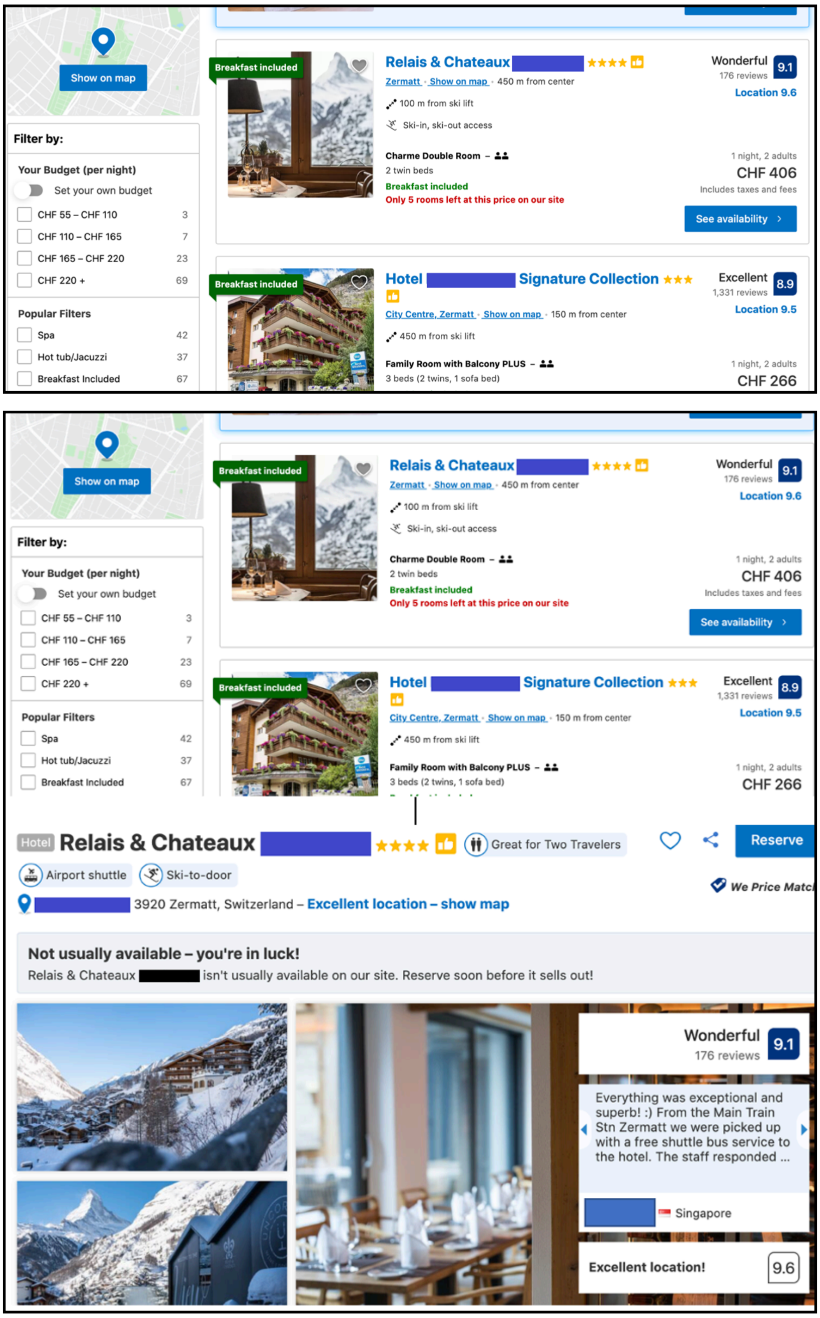

Figure 1. Example of search results (above) and hotel page (below) on Booking.com.
Notes: Above on the left, some of the different filters that can be used to adapt search result list. On the list, review scores are prominently displayed. On the image below that, one can see the indication "great for two travelers", the yellow thumb symbol that stands for the preferred partner program, and the highlighting of specific comments.

Two types of data sources

Data that goes into building evaluative categories come from two types of sources. On the one hand, there is information that is entered by hotels themselves, often based on some type of objective criteria.

Elements such as room size, facilities (is there parking, a pool, a spa) or stars rating are objectively verifiable and sometimes officially sanctioned features that are part of the evaluative infrastructure offered on the website. On the other hand, there is user-generated content that becomes part of the evaluative infrastructure, often through the intermediation of calculative devices built by the platform. After their stay, Booking.com automatically sends an email to customers, inviting them to review the establishment they stayed at. Users are asked to rate a number of separate dimensions (staff, cleanliness, location, quality/price, comfort, facilities, wi-fi) using a four category "smiley scale" and to leave positive and negative comments. Booking.com aggregates these reviews into an overall score and into separate scores for the different dimensions, using a scale from 1 to 10.[6] The scores are an essential part of the ranking and are prominently displayed on the search results list. Furthermore, the reviews[7] are also used to generate additional categories that become attributed to hotels as forms of valuation – ideal for couples, great location, etc. – which will appear prominently on the hotel's page (see Figure 1). It is here that the evaluative infrastructure is generative of new valuation categories (Kornberger et al. 2017), distinctive forms of digital valuation based on aggregated and algorithmically calculated user data.

Poles of valuation

The description of Booking.com's infrastructure and its affordances has highlighted the plurality of evaluative criteria that one finds on the platform. The platform generates new evaluative criteria and categories. Inviting consumer to leave feedback through the use of a pre-formatted questionnaire (and thus "incorporating the kinds of participation that the internet itself made possible" [Gillespie 2018: 15]), Booking.com produces online consumer reviews, which become visible on the platform as grades given in various categories as well as in the form of commentaries. Furthermore, based on these reviews but also on performance measures and on information that hotels themselves enter into the platform, algorithms produce new categories as well as scores and rankings. But these forms of valuation, generated by the platform itself and algorithmically refurbished, are not the only evaluative criteria that are visible on Booking.com. At the same time, all kinds of already existing evaluative criteria (stars, room facilities,

---

[6] In this transformation from a 1–4 to a 1–10 scale, it appears that Booking.com tends to increase review scores, leading to more positive ratings overall compared to other platforms (Eslami et al. 2017). It could be that this calculative increase of scores helps the platform increase the number of reservations.

[7] Probably along with other data such as reservation histories.

labels, etc.), which originate outside of the platform, are integrated into the platform and become part of its evaluative infrastructure. The dazzling plurality of evaluative criteria that is visible on the platform is a result of this process of integration.

Different forms of evaluation (of the evaluative infrastructure) belong to four different valuation "poles" (see Figure 2) which represent different types of judgment devices. Speaking of poles allows us to distinguish ensembles of devices responding to different principles and driven by different actor types, which occupy distinct positions in the evaluative space. It also allows us to conceptualize the interplay and overlap between these different positions that structure the evaluative infrastructure of platforms and are often in tension with each other. The *lay* valuation pole is the characteristic form of digital valuation relying on OCRs, which become visible on the platform in the form of grades (both aggregated and for specific categories) and in the form of comments. The actors behind this pole are consumers or users. The *"formulaic"* pole (using Orlikowsi and Scott's (2014) expression) refers to more or less objective evaluation criteria such as room size, facilities, services offered, etc. The stars rating, developed and usually controlled by national professional or sectorial associations, builds on such criteria to distinguish between different hotel classes. The *commercial* pole refers to evaluative criteria that are linked to the commercial and marketing practices of hotels. This includes special deals that are sometimes put forward and that may also be linked to consumer fidelity programs (on Booking.com, returning customers can obtain so-called "Genius levels" which give them access to special offers). Another example is the *preferred partner program* where hotels pay higher commissions to be better ranked. The commercial pole also includes qualifications of goods and services that aim to distinguish the offer by creating singularities designed to differentiate oneself from competitors. In the case of Booking.com, brands or quality labels are examples of evaluative criteria used to singularize hotels. The main actors behind the commercial valuation pole are thus hotels themselves. *Algorithmic* valuation, finally, refers to algorithmically generated rankings and categories produced by the platform. This valuation pole draws on the three others as it incorporates aspects from them into the calculation of new valuation forms. While lay evaluation is a critical component of algorithmic valuation, the latter cannot be reduced to a calculative operation of solely transforming lay judgments. Contrary to what is suggested by Orlikowsky and Scott's article (2014), which opposes algorithmic to formulaic evaluation, algorithmic valuation actually also draws in aspects from the commercial valuation pole (such as price, in particular) and potentially formulaic aspects (for instance, facilities).

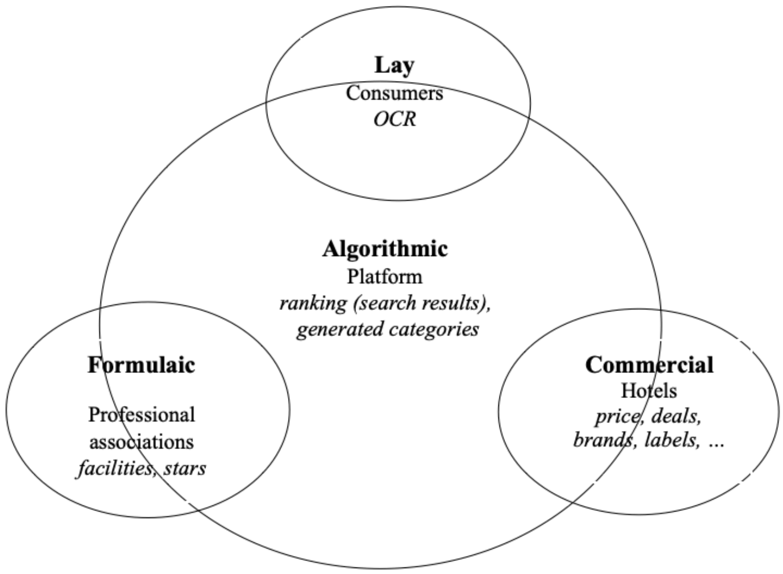

Figure 2. Valuation poles on Booking.com's evaluative infrastructure.
Source: Authors' own work.

Permissive hierarchical integration

As a "pivotal platform" in the hotel industry, Booking.com has the ability to "manipulate the processing and presentation of information to its own commercial advantage" (Lynskey 2017: 9). Its gatekeeper status derives from the control it exercises over "the flow and accessibility of information and structuring [of] the digital environment" offered on its platform (Lynskey 2017: 11). This has consequences for the way that the different valuation poles are made visible on the platform. The evaluative infrastructure of the platform mediates the valuation poles on three different levels.

First, by assembling evaluative criteria belonging to the different valuation poles, Booking.com has control over how and what is integrated, channeling what will be visible and what won't. As the description of Booking.com's infrastructure shows, the platform is very open with regard to this. It acts as a sort of "aspirator" of valuation forms and allows them to become visible, accessible, and searchable on its infrastructure. It is permissive and refrains from tightly controlling this information. This precisely corresponds to the characteristic coordination form of platforms, which is not based on tight, hierarchical control, but on distribution (Kornberger et al. 2017) and co-optation (Stark and Pais 2020). However, this should not hide the fact that this permissiveness is also a form of power exerted by the platform: it is at the discretion of the platform. To use Vallas and

Schor's (2020) apt characterization, platforms are "permissive potentates".

Second, the platform also mediates the valuation poles by setting algorithmic valuation as the privileged valuation form in the evaluative space. The ranking is the default option of search results, and the platform fully controls the way review scores are displayed and even calculated. There is thus a built-in hierarchy between the different valuation poles visible on the platform: the user who wants to privilege other criteria such as stars has to become active and adapt the search criteria or the selection. To be sure, this is enabled and even facilitated by the platform; its evaluative infrastructure is clearly plural in this sense. But the default option is always the one based on the platform-generated ranking (the algorithmic pole). The mediation operated between the different valuation poles is thus hierarchical and privileges one pole over the others. Thus the functioning of the platform will "reduce singularities to create comparabilities" (Esposito and Stark 2020, 127), in particular through the algorithmic pole. It should be added that the lay pole – which is also generated by the platform, albeit in a decentralized way – equally plays a privileged role. Compared to criteria belonging to commercial and formulaic valuation, lay valuation is particularly prominently displayed.

Finally, the platform mediates the different valuation poles not only by allowing them to be displayed and by curating the ways in which they are displayed on its infrastructure, but also by integrating aspects of lay, commercial and formulaic valuation into algorithmic valuation. The platform uses criteria from different valuation poles to create new ones. In this sense, the plural evaluative infrastructure is generative: the rankings or the new valuation categories generated by the platform (such as "ideal for couples") also integrate aspects from the other valuation poles. Search result lists do not only take into account review scores but also other elements such as facilities, price, deals, etc. In addition, these are highly personalized and depend also on the user – his/her data profile as it is known to the platform, along with reservation history. The comparability created by the platform is user-specific. Potentially this may lead to a kind of deterministic recommendation – users get to see the same or at least similar hotels, as is the case with recommendation algorithms (Seaver 2019).

Overall, Booking.com's plural evaluative infrastructure is thus mediated by the platform itself and takes the form of a permissive hierarchical integration of different valuation poles. Algorithmic valuation is clearly privileged and offers a kind of a summary of the other valuation forms; a platform specific judgment device that becomes dominant and subordinates the judgment devices controlled by other actors – be it the professionals or independent experts evaluating hotels, or the hotels themselves with their pricing but also with their marketing strategies (in Karpik's (2010) terms, *dependent* judgment devices). At the same time, the integration is permissive: it gives users the option to search for specific criteria and thus lets them customize their use of the platform at their will. It is in this hierarchical yet permissive integration of valuation poles that lies the evaluative innovation of platforms.

Destabilization of the evaluative landscape through digital valuation

In this final section, we analyze, through interviews with hotel owners and managers, the effects of the increasing significance of Booking.com (and other reservation platforms) for established routines of valuation. The interviews show that the platform provokes destabilization around three main issues: the opacity and resulting volatility of algorithmic valuation; the weakening of formulaic (in particular professional) valuation through the dominance of algorithmic valuation; and the issue of online singularity, i.e., the relationship between valuation and singularity on digital platforms. Hotels develop varied responses to each of them.

Destabilizing issue #1: The mystery of algorithmic valuation

In spite of the fact that Booking.com's evaluative infrastructure assembles different forms of valuation, reactions expressed by hotel managers clearly point to the centrality of the ranking, and therein of the review scores. From Booking.com's perspective, evaluation by customers is seen as promoting market transparency; rather than relying on the selection and advice of travel agents and professional experts, people can now do it all for themselves. From the company's point of view, developing rankings that maximize a potential for matches is beneficial to everyone:

> I think in the end it's just a customer need, and I think we tried to put the customer at the center of everything we do, and when we see that this is a need and that this helps the customer to make a decision, then we will focus on that point. (Booking.com official, January 13, 2020)

But hotel owners and managers mostly do not see it that way. In their view, rankings and in particular the way the platform uses customer evaluation to rank hotels, are anything but transparent. From their perspective, the centrality of algorithmic valuation creates volatility and insecurity. Many of them criticize the opacity of the algorithm and speculate about what goes into it.[8] The following excerpt is a typical expression of this:

> For [Booking.com], what's important is the number of rooms that one makes available, and especially all year. If you give a lot of rooms in low season and few in high season, then you are automatically worse than other hotels of the region. […] The second criterion is price comparison. The third criterion is the number of matches, the indicator of success. […] This is your rate of success, and if it's bad you will be low on the ranking with bad grades. And it will always be the bad grades that appear first. The bad comments and the bad grades. […] In the comments they will make visible the one that is related to the customer's criteria, to make sure he/she doesn't pick you. (Hotel manager, June 12, 2019)

This hotel manager is convinced that Booking.com manipulates overall scores for hotels that do not make many rooms available on the platform, and that the website makes negative reviews visible for hotels that are already ranked low. In his view, the rankings and the score are closely related: while he agrees that rankings are based on more than just OCRs, he also believes that the reviews made visible by the platform depend on the hotel ranking. While not all hotel managers share this level of suspicion, most agree that OCRs are a crucial issue. For instance, a manager tells us that while a difference between average scores of 8–8.5 does not differentiate hotels much, scores approaching 9 and especially above 9 become a clear distinguishing feature. Having a grade of 9 and above allows hotels to charge more for their rooms.

Given the importance they attach to OCRs, it is not surprising that customer review management has become an important aspect of hotels' activities. They have to keep track of reviews and will often post answers to them. Booking.com's "extranet", which is the interface to which hotels have access, actually provides them with a number of tools for review management:

> It's the system they can enter, where they have a lot of reports. Not only the reviews. We also treat them with machine learning, we read the reviews and give them reports like "Ok these are the things that have been talked about most, look at that", so we also give them some recommendations, and every hotel has access to this. (Booking.com official, January 13, 2020)

[8] According to a recent survey study, six Swiss hotel managers out of ten consider that OTAs do not communicate in a transparent and understandable manner how the rankings are established, nor how they can influence their position in these rankings (Schegg 2019).

Even with the help of such tools, management of OCRs takes up a lot of time and resources for keeping track and responding to reviews, especially because hotels are present on many websites doing OCRs. Aggregation services such as RealReview track reviews on all the major review websites (such as TripAdvisor, Google or Booking.com) in real time. Gathering this data, they use different machine learning technologies to produce daily scores but also more refined information such as what aspects customers talk about negatively or positively. Bigger hotels especially subscribe to such services, which summarize reviews and thus facilitate customer review management. One hotel manager we interviewed revealed that his hotel had even outsourced the activity of writing and publishing replies to customer reviews, working with a local service provider.

Besides keeping track of reviews and replying, in order to manage reputation (Kim and Velthuis 2021), hoteliers use OCRs to monitor their offers and services in order to improve future feedback. They may for instance use specific reviews in workforce managing, to motivate the staff, improve specific points or respond to critiques (observed already by Orlikowski and Scott 2014). But there are also more proactive strategies of review management that seek to secure better review scores. Most hoteliers solicit their clients directly and ask them to leave reviews at the end of their stay. Often, they assume that it is not the positive reviews that matter so much as the high number of reviews, as a mass of reviews will automatically drown out the few very negative ones:

> It's not even about "Please give us a good review", it's just "Review us – with a good grade we hope". The more comments we get, the more the bad one, the one that hurts us but that doesn't reflect reality, will be diluted, so it won't impact us that much. (Hotel manager, July 2, 2019)

Not all hoteliers trust this mechanism, though; many develop ways to incite customers to leave good reviews. For instance, by offering advantages to customers who book through Booking.com:

> On Booking.com, there are many more reviews because each customer automatically receives a request for evaluation, so the customer puts a comment. I know that there are hoteliers who systematically upgrade people who come via Booking.com, because of the reviews. (Hotel manager, May 14, 2019)

Without admitting the use of illicit practices, many hoteliers also point out how easy it is to cheat by writing fake reviews or purchasing them – a phenomenon that has long been studied in the restaurant industry (Luca and Zervas 2016).

All these techniques of review management aim at positively influencing OCRs, which hotels perceive as a crucial part of the digital evaluation infrastructure and of the ranking produced by Booking.com's search engine. But effects of the practices are uncertain; they are more a necessary reaction to the digital evaluative infrastructure and a testimony to its destabilizing effects than a sure path for a better position.

Destabilizing Issue #2: Weakening of formulaic valuation

Much of the literature on digital valuation has focused on the centrality of OCR and its tendency to gradually displace traditional forms of valuation. In this process, valuations built on OCRs come to replace, through competition and increased use, the formulaic categories of professional valuation. This process has been described in a number of studies on the hotel and restaurant industries (Mellet et al. 2014; Orlikowski and Scott 2014; Beuscart et al. 2016). Building on an egalitarian logic as opposed to an expert one, Booking.com's review scores also have such effects. OCR scores compete with the traditional stars rating, to the point where some hoteliers and representatives of professional associations speculate that the stars rating could soon be rendered obsolete:

> Sometimes I wonder whether customer reviews are going to replace the categories somewhat. In the sense that I think that a hotel which is a bit special could, in the future, not have a star-classification but simply an excellent review score which would make it an interesting hotel to visit. (Hotel manager, May 14, 2019)

> I personally think that the importance of these stars will diminish. [The professional association] will not say the same thing, but certain hotels already say "We distinguish ourselves through our concept or through additional services, not through stars". They also say "Stars are too strict" or "They do not correspond to our times". (Representative of a sectoral organization, June 18, 2019)

The views expressed by these two professionals point to the increased importance of OCRs for hotels. They diminish the relevance of stars rating which seems old-fashioned and rigid. Both interviewees look at it from the point of view of hotels, for which the stars rating has become less important. Others will point to the fact that the two classification systems are based on very different logics and therefore complementary, but these two comments show that, at least for specific hotel categories, the new OCR based ratings can render stars rating irrelevant. In addition, OCR based ratings create new singularities: hotels that have a very high overall grade will stand out from their competitors. For them, such a distinction can be much more useful than having a stars rating and might be privileged. This is not the case for all hotels, though, since the majority of them will have more average OCR ratings. The OCR ratings therefore also create new hierarchies, and as a consequence, not all hotels will appreciate them in the same way.

Interestingly, professional associations and sectoral organizations recognize the specific algorithmic valuation forms of platforms, and use them as an element in the construction of professional forms of qualification and valuation. The case in point here is the decision by the Swiss tourism industry organization to use the aggregated scores produced by RealReview as a criterion for inclusion into promotional campaigns on its official website, with the agreement of the professional association. Only hotels that have at least an 80% satisfaction rate (a score of at least 8 out of 10) are qualified to participate in this promotion. By integrating this form of lay judgment into categories developed by professional organizations, the legitimacy of OCRs is reinforced, as revealed by this statement from a representative of a sectoral organization:

> We are conscious that what other customers say is more and more important. That's why we work with RealReview. [...] It makes it possible for us to measure a bit the impact of already existing customer satisfaction. For hotels, but also for us it was important to be able to integrate this aspect of customer evaluation. (Accommodation marketing manager of a sectoral organization, May 21, 2019)

The rise of aggregators such as RealReview attests to the power of platforms and their characteristic valuation forms, i.e., the lay and algorithmic poles of the evaluation space. We see how aggregated OCR scores become a major component in assessing hotel qualities. Hotels and hotel associations integrate them into their formulaic or commercial valuation strategies. For some hoteliers, this undermines the very idea of professional valuation. As in cases documented for the tourism industry (Beauvisage et al. 2013; Cardon 2014; Beuscart et al. 2016), these hotel managers do not think that ordinary customers can really judge the quality of a hotel; such ratings should not be recognized by the profession:

> What I find pathetic – and I wrote this to [the professional association] – what I find really pathetic is that [the sectoral organization] is using these evaluations, through RealReview. That's a scandal! [...] What's bad is that the association that establishes professional norms, that has professional auditors who visit the establishments, puts this into the balance and decides that the most important things are customer evaluations by clients who don't know the evaluation criteria, who don't know anything at all. That's a disaster! It makes me say that they [the professional association] are

> worthless. Because they don't control the evaluation, they put in place a professional system and then afterwards they integrate things that have nothing to do with it, where one doesn't even have proof that commenters actually were customers. They take these evaluations without thinking. Why? Because they want to appear modern. (Hotel manager, June 12, 2019)

Similarly, another manager, in charge of a luxury hotel, denies the competence and legitimacy of customers to judge hotel quality:

> A client who always goes into two-star hotels and then once goes to a five-star establishment because he/she received a gift, he/she cannot evaluate a five-star hotel like a professional. Professionals know the expectations. They also have their personal opinion, but … they have a catalogue of criteria they have to refer to. That's why [comments and customer scores] cannot replace this. (Hotel manager, June 5, 2019)

But not all agree with this view. One interviewee, a hotelier who is also a member of the national association's executive committee and of the commission in charge of updating the evaluation criteria, sees this integration of OCRs into professional valuation categories much more positively:

> I think I'm the only one to think like that today, although it's gently emerging. What I explained to my colleagues in the [evaluation] commission is this: we nowadays have an important mass of comments per establishment. As for statistics, we know that it takes a certain time for statistics to have value, and I think it's exactly the same thing, after a certain number of comments, one can estimate that this value is credible and so one can integrate it into a system, and for instance say that all the hotels that are part of a given category need to have a minimum grade of 8. (Hotel manager, July 2, 2019)

This hotelier makes an argument of complementarity. He admits that customer reviews have a different focus than stars: rather than objective criteria such as room sizes or facilities, customer reviews reflect the quality of the service. But he sees it as positive. Moreover, a high number of reviewers is supposed to give such OCR scores an objective quality (Mellet et al. 2014: 61), which legitimizes their integration into professional valuation categories. Integrating OCR into other forms of valuation therefore improves their overall quality.

That the weakening and undermining of formulaic valuation by algorithmic valuation is perceived unequally reveals fault lines within members of the profession and the professional association between those pursuing a kind of "defensive professionalism" (Muzio and Ackroyd 2005) critical of algorithmic valuation and others who welcome it. Indeed, it can be beneficial for hotels to altogether bypass the professional valuation categories in favor of the less rigid platform-based valuations. One hotel manager, for instance, told us he explicitly refused to apply for a star classification, which would put the hotel into an undesirable three-star category, because he prefers the less rigid classification one can find on Booking.com combined with the branding of belonging to a specialized chain:

> Hotel manager: We don't have a star classification. The reason is simply that we would have fewer stars than we feel corresponds to our quality. […] The hotel is actually a four-star hotel, in terms of quality, service, rooms.
> Interviewer: And on Booking.com you appear as a four-star hotel.
> Hotel manager: Yes, but this is something they do themselves. We don't have four stars from [the professional association]. The problem is that when one wants to have a four-star rating, you need room service for instance, I think you even need a pool, and we don't have that. Which means we would be three-stars but we don't want that and that's why we don't have any stars at all. (July 11, 2019)

Destabilizing issue #3: Singularity in times of digital valuation

As we have seen, the platform creates new (lay and algorithmic) valuation forms which leads to new hierarchies and also new forms of singularities that become visible on the platform infrastructure (such as very high ratings, new algorithmically-generated stars rating, or other categories). According to Karpik (2010), singularities are goods and services that are multidimensional, of uncertain quality and incommensurable, and therefore cannot easily be grasped by standard methods of qualification. The relationship between singularities and valuation is always ambivalent. For producers, singularization is the ultimate form of valuation (Callon 2021); so-called "dependent judgement devices" (Karpik 2010) in the form of branding and marketing strategies allow them to distinguish their offer from all the others and make it unique. There is only one *Lausanne Palace;* a specific hotel in a specific town is never the same as any other hotel. On the other hand, judgment devices created by third parties (such as guide books) aim at reducing singularities in favor of comparability, often by creating scales, ratings or rankings that make it possible for consumers to compare different offers (Karpik 2010; Beckert and Aspers 2011). In the online environment of digital valuation, although algorithmic valuation does create new categories that can help hotels distinguish and singularize themselves, it also creates comparability along a nearly infinite number of criteria – leading to a state of near all-encompassing commensuration across hotel types as well as across geographical space. Overall, the trend is to make singularities less visible and to favor the algorithmic rankings. For hotels, this means that they are in a situation of heightened competition, which is another destabilizing effect of digital valuation.

However, as we have shown, the platform evaluative infrastructures are plural and permissive. Hotels use this permissiveness to revalue forms of commercial valuation building on singularity. In fact, singularization is frequently used as a means of bypassing platforms and increasing direct bookings (through the hotel's website, by email, phone or walk-in) by developing "value packages" such as special offers (Balsiger et al. 2022). Hotels seek to lure customers away from platforms by offering values that are only available through direct booking. With the rise of new forms of digital valuation and the increasing dominance of reservation platforms, such value packages have become crucial tools to divert consumers away from platforms. Thus, if the rise of reservation platforms has brought with it new forms of lay and algorithmic valuation, this process does not make commercial valuation forms built on singularity disappear. On the contrary, our observations indicate that it has actually led hotels to develop and promote various forms of valuation based on singularization.

While these valuation forms are used to encourage direct booking, they also find their way back on the platform when hotels make alternative forms of qualification and valuation visible on platforms through hotel *names*. To illustrate this point, we refer back to the manager of the hotel that decided not to apply for a star rating because he prefers the less rigid Booking.com classification. He went on to tell us that this is possible because of the hotel's brand, as the hotel belongs to a small chain of hip urban boutique hotels, recognizable by its name:

> The brand of course helps us in this [not needing a star rating]. As an individual hotel we probably wouldn't be able to afford this. The brand is sufficiently known in the German-speaking area and people know that it has a certain level of quality. (Hotel manager, July 11, 2019)

The chain's brand is a valuation that becomes a substitute for a star rating. Chains are part of those alternative forms of qualification and valuation that hotels have been pushing since the rise of OTAs, in order to increase the proportion of direct bookings. At the same time, because the brand is integrated into the very name of the hotel, this type of valuation is also clearly visible on booking platforms.

Rendering alternative forms of valuation visible on platforms is something that we also observe beyond those chain brands. Hotels use quality labels to attract customers through alternative, often "offline" channels in order to bypass reservation platforms. At the same time, however, they also seek to make these quality labels visible on the platform. But categories to make them visible do not readily exist on the platform. What hotels do, then, is include them in the very name of the hotel. A hotel will change its name to Hotel X and Spa, Biker Hotel Y, or *Relais & Châteaux* Hotel Z to stand out from the competition on the platform. In this way, certain quality labels, some of which would otherwise not be visible on the platform because the platform does not integrate them, appear online at the instigation of hotel managers. For instance, one hotel manager, who advertises his hotel on Booking.com as Hotel X *Relais & Châteaux,* explains the importance of this prestigious French label:

> *Relais & Châteaux* is one of the few independent labels that really puts the emphasis on quality, on cuisine. It's a good stamp for asserting the quality of a hotel. [...] We have a lot of clients who book through *Relais & Châteaux*. That's great and very interesting for us. (Hotel manager, June 5, 2019)

From a similar perspective, another manager adds the qualification "Art Boutique Hotel" to his hotel's name on Booking.com:

> We are a *hôtel de charme* in the mountains, chalet type or boutique hotel. Every room is different. We put "Art Boutique Hotel" [on Booking.com] because my wife has always been a little bit of an artist and a lot of the paintings and many of the furniture pieces are made by her. (Hotel manager, July 2, 2019)

The practices we observe here – naming one's hotel using a label or another quality category – are reactions to the destabilizing effects of digital valuation. They go beyond using the permissiveness of the platform, since they bring in categories that would otherwise not be visible. Such forms of singularization are a form of "gaming" the platform through a creative use of its affordances, which shows that hotel managers have achieved a certain mastery of this. More generally, it shows how branding, marketing, and other forms of techniques of singularization, persist in the age of digital valuation. They are used to bypass digital valuation, but they are also brought into the platform infrastructure.

Discussion and conclusion

Markets are characterized by different kinds of valuation that preexist the rise of digital platforms. These valuations are often controlled by professionals, experts, taste makers or other third parties (Karpik 2010). The rise of digital platforms and their user-generated as well as algorithmic valuation challenges the role of these other forms of valuation, and modifies market valuation (Mellet et al. 2014). In this article we seek to advance the analysis of valuation on platforms through a case study of Booking.com. Previous studies have either focused on the role of lay valuation through online consumer reviews – showing how they increasingly compete with and displace traditional forms of valuation by experts or professionals – or have described platforms as plural evaluative infrastructures. We draw on both these perspectives to characterize the evaluative innovation that platforms constitute. Our analysis suggests that digital valuation on platforms is constituted by a combination of valuation poles: lay, commercial, formulaic and algorithmic valuations coexist on digital platforms and are all highly visible, searchable and comparable. However, the combination is structured by what we call "permissive hierarchical integration". Pursuing the goal of maximizing the number of potential transactions, the platform decides which qualities become visible, how they are displayed, and which ones are shown by default. It privileges lay and especially algorithmic valuation, the latter summing up the other valuation forms and specifically customized to individual consumers' data profiles. At the same time, evaluative criteria are plural: users have the option to search all possible criteria and use the infrastructure at their own will, making the integration of valuation poles not only hierarchical but also permissive.

Because an increasing number of transactions take place on platforms, the digital evaluative infrastructure that they create destabilizes the overall valuation landscape. It provokes volatility for hotels that particularly struggle with the lack of transparency of algorithmic valuation and the centrality of customer ratings therein. Algorithmic valuation also tends to compete with and even undermine formulaic forms of valuation, as shown by early studies on OCRs in the tourism industry (Beauvisage et al. 2013; Cardon 2014; Orlikowski and Scott 2014; Beuscart et al. 2016). Yet not all hotel owners or managers see this as problematic: parts of the profession even actively *support* algorithmic and lay valuation in its competition with formulaic valuation. They see value in integrating aggregated customer review scores and find advantages in the less rigid valuation categories offered on platforms. Finally, the digital evaluative infrastructure also destabilizes the valuation strategies of hotels seeking singularity. Although the evaluative infrastructure is permissive and customizable, it also constitutes an environment of all-encompassing comparability and commensurability. This goes to the detriment of the creation of distinctive, singular qualities. But we see that such valuation strategies of singularization do not disappear. While hotels mostly use quality labels or brands as a way to favor direct bookings, they also find ways to make them visible on the platform – for instance by adapting their name. This is, of course, a somewhat artisanal way to make certain specific forms of valuation visible on the platform. Quite clearly, this rather rudimentary instrument illustrates both the power and the allure of platforms (Kenney and Zysman 2016).

This study contributes to better understanding how, in digital valuation, different forms of valuation are in play. It goes beyond a view that sees digital valuation only as a threat to traditional, especially formulaic valuation, to show how digital platforms lead to a reassembly of different forms of valuation that belong to different valuation poles. Digital valuation thus contains all these other forms of valuation. And while it does increase the weight of lay valuation (through online consumer reviews), this does not make other valuation forms disappear. Our study shows, indeed, how the algorithmic valuation produced by the platform also incorporates commercial and even formulaic valuation.

Two further aspects of the characteristic interplay between valuation poles in digital valuation processes are particularly noteworthy. First, the use of valuation poles and their interplay is fundamentally shaped by the platform's commercial interests. The platform's main interest is to maximize concluded transactions, and it will seek to assemble and tweak valuation poles in its favor. This is achieved through an algorithmically produced search result list, which is supposed to show the most likely matches first for a given customer. This search result list, which appears as a ranking, is thus guided by a commercial logic of maximizing transactions. This is a difference from review websites like TripAdvisor or Yelp, on which the seminal studies of digital valuation were based (Jeacle and Carter 2011; Scott and Orlikowski 2012; Mellet et al. 2014; Orlikowski and Scott 2014). Such review websites produce ratings but without making transactions; they are "only" judgment devices. In the case of Booking.com, the search results, which prominently display OCRs, also appear as a sort of judgment device of hotel qualities and therefore of the hotel market at large. Yet they first and foremost serve to produce transactions, making Booking.com both judge and jury. This aspect is most problematic for hotels, as algorithmic uncertainty and its stakes are heightened.

At the same time, hotels find other means to pursue their commercial interests in this evaluative landscape, which are also facilitated by the platform environment. The goal of maximizing transactions leads the platform to allow manifold possibilities for users to search according to other criteria of evaluation. It is not a fully commensurate space with rankings that have overwhelming force. The plasticity of the interface leaves space for multiple manipulations, from which hotels can also benefit.

Second, the study enriches the literature on digital valuation which so far has failed to distinguish between the rejection of lay valuation and algorithmic valuation by putting them together in the category of "platform" valuation. Our study finds that when lay valuation is opposed to formulaic valuation, algorithmic valuation becomes somehow acceptable: the mass of OCRs aggregated by algorithms is seen as statistically relevant.[9] It shows the ambivalence of market actors who are forced to play along with the rules of the game, when confronted with the plural evaluative infrastructure of platforms.

This type of case study invites further studies that look into valuation on different kinds of platforms and in different markets. How do other platforms (for instance for food delivery or music streaming) handle the integration of a variety of valuation poles? Is formulaic and commercial valuation present on other platforms, and to which extent is it challenged by lay and algorithmic valuation? It is likely that whereas all platforms function as evaluative infrastructures, the way the integration of valuation poles is structured will differ. In turn, this differential integration affects the forms of reactivity of market participants and their possibilities of singularization when facing the platform infrastructure.

Acknowledgment: The authors would like to acknowledge the Swiss National Science Foundation and its Digital Lives program for generously funding the research on which this article is based. Many thanks also to Muriel Surdez, who has commented on an early draft of this paper, and to the two anonymous reviewers and the managing editor at *Valuation Studies*. Their very constructive comments were extremely useful in advancing the analysis and theoretical argument.

References

Ahrne, Göran, Patrik Aspers, and Nils Brunsson. 2015. "The Organization of Markets." *Organization Studies* 36(1): 7–27.

Balsiger, Philip, Thomas Jammet, Nicola Cianferoni, and Muriel Surdez. 2022. "Coping with Digital Market Re-Organization: How the Hotel Industry Strategically Responds to Digital Platform Power." *Competition & Change,* online first.

Beauvisage, Thomas, Jean-Samuel Beuscart, Vincent Cardon, Kevin Mellet, and Marie Trespeuch. 2013. "Online Consumer Reviews. Markets Facing Common Opinion." *Reseaux* 177(1): 131–161.

Beckert, Jens, and Patrik Aspers (Eds.) 2011. *The Worth of Goods. Valuation & Pricing in the Economy.* Oxford: Oxford University Press.

Bessy, Christian, and Pierre-Marie Chauvin. 2013. "The Power of Market Intermediaries: From Information to Valuation Processes." V*aluation Studies* 1(1): 83–117.

Beuscart, Jean-Samuel, Kevin Mellet, and Marie Trespeuch. 2016. "Reactivity without Legitimacy? Online Consumer Reviews in the Restaurant Industry." *Journal of Cultural Economy* 9(5): 458-475.

[9] We would like to thank one of the reviewers for pointing out this observation.

Callon, Michel. 2021. *Markets in the Making. Rethinking Competition, Goods, and Innovation*. Princeton, NJ: Princeton University Press.

Cardon, Vincent. 2014. "Des chiffres et des lettres: Évaluation, expressions du jugement de qualité et hiérarchies sur le marché de l'hôtellerie." *Réseaux* 1(183): 207–245.

Curchod, Corentin, Gerardo Patriotta, and Laurie Cohen. 2020. "Working for an Algorithm: Power Asymmetries and Agency in Online Work Settings." *Administrative Science Quarterly* 65(3): 644-676.

Eslami, Motahhare, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "'Be Careful; Things Can Be Worse than They Appear': Understanding Biased Algorithms and Users' Behavior around Them in Rating Platforms." *The International AAAI Conference on Web and Social Media*.

Espeland, Wendy Nelson, and Michael Sauder. 2007. "Rankings and Reactivity: How Public Measures Recreate Social Worlds." *American Journal of Sociology* 113(1): 1–40.

Esposito, Elena, and David Stark. 2020. "What's Observed in a Rating? Rankings as Orientation in the Face of Uncertainty." In *The Performance Complex. Competition and Competitions in Social Life*, edited by David Stark, 123–143. Oxford: Oxford University Press.

Fradkin, Andrey. 2017. "Search, Matching, and the Role of Digital Marketplace Design in Enabling Trade: Evidence from Airbnb." SSRN Scholarly Paper ID 2939084. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.2939084.

Gillespie, Tarleton. 2014. "The Relevance of Algorithms." In *Media Technologies*, edited by Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot, 167–194. Cambridge, MA: The MIT Press.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven, CT: Yale University Press.

Jeacle, Ingrid, and Chris Carter. 2011. "In TripAdvisor We Trust: Rankings, Calculative Regimes and Abstract Systems." *Accounting, Organizations and Society* 36(4–5): 293–309.

Just, Natascha, and Michael Latzer. 2017. "Governance by Algorithms: Reality Construction by Algorithmic Selection on the Internet." *Media, Culture & Society* 39(2): 238–258.

Karpik, Lucien. 2010. *Valuing the Unique: The Economics of Singularities*. Princeton, NJ: Princeton University Press.

Kenney, Martin, and John Zysman. 2016. "The Rise of the Platform Economy." *Issues in Science and Technology* 32(3): 61–69.

Kim, Bomi, and Olav Velthuis. 2021. "From Reactivity to Reputation Management: Online Consumer Review Systems in the Restaurant Industry." *Journal of Cultural Economy March,* 1–19.

Kirchner, Stefan, and Jürgen Beyer. 2016. "Die Plattformlogik Als Digitale Marktordnung." *Zeitschrift Für Soziologie* 45(5): 324–339.

Kirchner, Stefan, and Elke Schüssler. 2019. “The Organization of Digital Marketplaces: Unmasking the Role of Internet Platforms in the Sharing Economy.” In *Organization Outside Organizations. The Abundance of Partial Organization in Social Life,* edited by Göran Ahrne and Nils Brunsson, 131–154. Cambridge, MA: Cambridge University Press.

Kornberger, Martin, Dane Pflueger, and Jan Mouritsen. 2017. “Evaluative Infrastructures: Accounting for Platform Organization.” *Accounting, Organizations and Society* 60: 79–95.

Luca, Michael, and Georgios Zervas. 2016. “Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud.” *Management Science* 62(12): 3412–3427.

Lynskey, Orla. 2017. “Regulating ‘Platform Power.’” SSRN Scholarly Paper ID 2921021. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.2921021.

Mellet, Kevin, Thomas Beauvisage, Jean-Samuel Beuscart, and Marie Trespeuch. 2014. “A ‘Democratization’ of Markets? Online Consumer Reviews in the Restaurant Industry.” *Valuation Studies* 2(1): 5–41.

Muzio, Daniel, and Stephen Ackroyd. 2005. “On the Consequences of Defensive Professionalism: Recent Changes in the Legal Labour Process.” *Journal of Law and Society* 32(4): 615–642.

Orlikowski, Wanda J., and Susan V. Scott. 2014. “What Happens When Evaluation Goes Online? Exploring Apparatuses of Valuation in the Travel Sector.” *Organization Science* 25(3): 868–891.

Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information.* Cambridge, MA: Harvard University Press.

Schegg, Roland. 2019. “Stabilisation de la vente directe dans l’hôtellerie. Résultats d’une enquête en ligne sur la situation de la distribution dans l’hôtellerie Suisse pour l'année 2018.” Sierre: Haute École de Gestion, HES-SO.

Scott, Susan V., and Wanda J. Orlikowski. 2012. “Reconfiguring Relations of Accountability: Materialization of Social Media in the Travel Sector.” *Accounting, Organizations and Society* 37(1): 26–40.

Seaver, Nick. 2019. “Captivating Algorithms: Recommender Systems as Traps.” *Journal of Material Culture* 24(4): 421–436.

Srnicek, Nick. 2017. *Platform Capitalism*. Cambridge: Polity Press.

Stark, David, and Ivana Pais. 2020. “Algorithmic Management in the Platform Economy.” *Sociologica* 14(3): 47–72.

Vallas, Steven, and Juliet B. Schor. 2020. “What Do Platforms Do? Understanding the Gig Economy.” *Annual Review of Sociology* 46(1).

Wang, Tao, Filippo Carlo Wezel, and Bernard Forgues. 2016. “Protecting Market Identity: When and How Do Organizations Respond to Consumers’ Devaluations?” *Academy of Management Journal* 59(1): 135–162.

Philip Balsiger is professor of economic sociology at the University of Neuchâtel. He is interested in markets, market change and valuation, in particular by studying the consequences of market contention and digitalization processes. His work has been published in outlets such as *Socio-Economic Review, Competition & Change, Journal of Consumer Culture, and European Journal of Sociology*. He is the co-editor of *The Contested Moralities of Markets* (Emerald Publishing, 2019) and author of *The Fight for Ethical Fashion* (Routledge, 2016).

Thomas Jammet is scientific assistant at the School of Social Work Fribourg (HES-SO) and Research Associate at the University of Neuchâtel. His work focuses on the digital communication strategies of organizations and on the transformation of professional activities through digitalization. He has published several articles in French-language peer-reviewed social science journals and contributed to two projects funded by the Swiss National Science Foundation.

Valuation Studies 9(1) 2022: 79-107

Theme issue contribution

What Does It Mean to Measure a Smile? Assigning numerical values to emotions

Maria Arnelid, Ericka Johnson and Katherine Harrison

Abstract

This article looks at the implications of emotion recognition, zooming in on the specific case of the care robot Pepper introduced at a hospital in Toronto. Here, emotion recognition comes with the promise of equipping robots with a less tangible, more emotive set of skills – from companionship to encouragement. Through close analysis of a variety of materials related to emotion detection software – iMotions – we look into two aspects of the technology. First, we investigate the how of emotion detection: what does it mean to detect emotions in practice? Second, we reflect on the question of whose emotions are measured, and what the use of care robots can say about the norms and values shaping care practices today. We argue that care robots and emotion detection can be understood as part of a fragmentation of care work: a process in which care is increasingly being understood as a series of discrete tasks rather than as holistic practice. Finally, we draw attention to the multitude of actors whose needs are addressed by Pepper, even while it is being imagined as a care provider for patients.

Keywords: care robots; emotions; emotion detection software; care; digital valuation

Maria Arnelid is a PhD student in Gender Studies, Linköping University.
Ericka Johnson is Professor of Gender and Society, Linköping University.
Katherine Harrison is Associate Professor in Gender Studies, Linköping University.

© 2022 The authors This work is licensed under a *Creative Commons Attribution 4.0 International License*.
https://doi.org/10.3384/VS.2001-5992.2022.9.1.79-107
Hosted by Linköping University Electronic press
http://valuationstudies.liu.se

Introduction

In a children's care ward in a Toronto Hospital, Pepper – a white, plastic, child-sized robot with blue-ringed eyes and a tiny smiling mouth – has been introduced as part of the staff. A news story covering this development shows montages of Pepper and children dancing together in the hospital lobby, people taking smiling selfies with Pepper in a hallway, Pepper letting children navigate the iPad-like interface attached to its chest, Pepper making a sick child in his hospital bed laugh out loud.[1] All the while, the narrator, together with hospital administrators and parents, is talking about how useful Pepper is for calming children down, alleviating stress and anxiety, and simply making a stay in hospital more fun.

Pepper is presented as a robot that can speak with emotion and be a friendly conversation partner, but Pepper's smile is an unmoving, sculpted part of its robot face. Yes, Pepper smiles, but continuously and unchangingly, like Barbie, Buddha, or the Mona Lisa. However, Pepper is equipped with facial recognition software and can use the sensors in its[2] 'eyes'[3] to detect human emotional expression (from human faces) and adapt its behaviour based on the perceived mood of the human it is interacting with. In this sense, Pepper responds to human smiles, which means that Pepper's algorithms identify smiles, interpret their meaning, and change Pepper's responses and conversation accordingly. In the hospital in Toronto, this is done specifically to help lift the mood and engage the emotions of the person Pepper is interacting with – Pepper tries to make sick children happy. Pepper wants them to smile. The slippage between these two – making people happy and making people smile – is the topic of this article.

Robots like Pepper are increasingly being used in different kinds of care settings from schools, to care facilities for older adults, to

---

[1] Adam (2018) "Meet Pepper – Canada's first emotionally sensitive robot for sick kids". Global News, 6 May. https://globalnews.ca/news/4180025/pepper-canada-robot/, accessed 21 March 2021.

[2] In the video clip described at the beginning of this article, the issue of gender is initially addressed, with the narrator saying that Pepper prefers the pronoun 'it' (we note the interesting ascription of agency to Pepper in this sentence); but then, in the combined interviews throughout the rest of the clip, Pepper is referred to as 'he'. This tendency to (mis)gender humanoid robots in general and Pepper in particular is discussed in Robertson (2010, 2017); Søraa (2017); Kennedy and Strengers (2020), and it has been noted that Pepper is ascribed non-human, non-binary, male and female genders in different situations. For the purpose of simplification and clarity, we will refer to Pepper as 'it', but remind the reader and ourselves that this is just a placeholder for something much more complex and slippery, particularly in the context of care robots and the gendered understandings evoked by care practices.

[3] Noteworthy, though, is that Pepper's camera is located slightly above the robot's eyes, in its forehead.

hospitals.[4] Unlike industrial robots designed to manipulate physical objects in precise, measurable ways, these robots are programmed to perform a less tangible, more emotive set of skills – skills that range from simple encouragement to companionship. This article focuses on the growing use of social robots like Pepper to provide different kinds of emotional care, and more precisely, on how these human–robot social interactions depend on the existence of a particular set of measurements. These measurements make possible such interactions by allowing the robot to read the faces of children (or others), assessing and assigning emotions to them, and responding appropriately. The measurements also make it possible to monitor and account for the effects of such an interaction. Or, as Andrew McStay (2018) puts it: 'The industrial significance is this: if one can affect emotions and make people feel a certain way, an organization has an increased chance of capturing attention, making a desired impression and affecting decision-making. This gives emotional life economic value' (2018: 17).

The work being done by robots such as Pepper plays an important role in making visible the often 'invisible' labour (Star 1999; Duffy 2011; Allen 2013) performed in institutional care settings. The notion of 'invisible' labour is often applied to tasks that are essential to the smooth running of an organization, technology, or relationship, but which go unnoticed (and consequently are undervalued). As such, measurements associated with robotic care can make clearer how the emotional labour of care work is valued by those who receive it and those who are responsible for administrating it. Making care accountable[5] requires a quantifiable measurement of whether these robots are effective – both in accurately reading the humans to whom they are assigned for care and in responding appropriately to these humans such that they improve their quality of life in some way. But what does it do to care interactions when they become programmable and accountable in this way?

---

[4] See for example: BBC News, "Pepper robot to work in Belgian hospitals", 14 June 2016. https://www.bbc.com/news/technology-36528253, accessed 3 February 2021.
Do (2018), "Meet Pepper: An AI robot that will reduce wait times in hospitals", 31 October (2018). https://news.engineering.utoronto.ca/meet-pepper-an-ai-robot-that-will-reduce-wait-times-in-hospitals/, accessed 3 February 2021.
Bayern 2020, "How robots are revolutionizing healthcare". ZDNet, 1 July 2020. https://www.zdnet.com/article/how-robots-are-revolutionizing-healthcare/, accessed 3 February 2021.

[5] Here we understand "accountable" to mean both economic valuing of labour by institutions and commercial organizations, and making the development of care robots morally and ethically accountable.

This question is at the heart of the broader research project surrounding this article,[6] where we explore different ways in which concepts such as empathy, affect, touch, and care are incorporated and shaped in the development of social robots for care work. This collective research approaches how different disciplinary understandings of what makes good social interaction between human and robot are brought into dialogue in the context of care robot development. In this article, we want to bring this kind of approach to a study of Pepper as caregiver. We want to investigate the practices involved in enabling Pepper as an emotional caregiver, and ask what happens to emotions, as well as care, when they are programmed to be delegated to Pepper? Additionally, and to address concerns about norms and values engaged by care robots, we will be thinking through how Pepper and care robots in general highlight context and emphasize the specificity of contextualized technology. Technology does different things in different places. Pepper is no different. Pepper is a robot which can be used to interact with humans in many different contexts and for many different reasons. As Pepper moves into a new context, for example, into a children's ward, this move exposes concerns that Pepper's presence is addressing in that particular place and time, with those particular people or groups. And it triggers a question about what roles social institutions, like hospitals, play in Pepper's placement.

To grasp the many agendas, discourses, hopes, and practices entangled in the introduction of Pepper as a caring companion into a hospital environment, we take inspiration from Donna Haraway's notion of the 'imploded knot'. A different kind of cyborg from the one that Haraway uses so provocatively in her manifesto, Pepper nonetheless represents the 'implosion of the technical, organic, political, economic, oneiric, and textual that is evident in the material-semiotic practices and entities in (…) technoscience' (Haraway 1997: 12; see also Dumit 2018). Haraway's attention to the multiple strands that knot together in technoscientific practices and entities is useful here in lifting the very sets of thinking and scholarship that inform our reading of Pepper. As such, we find it helpful to offer the reader a brief overview of some of the key ideas from scholarly literature.

Getting into conversation

In this section, we will introduce some central scholarly discussions that inform and inspire our discussions around Pepper as caregiver. The section is organized into two subsections. The first focuses on

---

[6] The authors are part of an interdisciplinary research project that brings together robot designers, computer scientists and science, technology, and society theorists experienced in ethnographic studies of affective human–machine interactions. The team is exploring cases of robots in the iterative design/early testing phase.

different literatures associated with emotions and AI and connects this to the concept of norms. This is followed by a second subsection introducing the critical scholarship around care that has informed our analysis of Pepper in this article.

Emotions and AI

The notion of social interaction with robots and AI has long been a theme in science fiction. The present moment, however, is marked by a turn to the real with digital assistants such as Siri[7] or Alexa[8] increasingly becoming part of our households, where daily chores such as vacuuming[9] or lawn mowing[10] are taken care of by robots. Meanwhile, robot teaching assistants are being trialled for use in schools[11] and robotic animals such as dogs[12] or seals[13] are widely touted as bringing invaluable companionship to older adults or those who are ill. New understandings of relationships in which robotic non-humans become part of private, intimate life are as urgently required as the ethical and legal frameworks demanded in order to keep them accountable. As such, this article is part of increasing attention to the promises and challenges of 'emotional AI'.[14]

Scholarship around emotional AI can be dated back to Rosalind Picard's work in the mid-1990s on affective computing:

> I have come to the conclusion that if we want computers to be genuinely intelligent, to adapt to us, and to interact naturally with us, then they will need the ability to recognize and express emotions, to have emotions, and to have what has come to be called 'emotional intelligence'. (Picard 1997: x)

---

[7] Apple, 'Siri'. https://www.apple.com/siri/, accessed 20 September 2021.

[8] Amazon, 'What is Alexa?'. https://developer.amazon.com/en-US/alexa, accessed 20 September 2021.

[9] iRobot, 'Roomba'. https://www.irobot.se/roomba, accessed 20 September 2021.

[10] Worx, 'Is Worx Landroid the best robot mower money can buy?' https://eu.worx.com/landroid/en/, accessed 20 September 2021.

[11] Cookson (2019). 'Robot trained to be useful teaching assistant in three hours'. Financial Times, 23 October 2019. https://www.ft.com/content/5458d814-f4bd-11e9-b018-3ef8794b17c6, accessed 20 September 2021.

[12] Joy For All, 'Lifelike Robotic Pets for Seniors'. https://joyforall.com/, accessed 20 September 2021.

[13] Paro, 'Paro therapeutic robot'. http://www.parorobots.com/, accessed 20 September 2021.

[14] For a useful overview and unpacking of "emotional AI" see McStay (2018).

Picard's work – together with that of Cynthia Breazeal – made emotional interaction with a robot or AI impossible to ignore. Picard herself acknowledged in the introduction to her seminal 1997 volume *Affective Computing* that the idea of computers having emotions might sound 'outlandish'. However, as she explains, this response was grounded in the prevailing notion that rationality and emotion are two completely distinct and independent mechanisms. Instead, she makes the argument that emotions 'influence the very mechanisms of rational thinking' (1997: back cover) and proceeds from there to argue that: 'Computers do not need affective abilities for the fanciful goal of becoming humanoids; they need them for the meeker and more practical goal: to function with intelligence and sensitivity toward humans' (Picard 1997: 247).

In a chapter titled 'Recognizing and Expressing Affect', Picard details the various models available for recognizing and expressing affect. Many of the models she describes have developed significantly since this book was published. For example, she details real time processing as a major stumbling block with facial recognition – something that contemporary facial recognition software claims to have resolved. However, Picard's work laid the foundation for two key premises: (i) emotions as integral to intelligence, and (ii) emotions as tangible, measurable, and accurately reproducible. The possibility for a machine to read accurately the emotional expression of a human (thus allowing the next 'step' in terms of programming an appropriate response) is a key part of creating the conditions for the 'natural' feeling interaction of which social roboticists dream.

One of the models that Picard details is Paul Ekman's 'Facial Action Coding System' which is also the basis for one of the best-known facial-recognition softwares, Affectiva.[15] Ekman's work codified facial expressions for a series of emotions (see Ekman 1976; Ekman and Rosenberg 2005), identifying expressions and muscle movements which are claimed to be relevant across many different cultures and contexts. This approach leans heavily on a Darwinian understanding of emotional expression as part of biological evolution, more basic and universal than local cultural expressions. Indeed, Ekman and Friesen (1978) termed these facial expressions 'basic emotions' and posited that they would be relevant for all humans. This system has been widely used by psychology researchers, computer and AI developers and, not least, animators, to read and/or reproduce emotional expression in faces.

The idea of 'basic emotions' has also been the subject of much discussion by social sciences-oriented scholars interested in the turn towards digital affect. This literature focuses on exploring issues such

---

[15] Notably, Picard is one of the founders of Affectiva. https://www.affectiva.com/who/about-us/, accessed 18 May 2022.

as the 'range of phenomena encompassed by terms such as *affect, emotion, feeling* and *mood*' (Stark 2019: 118, emphasis in original; see also Papoulias and Callard 2010), and the consequences of adopting particular models, for example Rhee's (2018) reminder that '(e)motional labor, which emerges from uneven power relations, insists on the expression of normative emotions, in many instances as evidence of humanness or citizenship' (2018: 101). Overall, this body of scholarship leans more generally towards 'the messier idea that emotions might not be fixed objects, but culturally constructed experiences and expressions defined through historical and situational circumstances' (McStay 2018: 4). We will return to these disciplinary differences in understanding emotions throughout the text.

In this article, we are interested in exploring the role that the measuring of emotions plays in care work done by social robots. This is a significant part of our article because the *how* – what practices and what science are part of the measuring of emotions – is where the work happens. We are looking at the nitty gritty of valuation practices (inspired by science and technology studies (STS) work on users, manuals, instructions and classification [Akrich 1992; Goodwin 1994; Bowker and Star 1999]). We are doing this against a background of how the technology is used, which provokes questions about why this work is being done, and also points to larger questions about the interaction of this work with other structures and norms (inspired by work practice researchers like Cockburn 1983, Orr 1996, and Suchman et al. 1999).

When we are talking about the relationship between norms and technology in the context of emotion recognition, we are beginning from an understanding of norms that takes inspiration from early sociological work (Parsons 1951; Joas and Knöbl 2009 [2004]) and examines how social norms shape what is possible and acceptable, particularly in institutional settings like care homes and hospitals. This understanding of norms is also relevant for seeing how they shape practices in medicine and science (Merton 1942, 1973; Bucchi 2015). However, our work with norms is also highly influenced by the way they have been examined in STS, both as reproduced and materialized in technology (Winner 1980); visible in the discourses and tropes used to describe technology (Haraway 1997; Johnson 2019); internalized in our responses to technology (Rose 2007); and conscientiously challenged through design (Disalvo 2012; Ehrnberger 2017; Escobar 2017).

This means that we consider that smiles are not only valued, measured, counted, but also what normative work they are doing and what power dynamics are at play: which smiles, whose smiles, and where. From this we will try to read which norms can be articulated in the use of emotion (or at least smile) detection technology to assign values in human–robot interaction.

(Robotic) care

Since Pepper is supposed to be providing care, critical reflections that have been carried out in STS on care as a theoretical category are relevant in framing some of the observations we make. In what follows we draw particular attention to some themes within this literature, namely valuation of care, emotional labour, power and care fragmentation.

As a sociological term, care has a long history, often related to the ethics of care in professions like nursing. It often draws upon and teases out a universalizing (and naturalized) understanding of care as something 'good' which can nonetheless be dissected into parts, categorized, and then taught to those who are supposed to deliver care (Duffy 2011; Allen 2013). In more critical discussions of the term, Joan Tronto (1993) is often referenced, with her early critique of tendencies to imagine care as (universally) feminine. She defines care as: 'Everything that we do to maintain, continue and repair "our world" so that we can live in it as well as possible. That world includes our bodies, ourselves, and our environment, all that we seek to interweave in a complex, life sustaining web' (Tronto 1993: 103).

Importantly, Tronto's political argument is that care is often carried out by underprivileged groups, which serve to maintain systems of privilege for others. This line of theorizing has frequently seen care equated with devalued labour, often similar to what is called invisible labour, and which can overlap with the 'dull, dirty and dangerous' work that it is often imagined will be assigned to robots in the (near) future (Suchman 2007; Rhee 2018; DeFalco 2020). This also resonates with the emotionally subservient role assigned to Pepper as it is engaged in making sick children happy.

Another line of research into care deals with the way emotions are enacted as care practice. Here, attention is paid to emotional labour and the management of feeling, as developed initially in studies of workplace expectations, service industries, and the 'commercialization of feeling' (Hochschild 1983). It later appeared in classics like James's definition: 'Care = organization + physical labour + emotional labour' (James 1992: 2). Such studies quickly became a staple of sociological research on nursing practices (see Allen 2013) often highlighting structural issues of care provision, labour relations, and the expected 'doing' of care, physically and emotionally.

The power dynamics and politics of care have been brought into an STS discussion around the word which has blossomed in the last decade. Here, too, one is reminded that care is multifaceted and not necessarily benign or positive. Not all care is good. This conversation started with Puig de la Bellacasa's attempt to encourage an ethos of care in STS research as a response to, and extension of, Latour's suggestion that the field engage matters of concern (Puig de la Bellacasa 2011). While Latour was suggesting *matters of concern vs* *matters of fact* as a way of addressing a dreaded turn away from 'truth' or belief in science that a constructivist approach was by some thought to produce (and written before our current mayhem of alternative facts aftermaths) (Latour 2004), Puig de la Bellacasa (2011) was suggesting that a discussion of care would carry with it a critical edge, one attuned to exclusions and power dynamics in stratified worlds (2011: 86). As she points out, 'care' is a stronger word than 'concern' and can also be easily turned into a verb, 'to care'. This is important because '[u]nderstanding caring as something we do extends a vision of care as an ethically and politically charged *practice*, one that has been at the forefront of feminist concern with devalued labours' (Puig de la Bellacasa 2011: 90).

Taking this further, Martin et al. (2015) addressed the way that care 'is both necessary to the fabric of biological and social existence and notorious for the problems that it raises when it is defined, legislated, measured and evaluated' (2015: 625). They, too, point out that care is not always positive, it can have a darker side, lack innocence, and induce violence. It is selective – it can cherish some things and exclude others. And the power of care includes the power to define what counts as care and how it should be administered. Likewise, it can:

> render a receiver powerless or otherwise limit their power. It can set up conditions of indebtedness or obligation. It can also sediment these asymmetries by putting recipients in situations where they cannot reciprocate. Care organizes, classifies, and disciplines bodies. Colonial regimes show us precisely how care can become a means of governance. It is in this sense that care makes palpable how justice for some can easily become injustice for others. (Martin et al. 2015: 627)

These aspects, too, are very relevant to our discussion of robots. One can ask what bodies are being organized, and how, by care robots like Pepper. That question easily reshapes into a question of which organizations (hospitals? nursing homes?) are tasked with caring for which bodies, and therewith what power dynamics the use of Pepper is reproducing.

We suggest that keeping these critical stances to the concept of care in mind can remind us that an analysis like ours has political implications for many people, care givers and care recipients alike, not just for the development of robots or their integration into care provision budgets. Furthermore, they make clear the necessity of critically examining newly emerging modes of caregiving (such as Pepper), with particular attention to how digitization may reconfigure understandings and practices of care. In the case of Pepper, for example, this includes tracing a line from earlier analyses of how care was dissected and categorized in order to be taught to humans, to the dissection of care work necessary for it to be programmed into a robot. One of the most tangible implications of this critical stance can be a recognition of the politics of care fragmentation.

'Care fragmentation' (Vallès-Peris and Domènech 2020) refers to a process in which the various elements of care work go from being understood as a holistic practice to being understood as a series of tasks that can be thought of, and carried out, separately. Vallès-Peris and Domènech (2020) use the term care fragmentation in an interview study investigating roboticists' imaginaries of care robots. Here, care fragmentation refers to the process in which care becomes 'conceptualized as a set of tasks that can be separated in[to] pieces made of different tasks, with some of these pieces being able to be delegated to the robot and others not' (2020: 165). More specifically, they describe how the physically demanding and strenuous tasks of care work become separated from the 'affective tasks' (2020: 165) (specified in the article for example as conversations, quality time, and creative interaction).

Similar to Vallès-Peris and Domènech, we will argue that articulation of Pepper as a caregiver depends on practices of care fragmentation. However, as we will show, care fragmentation in our case differs from that in Vallès-Peris and Domènech's study. It does so since Pepper comes with the promise precisely of managing the kind of 'affective tasks' that roboticists placed outside of care robots' range in their study. What we will show is that emotion recognition took part in another kind of care fragmentation, one which aimed at making possible affective care interactions between Pepper and patient.

Brochures, demos, and implosions

The authors of this text were introduced to Pepper thanks to collaboration with the *Machine Perception and Interaction Lab* at Örebro University. In this lab, Pepper has been tested as a care robot for older adults, coaching residents in a care home for older adults in exercising (Akalin et al. 2019). Pepper has also been a part of experiments exploring topics such as a sense of safety and security in human–robot interaction (Akalin et al. 2017). At one point during our collaboration, a robotics professor showed us the facial recognition software that they run on Pepper: Affectiva. The professor demonstrated how the software assigned emotional interpretation of her various facial movements while she went through a roster of smiling, frowning, looking confused.

Demonstration of the Affectiva software made us curious about the role that it played in our colleagues' research and that of others working with Pepper. The encounter therefore inspired us to go further with investigating the use of emotion detection in Pepper. Having made this decision, we started looking into media coverage depicting Pepper in use, as a way to gain insight into how this technology is demonstrated to the public. Reading news articles, such as the story that opens this article (showing how Pepper, equipped with emotion detecting software, is used in hospital environments) was one way of learning more about how this technology is introduced.

However, we also wanted to understand the workings of the software and learn *how* it measures emotions. To do so, we turned to Affectiva's web page. Affectiva software is offered for use in several different settings. However, when used for the purpose of emotion recognition, the web page directs you to their partner, iMotions. iMotions is one of several companies and many research groups using Affectiva. iMotions uses the Affectiva technology – facial movement tracking enabled by a vast database of faces,[16] deep learning, and machine learning – as part of software that helps researchers with all stages of their emotion-tracking studies. iMotions software includes, besides the Affectiva technology, survey tools, data visualization, and a library function for previous studies. We have engaged with iMotions in two different ways. Initially, we analysed different materials made available through the iMotions website which describe the functioning of the software. One important such resource, which will reappear throughout the text, is an introduction brochure titled 'Facial Expression Analysis: The Complete Pocket Guide' (2017).

The guide is twenty-seven pages long, freely available as a pdf to download from the iMotions website, and divided into three main sections. The first section entitled – 'The Basics … and Beyond' – outlines the theory behind Facial Expression Analysis. It references Darwin, evolution, and Ekman's work on cross-cultural emotion recognition (Ekman and Friesen 1978). Put simply, the theoretical basis relates to the perceived causal relationship between facial movements and emotions exemplified in the paragraph above – the view of emotions as 'readable' (to use iMotions' own terminology) through facial movements. The second section – 'Getting Started with Facial Expression Analysis' – goes into the practicalities of using the iMotions software. It explains which types of technology are needed to use it (most importantly, a working web camera) and gives general instructions about how to set up the software. Finally, 'Facial Expression Analysis … Reloaded' deals with other measurements of emotions that can be used to complement facial expression analysis. For example, it can be complemented with a device that can be attached to one's index and middle fingers which measures the valence of one's emotional response (how strongly one experiences the emotion) using sweat measurements. The brochure appears to be aimed at beginner- or intermediate-level users of the software due to user-friendly terms (the fact that we, three social scientists, could

[16] Called the Affdex database, which contains the world's largest data set of human faces (McStay 2018:61).

understand it is a testament to that). The text is complemented throughout with pictures of demonstration faces and results charts.

The second way in which we engaged with iMotions was through a meeting where the software was introduced to us. During our exploration of the website and the brochure, a chat prompted us to make direct contact with the iMotions team for a one-hour online demonstration. This seemed like a good opportunity to ask some direct questions about the software, so we decided to participate in the demonstration as potential buyers/users of the product. In this meeting, we made clear that we were researching emotion detection and wanted to learn more about the practicalities of using the iMotions software. We were walked through the interface of the software: what one sees when one is using it. During the demonstration, a company representative illustrated how to carry out a facial movement recording by using her own web camera, 'reading' her own emotions, to use iMotions's terminology.

All of the materials detailed above – news stories such as the one about Pepper in the children's ward, the material gained from Affectiva's and iMotions' websites, the iMotions Facial Expression Analysis brochure, and the field notes that we took during the demo with the iMotions team – make up the empirical material for this article. Each of the materials has been integrated differently into our analysis. The media coverage fed our initial curiosity and helped us to understand why it is so important to pay attention to this measurement software; witnessing the embodied connection between measurement algorithms, bodies, and affects as it plays out in the children's ward of a hospital. The specific story about Pepper being introduced in the Toronto hospital also inspired the case that we return to throughout the article. The brochure helped us to understand how iMotions situate and formulate both their software and emotion detection in general. Taking part in the demonstration that iMotions gave us helped both in understanding how the software works (seeing the yellow dots on the face) as well as triggering wider curiosity about the broader scientific context in which the technology is situated.

In the article, we will 'implode' (Haraway 1997; see also Dumit 2018) the various empirical materials that we introduce above. This means that we aim to disentangle the different discourses, media, and intellectual heritages that are knotted together in the materials, all contributing to the articulation of emotion detection. We will implode the intellectual heritage of the system in the form of Paul Ekman and Wallace V. Friesen's theory of 'basic emotions' (1978), the technical details of how to make the system work, the set of bodily norms required for recognition, and the commercial rhetoric designed to increase the appeal of the package offered by iMotions. In doing so, we take inspiration from STS analyses that pay close attention to the 'interdependence of technical networks and standards, on the one hand, and the real work of politics and knowledge production on the other' (Bowker and Star 1999: 34). Approaching the material in this way highlights the 'non-innocence' (Haraway 1991 [1988]: 157) of emotion recognition software by examining its heritage and connecting it to its contemporary consequences.

iMotions and a smile

> As we laugh or cry we're putting our emotions on display, allowing others to glimpse into our minds as they 'read' our face based on changes in key face features such as eyes, brows, lids, nostrils, and lips. Computer-based facial expression analysis mimics our human coding skills quite impressively as it captures raw, unfiltered emotional responses towards any type of emotionally engaging content. (iMotions 2017: 2)

The quotation above is from iMotions' promotional material and offers an intriguing introduction to the phenomenon of emotion detection. Software such as that offered by iMotions is one example from a range of programs currently available to do the work of 'reading' emotional responses. The idea that facial expressions can be reliably correlated to emotional responses, and that such expressions can be accurately read and analysed by software (in real-time) is a premise upon which social robotics is built (Picard 1997; Ekman and Rosenberg 2005). It is a crucial part of the claim that robots like Pepper can act with and understand emotions. The ability to measure emotions accurately is important to those both for whom the robots are tasked with caring (ensuring or promising that the robot will be able to respond to their needs and feelings), and for the programmers of such robots who are trying to develop the performance of emotional responses by robots (to facilitate a smoother bond between human and robot).

The iMotions brochure uses a variety of strategies to connect facial expressions and emotions: scientific references, a particular kind of rhetorical register, different theories, and illustrations in the brochure of scientific-looking diagrams. In the demonstration meeting with iMotions, this material was meshed with the technical specificities of the software. Through the online introduction, and from the questions we were able to ask the developers and demonstrators, it became clear that all these elements are part of enabling the emotionally sensitive care that Pepper is argued to give to children at the Toronto hospital.

Below, we will illustrate how this software works and was demonstrated to us. We will argue that emotion detection such as that developed by iMotions, when used in care robots, can be understood as part of a larger process of 'care fragmentation' (Vallès-Peris and Domènech 2020): a process in which the various elements of care work go from being construed as a holistic performance of physical as well as emotional labour, all wrapped up in a professional identity – to being seen as a series of discrete tasks that can be performed by anyone, specifically a robot in our case. We will show how the promise of affective robotic care requires this kind of fragmentation where, in order for it to be manageable by a robot such as Pepper, affective care practices become about scanning the slightest movement of the corner of the mouth in order to calculate how probable it is that the patient is smiling. And we will discuss how, as this is done, some smiles are made visible and are valued, while others are not.

Using iMotions - how it works

A robot is not necessary for using the iMotions software. This became clear to us as the software was demonstrated in the meeting with iMotions. For many users, the only equipment needed (besides the software) is a computer with a standard web camera, directed at the study participant's face. When using the software, real-time images of the face are picked up by the camera and adorned: the outline of the face is marked by a thin yellow line that follows head movements. Within this outline, small yellow dots mark the ends and middle points of the eyebrows and eyes, as well as the edges of the nostrils, lips, and chin. All in all, the software covers the face in about thirty yellow dots. These dots mark crucial 'landmarks'– facial areas that are seen as containing information about emotional states, especially when read in relation to other dots as one's muscles move. The movement of these landmarks in response to a person using their facial muscles constitutes the basic data on which iMotions relies.

Then those data, i.e. the movement of facial landmarks, are fed into a classification algorithm: a type of algorithm that is used to assign data to predefined categories. In the iMotions software, these categories are made up of different emotions. The algorithm returns a numerical score which corresponds to the likelihood that one is, for example, happy. This can be judged partly by how wide the algorithm perceives one's smile to be. This shows that there is a strong connection being made here between a smile and happiness: if one smiles, there is a strong likelihood that the software will perceive that one is happy. If the algorithm returns the number 0, this means that there is 'no expression' of happiness. If the algorithm returns the number 100, the expression of happiness is suspected to be 'fully present'. In other words, emotions are measured on a scale from zero to hundred.

In our observations of the iMotions software, we have noticed some key processes of standardization that are involved in iMotions' emotion detection – standardization processes that are, by extension, key to enabling the promise of emotional competence in robotic care Below, we will delve into the emotion classification systems that iMotions uses as a basis for their software, discussing both the science that enables it and the ways in which researchers and participants have to adapt in order to fit into the system. In doing so, we pay attention to the intersection of quantification, classification, and standardization (Bowker and Star 1999) in the context of emotion detection.

iMotions (and similar competing) software requires a standardized and widely accepted conceptual framework for classifying emotions. iMotions' Facial Expression Analysis software depends on earlier, analogue, methods of classifying and cataloguing emotional states. One such system is Paul Ekman's earlier mentioned 'Facial Action Coding System' (FACS), which is explained by iMotions in the following terms: 'a fully standardized classification system of facial expressions for expert human coders based on anatomic features' (iMotions 2017:16).

As is explained in the brochure, FACS forms an important foundation for the iMotions software. FACS and the basic emotions theory, on which FACS is founded, is described in detail in the brochure. For example, the brochure dedicates several pages to illustrating the seven basic emotions that Ekman and Friesen (1978) discuss: joy, anger, surprise, fear, contempt, sadness, disgust. In the brochure, these emotions are complemented by images of different (albeit all white) faces illustrating these emotions in what could only be seen as fairly exaggerated ways (for example, the man illustrating the emotion 'fear' is depicted opening his mouth wide as if screaming).

It is these seven basic emotions that the iMotions software 'finds' through measurement of the tiny movements of landmarks in a person's face – or that Pepper looks for in faces of children in the Toronto hospital. This classification system is therefore a key component of the care fragmentation described in the previous section: basic emotions theory has laid the foundation for establishing clearly distinguishable components of an emotional spectrum, thus helping robots like Pepper in assigning an emotion to people they encounter.

But can you measure an emotion?

So far, we have described and discussed the process in which iMotions measures emotions digitally. However, emotion detection has also been associated with the important methodological question of whether emotions *can* be measured (Davies 2017). Is there a difference between other real-time tracking technologies such as step counters and sleep trackers (see for example Lupton 2016; Salmela et al. 2019) – measuring numbers of steps walked/ numbers of hours slept – and emotion detection technologies? Arguments both for and against this inevitably circulate around questions of the very nature of emotions: are emotions something that you can 'read' on a face? Are there categories of emotions that can be distinguished from each other? Are emotions something that can be represented using numbers? And if so, what is the best way to achieve this?

In his work on mood tracking apps, Davies (2017) takes up this question, arguing that there is something special to measuring emotions. According to Davies, the emphasis on catching emotions in real time that is inherent both in the mood tracking apps that Davies focuses on and in the work of iMotions implies a view of emotions as subjective and constantly shifting experiences that exist in the moment (2017: 45). Davies sees the view of emotions as in the moment, and simultaneous attempts to somehow 'capture' (2017: 39) them, as a philosophical point of conflict underlying mood tracking: 'There is a philosophical contradiction here between the privileging of immediate, unreflective experience as the essence of value and the attempt to represent it in calculable, objective form for purposes of evaluation' (Davies 2017: 39).

What Davies understands as a philosophical contradiction, between the here-and-now view of emotions in mood tracking and the simultaneous wish to 'capture' these moods, could perhaps also be framed as a disciplinary one. As we have mentioned, the approach to emotions taken by iMotions – as visible in facial movements, measurable and quantifiable – has been understood as related to a larger tendency in AI research to consider emotions as quantifiable and replicable 'discrete states' (Suchman 2007: 232–234). The view of emotions as measurable and quantifiable is, quite obviously, compatible with research in which emotions are measured and represented in a calculable and objective form. As noted earlier, this view of emotions can be contrasted with discourses in the social sciences that are more attuned to the contexts and practices of the production and expression of 'affect' (Ahmed 2004; Pellegrini and Puar 2009). Here, emotions are thought of not as discrete psychological states but rather as social and cultural practices (Ahmed 2004: 9). This view of emotions aligns less well with emotion detection practices, such as those carried out by iMotions, in which emotions are seen as residing in the physical body and possible to read through muscle movements. In other words, Davies's discussion on the philosophical conflict underlining mood tracking – between on the one hand, viewing emotions as subjective and 'in the moment' and, on the other, wanting to 'capture' them – could be connected to wider discussions on the nature of emotions and the methodological implications for how to study them.

The approach used by iMotions and other similar emotion-tracking technologies – to trace physical signs of emotions in the body – is put forth as one way out of the struggles of capturing 'real experience' without disrupting the flow of experience. In these cases, physical data is used as a tool to 'avoid any perceivable engagement with the quality or quantity of subjective experience' (Davies 2017: 45). In other words, technologies such as iMotions could be seen as using physical data as a tool to navigate between an emphasis on emotions as a subjective experience, and a wish to quantify and calculate them. To do this they rely on discrete, observable physical changes in facial muscles, pulse rate, skin temperature, etc. As such, technologies seem to hold the promise of providing accurate and reliable access to 'real' emotions unmediated by limits of language, using a system that guarantees quantifiable data.

Interestingly, the methodological challenge raised by Davies leads us back to the different disciplinary understandings of emotions that we noted earlier. The process of identifying emotional responses (by connecting physiological responses to predefined categories of emotions through a computer algorithm) *does* involve mediation by language in order for emotions to become legible to others; it is just that this happens 'out of sight'. More precisely, the promise of iMotions to capture emotions directly by tracking physiological responses is – on closer inspection – undone by the background framework of 'basic emotions' on which it depends. The algorithm that connects muscle movement with emotion functions as a kind of black box, in which particular facial expressions have already (thanks to FACS) been categorized and *named*.

The methodological challenge with iMotions then lies not in how to handle the 'the flow of experience' but rather in recognizing the effects of the interpretative framework imposed by FACS. Fundamentally, reliance of such facial recognition software on language as a way to mediate the experience is in tension with the idea of affect as 'visceral forces beneath, alongside, or generally other than conscious knowing, vital forces insisting beyond emotion – that can serve to drive us toward movement, toward thought and extension' (Gregg and Seigworth 2010). In this latter conceptualization, the emotional/ affective responses are understood to reside outside language.

Developing this further, we suggest that it is also worth paying close attention to the experimental framework that surrounds the iMotions capture of emotions and which is necessary in order to deliver on its promise of reliable emotional feedback. This experimental framework differs depending on the type of emotion detection, but in the case of iMotions includes the materials, algorithms, and science that emotions are funnelled through in order for the software to 'read' them. This framework cannot be considered a neutral transmitter of emotions, but rather shapes them in various ways in order for them to be legible, or readable. The process of making emotions readable depends on specific types of standardization of emotions which we will investigate below.

Whose smile?

We have so far addressed how differing understandings of emotions may shed light on the limitations of facial recognition software that is premised in Ekman's work on basic emotions (1976). However, a less considered aspect – and one which turns us now towards exploring some of the norms involved in care work – is the question of which faces (and emotions) constitute 'valid' subjects. Thinking about care and norms is often connected to critical studies of care, which examine what types of care are institutionalized and how, which bodies receive institutional care, and which provide them (Duffy 2011; Allen 2013), and what the power dynamics of those norms involve (Murphy 2015; DeFalco 2020). These studies inform our concern for both the historical basis for emotion recognition as well as contemporary practices of it.

An interesting aspect of the practice of recording emotional states is who is considered a 'good and reliable subject' (i.e. a subject whose emotions are worth recording) – and who is not. This is approached by Lucy Suchman in her work on affective computing (2007). Referring to the historian of medicine Otniel Dror (2001), Suchman draws a parallel between affective computing and the practice of recording, cataloguing, and enumerating emotions in laboratory sciences – practices that Dror traces back to the late nineteenth and early twentieth centuries. The aim of such cataloguing practices was to produce clear representations of 'emotional states' such as anger, fear, or excitement, and these were used in a variety of practices.

In Suchman's analysis on Dror's work, she discusses how those who were considered good subjects for representing emotional states – and chosen to be included in the research – were those who displayed 'clearly recognizable emotions on demand' (Suchman 2007: 233). These subjects can be contrasted with those who were more 'ambiguous' in their emotional expressions and therefore difficult to classify, who were excluded from research. As we will show below, similar categorizations of 'good' and 'bad' subjects (in our case, subjects for producing detectable emotional states) come to light when looking closer at the technological limitations of the iMotions software.

Whose emotions can be detected? The iMotions system is significantly limited in terms of what can be reliably captured by the material technology itself. This is visible for example in the technical requirements listed in the iMotions brochure to ensure best results: 'For online automatic facial coding with webcams, keep the following camera specifications in mind' (iMotions 2017: 24). There follows a list of five guidelines about resolution, frame rate, lens, and so on, specifying the quality of equipment required. In addition, there is a long list of 'Respondent instructions' in which the ideal set-up for capture is achieved by positioning, illumination, visibility of face, and mobility of face. As the brochure goes into increasing detail about the experimental procedure, focus turns to ways in which the more temporary aspects of a participant's appearance may affect the measurements: 'Facial expression analysis requires the visibility of emotionally sensitive facial landmarks such as eyebrows, eyes, nose, and mouth. If any of these are occluded, the face tracking and expression analysis may lead to only partial results' (2017: 27).

This then is different from the practices of recording and cataloguing emotions (like the FACS), discussed above, in which particular types of individuals are considered to produce the 'correct' performance of certain emotions. This standardization is connected rather to the practical aspects of capturing emotions digitally. Accurate reading of an emotion requires the camera to be able to clearly 'recognize' the landmarks of the face to be measured. If these are obscured by features such as large glasses, long beards, facial jewellery, hats, or side-swept bangs, then the camera is unable to place the small yellow dots which it relies on to take measurements of facial movements.

In other words, the cheery statement, 'there are only a few things to consider before you get going with your study' (2017: 26), sits somewhat at odds with the pages of detailed instructions about how to set up the procedure for capturing emotions. There are quite a few chances for things to 'go wrong'. We could understand these lists of instructions as a way to mitigate an underlying fragility in the system. The opportunity to receive a reliable reading of 'raw unfiltered emotional responses' (iMotions 2017: 2) is revealed as actually highly precarious in technical terms and dependent on participants fulfilling certain bodily norms and the experiment fulfilling certain technical norms. The brochure guidelines are based on bodies appearing in particular ways in order to be quite literally recognizable, and also demand a particular quality of camera and lighting in order for the image to be of high enough quality to apply the yellow dots that show the necessary facial landmarks. The iMotions software could for this reason be seen as a site of normative tensions (Grosman and Reigeluth 2019) where different and at times conflicting bodily and technical normativities are enacted and handled (2019: 10).

Let us return to the story that we opened with: the story of Pepper caring for sick children in a Toronto hospital, to illustrate what Pepper's use of emotion detection software might look like in practice. Imagine Pepper moving into the room of one of the patients. If using the iMotions software, Pepper would use its camera (which is, as earlier mentioned, located on Pepper's forehead) to scan the child's face, marking its landmarks, running the information through the software's classification algorithm that calculates which emotion(s) that the child appears to be showing. Based on this information, Pepper could adapt its behaviour accordingly. If the child appears to be sad, for example, Pepper might dance for the child in the hopes of making them smile.

As is made clear in the video clip that started out this text, this aspect of Pepper's caregiving – making children smile – is seen as crucial. Inducing smiles is put forth as a key part of the care practice carried out by Pepper. And, as we attempt to detail above, Pepper uses advanced software to be able to carry out this specific component of its care work – software which, in turn, has to divide the smile (and the associated emotion, happiness) into even smaller fragments: the landmarks on the child's face and slight movements of the corners of their eyes and mouth. Throughout the steps of emotion detection, the many separate tasks that go into calculating smiles create distance between the smile, the emotion that it is associated with, the emotional competence that emotion detection is associated with, and the care work to which this emotional competence is seen as crucial. We have already argued that this slippage, from care work to software pointing out landmarks in the face, could be understood as a form of care fragmentation where care practices are divided into separate fragments that are seen, and carried out, as separate entities (Vallès-Peris and Domènech 2020), and which has significant implications for the regulation and provision of professional care (James 1992).

In Vallès-Peris and Domènech's (2020) study of roboticists' imaginaries of care robots, they highlight the role of care fragmentation in shaping these imaginaries. One example was the division of care practices into physical care tasks and affective care tasks. While physical tasks were seen by roboticists as tasks that could be delegated to a robot, it was considered crucial that human staff remain in charge of the latter, affective tasks of care work. What we find happening in the case of Pepper in the children's hospital contributes to another dimension of robotic care fragmentation, since Pepper is, indeed, being delegated the affective elements of care work. The care fragmentation that we describe rather has to do with what goes into making that delegation possible: how the affective elements of care work are spliced up into smaller and smaller components, or tasks, in order for a robot to be able to do them.

As we suggest above, reliable data on emotional responses can only be provided for a particular subset of participants, using a carefully defined suite of technologies and within a framework of emotional classification that was generated from a subset of participants who displayed emotions in particular ways; in other words, it is contingent, that is context and participant specific. In this paradigm, value is afforded to particular emotional expressions by virtue of being able to recognize them and accurately measure them. There are numerous other faces and emotions which remain outside the classification process and thus outside valuation.

Perhaps, the normative tension of whose emotions can be read could be seen as a side effect of robotic (affective) care. As robots are programmed to carry out the emotional labour that is arguably a significant part of care work, several steps of care fragmentation are required – breaking the emotional labour into smaller and smaller units. The emotional classification system that is used by iMotions is one part of this. The measurement of facial landmarks and algorithmic classifications another. And both of these begin with the idea that the robot is providing care when producing smiles.

Hierarchies of care?

Earlier discussions of care, especially care provision in (often poorly) paid labour relations can help us to better understand Pepper's role in processes of valuing, evaluating, and devaluing care. Feminist sociology, for example, makes clear how the dependency hidden in the term care is mitigated by existing hierarchies of race, gender, class, and humanism (Star 1991; Tronto 1993; Haraway 1997; Puig de la Bellacasa 2011). And possibly there are dependencies, and hierarchies of care, hidden in some of the smiles we observe with robots (cf. DeFalco 2020). Perhaps it is not a coincidence that Pepper is *always* smiling (replicating a deferential and affective understanding of care provision), and that Pepper's software is concerned about the smiles Pepper is producing (attuned to the responses of care recipients, rather than the feelings of the care provider, thus the unchanging smile on Pepper's face).

As Puig de la Bellacasa points out, seeing the dynamics of care requires paying attention to it on the ground, in the details of practice, in the situatedness of care – something feminist STS is accustomed to doing (Puig de la Bellacasa 2011: 100). This is part of what makes us curious about the practices involved in emotion detection as well as which faces are most legible to robots. We suggest that these differences in legibility are not neutral, incidental artefacts of programming difficulties (even if programming difficulties are one layer in the onion peel of explanations surrounding them), but rather expected results from technological obduracy – the way existing technologies and their cultural genealogies impact the possibilities available for the development and deployment of new technologies. In our example, we see the legacy of Ekman's (1976) work and understanding of emotions intersecting with robotic technologies for care. The theory of basic emotions (Ekman and Friesen 1978) enables the algorithmic classification that allows iMotions software to associate a mouth's muscle movements with happiness. However, as we show above, this legacy at the same time contributes to producing categories of legible and illegible emotions, causing some smiles to count and others not.

This discussion is easily connected to Suchman's analysis of technologies of the service economy, like smart assistant interfaces or, as here, robots meant to care. Such technologies embody a 'just visible enough' worker ethos, one which is '… autonomous, on the one hand, and just what we want, on the other. We want to be surprised by our machine servants, in sum, but not displeased' (Suchman 2007: 217–220; see also Kennedy and Strengers 2020). This is in relation to the standardized expressions of emotions explored above. We want care robots to be able to understand how we are feeling, but from a limited number of ways to visually express this on our faces. Remember, from above, that only some, standard, legible expressions of emotions were selected to produce the basic knowledge about how to read faces. Caregivers are tasked with recognizing and responding to standardized care recipients's emotions, which triggers questions about the politics of servile self-erasure in care work, as Tronto did for humans, and which Suchman does in human/non-human relations; and which Pepper's moulded, unchanging smile embodies. Smiles do different things in different relations. They can erase a person's self by feigning compliance and pleasantries or produce an affected subject by expressing a reaction. But if only some reactions are legible/readable/recognizable, only those responding in that way become subjects.

Our initial analysis suggests that – through unpacking the development of iMotions software and its deployment through Pepper – some familiar hierarchies of care are re/produced or remediated in the emergence of robotic care provision. Pepper illuminates the previous 'sedimented arrangements of valuation and devaluation' (Murphy 2015: 722), and assists us in asking: which valuations have previously been put on the bodies which were able to recognize those smiles before Pepper arrived on the scene? Pepper also prompts new, more future-oriented questions, such as what value is being placed on making the robot capable of assessing a patient's smile? As a non-human who can be switched off, Pepper has no access to discussion around its own value, and it is unsettling to think that this development might lead to broader devaluation of care workers of all kinds. Relatedly, the care practices that Pepper provides (and this relates particularly to emotional labour) are devalued by virtue of being provided by a non-human.

The tension here – that Pepper reminds us that other bodies have previously been expected to do the same work – also highlights a potential critique of Pepper's expertise, however skilled Pepper or the next generation of care robots may soon be. Intimate care proffered by non-humans is often valued differently or less so than care proffered human-to-human, and the reason for this lies in 'an inbuilt and little discussed expectation/requirement of "authentic" intimacy: humanness' (Harrison 2019). Perhaps this is why the introduction of Pepper to a children's ward, and the smiles that Pepper generates, are considered newsworthy. In her article, DeFalco organizes her argument around deconstructing the connection between authenticity and humanity within the field of care, challenging the dominant paradigm in which 'a wilfully anthropocentric perspective (…) makes 'real' or 'authentic' care the exclusive domain of human animals' (DeFalco 2020: 3).

DeFalco's arguments are politically motivated by a desire to interrogate the category of 'human' upon which the care she describes is based: 'by claiming that good care is human care, one is tacitly assuming the transparency of the category human' (DeFalco 2020: 5). As such, her work helps to tease out what it means to be able/allowed to do care. Defalco's discussion on care robots further illustrates how a devaluing of care is entangled with other valuation practices. Care work, in the fragmented sense we see demonstrated by the use of Pepper that we examine here, is also being *assigned* value in that it is connected to quantified understandings of emotion as read through facial recognition technology. As particular care practices produce a visible, legible emotional response in the care receiver, they are simultaneously fragmented (identified as a composition of discrete care tasks) and valued. The production of positive emotional response is seen as an essential and unique kind of care labour that is – as is illustrated by the triggering effect that the image of a robotic caregiver can have – very important to get right.

The value of a smile – still, whose?

In this work, we have asked how the technology – in this case a recognition program embodied by a robot – is refracting social norms and values about smiling and about care. We have paid attention to whose voices and concerns are whispering in the muddle of discourses about recognizing smiles that we see in the iMotions material but also in news reports and publicity about Pepper. We are trying to articulate which concerns are not merely whispering, but are speaking clearly and loudly. We have reflected on which institutional or structural values they are conveying, what end goals are given, which expectations and hopes are expressed, and which are silenced.

Returning to the presence of Pepper in the Toronto children's ward, these questions could ask why Pepper is there, with sick children. Who has decided this is important? So important that someone wants to make a video report about it? In the brief film about Pepper in the ward, the voices (concerns, reflections, and jubilations) of hospital administrators, parents, and children are all heard talking about how useful the robot is because it is helping the sick children have fun. And of course, if developers want to make sure that children are having fun when interacting with Pepper, it is good if Pepper can see the children smile when caring for them.

Underlying this discourse is expression of a very strong norm (that makes certain things possible and desirable in an institutional setting, and which is contingent on some places and times; apparently found on the Toronto children's ward in 2020) that a child should have fun (play, dance, laugh) even when sick in hospital. Within practices influenced by that norm, one element of providing care also includes providing the opportunity for fun, for play, for smiles.

This article has examined software developed to 'read' emotions by tracing facial human movements and mapping that onto a pre-existing cartography of emotion. This software is used in care robots like Pepper, who are imagined to be able to read the emotional responses of humans for whom they are caring. This scenario of care, bound up with the unidirectional expression of emotion which is thought to be important for the robot to process, also awakens questions about the human/non-human relation and the imaginaries of care which are found in it. The iMotions brochure proffers an understanding of 'authentic' emotions not validated or recognized by human judgement, but rather scientifically rigorous 'objective' measurements performed by facial recognition software. The ability to make such measurements means the possibility of being able to program a robot to recreate them and generate 'authentic' emotional reactions.[17] We followed the journey of a smile from the face of the human, through Pepper's camera eyes, as it is fed into an algorithm and transformed into a number that makes sense to Pepper, but also to those humans who own Pepper and assess its usefulness through value metrics that are expressed in numbers rather than warm and fuzzy feelings. We have used the concept of care fragmentation to make sense of this journey of compartmentalizing and translating care practices.

And it is here we end, with a reflection on care presented by STS researchers Martin, Myers and Viseu, who write: 'A critical practice of care would insist on paying attention to the privileged position of the caring subject, wary of who has the power to care, and who or what tends to get designated the proper or improper objects of care' (Martin et al. 2015: 636). Reviewing the video from the Toronto Children's ward with which we started this article, we can see care providers in the background. Hospital administrators are being interviewed – and smiling – in scenes intercut with sick and happy children. The smiles of care recipients may be what Pepper is measuring, but the purpose of measuring them is to produce smiles (or at least reduce headaches and stress) in the administrators, policymakers, and care providers who have been tasked with caring for the bodies that are interacting with

[17] There is, however, a crack in the argument that iMotions make: their correlation between facial measurements and emotions hangs on the following phrase: 'facial expressions and emotions are closely intertwined' (iMotions brochure, p.7). Note 'closely' – not 'completely' or 'accurately', but 'closely'.

Pepper. And these are the people (job categories, structures) that are paying for Pepper and its continued development.

Thus, we leave with a final unsettled feeling, and suggest that it is important to consider not only how a smile is being valued, but who (and what structures and systems) have decided that that particular smile is valuable. Who cares about the smiles, and what is the privileged position of that caring person?

Acknowledgements: This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation.

References

Adam, Aalia. 2018. "Meet Pepper – Canada's first emotionally sensitive robot for sick kids." *Global News* 6 May. https://globalnews.ca/news/4180025/pepper-canada-robot/, accessed 21 March 2021.

Affectiva, "About us." https://www.affectiva.com/who/about-us/, accessed 31 August 2021.

Ahmed, Sara. 2004. *The Cultural Politics of Emotion*. Edinburgh: Edinburgh University Press.

Akalin, Neziha, Andrev Kiselev, Annica Kristoffersson, and Amy Loutfi. 2017. "An Evaluation Tool of the Effect of Robots in Eldercare on the Sense of Safety and Security." In *Social Robotics,* edited by A. Kheddar et al., ICSR 2017. Lecture Notes in Computer Science, vol 10652. Springer, Cham., accessed 19 May 2022.

Akalin, Neziha, Annica Kristoffersson, and Amy Loutfi. 2019. "The Influence of Feedback Type in Robot-Assisted Training." *Multimodal Technologies and Interaction* 3: 67.

Akrich, Madeleine. 1992. "The De-scription of Technical Objects." *In Shaping Technology/ Building Society, Studies in Sociotechnical Change*, edited by W.E Biker, and J. Law, 205–224. Cambridge, MA: MIT Press.

Allen, Davina. 2013. "Introduction – The Sociology of Care Work (Special Issue)." *Sociology of Health & Illness* 35(6): 928–938.

Amazon, "What is Alexa?" https://developer.amazon.com/en-US/alexa, accessed 20 September 2021.

Apple, "Siri." https://www.apple.com/siri/, accessed 20 September 2021.

Bayern, Macy. 2020. "How robots are revolutionizing healthcare." ZDNet 1 July 2021. https://www.zdnet.com/article/how-robots-are-revolutionizing-healthcare/, accessed 3 February 2021.

BBC News, "Pepper robot to work in Belgian hospitals" 14 June 2016. https://www.bbc.com/news/technology-36528253, accessed 3 February 2021.

Bowker, Geoffrey C., and Susan Leigh Star. 1999. *Sorting Things Out*. Cambridge, MA: MIT Press.

Bucchi, Massimiano. 2015. "Norms, Competition and Visibility in Contemporary Science. The Legacy of Robert K. Merton." *Journal of Classical Sociology* 15(3): 233–252.

Cockburn, Cynthia. 1983. *Brothers: Male Dominance and Technological Change*. London: Pluto Press.

Cookson, Clive. 2019. "Robot trained to be useful teaching assistant in three hours." *Financial Times* 23 October 2019. https://www.ft.com/content/5458d814-f4bd-11e9-b018-3ef8794b17c6, accessed 20 September 2021.

Davies, William. 2017. "How Are We Now? Real-Time Mood-Monitoring as Valuation." *Journal of Cultural Economy* 10(1): 34–48.

DeFalco, Amelia. 2020. "Towards a Theory of Posthuman Care. Real Humans and Caring Robots." *Body and Society* 26(3): 31–60.

DiSalvo, Carl. 2012. *Adversarial Design*. Boston, MA: MIT press.

Do, Liz. 2018. "Meet Pepper: An AI robot that will reduce wait times in hospitals." *University of Toronto Engineering* News 31 October 2018. https://news.engineering.utoronto.ca/meet-pepper-an-ai-robot-that-will-reduce-wait-times-in-hospitals/, accessed 3 February 2021.

Dror, Otniel E. 2001. "Counting the Affects: Discoursing in Numbers." *Social Research* 68(2): 357–378.

Duffy, Mignon. 2011. *Making Care Count. A Century of Gender, Race and Paid Care Work*. London: Rutgers University Press.

Dumit, Joseph. 2018. "Writing the Implosion." *Cultural Anthropology* 29(2): 344–362.

Ehrnberger, Karin. 2017. *Tillblivelser. En trasslig berättelse om design som normkritisk praktik*. [Genesis. An entangled story about design as norm critical practice] KTH PhD dissertation. Stockholm: KTH.

Ekman, Paul. 1976. *Pictures of Facial Affect*. Palo Alto, CA: Consulting Psychologists Press.

Ekman, Paul, and E. Rosenberg. 2005. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System* (FACS). Oxford: Oxford University Press.

Ekman, Paul and Wallace V. Friesen. 1978. *Facial Action Coding System*. Palo Alto, CA: Consulting Psychologists Press.

Escobar, Arturo. 2017. *Designs for the Pluriverse. Radical Interdependence, Autonomy, and the Making of Worlds*. Durham, NC: Duke University Press.

Goodwin, Charles. 1994. "Professional Vision." *American Anthropologist* 93(3): 606–633.

Gregg, Melissa and Gregory Seigworth (Eds.). 2010: *The Affect Theory Reader*. Durham, NC: Duke University Press.

Grosman, Jérémy, and Tyler Reigeluth. 2019. "Perspectives on Algorithmic Normativities: Engineer, Objects, Activities." *Big Data & Society* 6(2): 1–12.

Haraway, Donna. 1991 [1988]. "Manifesto for Cyborgs: Science, Technology and Socialist Feminism in the 1980s." In *Simians, Cyborgs, and Women: The Reinvention of Nature,* 149–183. New York: Routledge.

Haraway, Donna. 1997. *Modest_Witness@Second_Millennium.FemaleMan© Meets:OncoMouse™* New York: Routledge.

Harrison, Katherine. 2019. "Automated and/or Authentic Intimacy: What Can We Learn about Contemporary Intimacy from the Case of Ashley Madison's Bots?" *First Monday* 24(10)

Hochschild, Arlie Russell. 1983. *The Managed Heart. Commercialization of Human Feeling.* Berkeley, CA: University of California Press.

iMotions. 2017. "Facial Expression Analysis – the Complete Pocket Guide." https://imotions.com/blog/facial-expression-analysis/, accessed 1 September 2021.

iMotions, "Facial Expression Analysis." https://imotions.com/guides/facial-expression-analysis/, accessed 1 September 2021.

iMotions, "iMotions Research Software." https://imotions.com/platform/, accessed 1 September 2021

iRobot, "Roomba." https://www.irobot.se/roomba, accessed 20 September 2021.

James, Nicky. 1992. "Care = organization + physical labour + emotional labour." *Sociology of Health & Illness* 14(4): 488–509.

Joas, Hans, and Wolfgang Knöbl. 2009 [2004]. *Social Theory. Twenty Introductory Lectures.* Cambridge: Cambridge University Press.

Johnson, Ericka. 2019. *Refracting Through Medical Technologies.* London: Routledge

Joy For All, "Lifelike Robotic Pets for Seniors." https://joyforall.com/ accessed 20 September 2021.

Kennedy, Jenny, and Yolanda Strengers. 2020. *The Smart Wife. Why Siri, Alexa, and Other Smart Home Devices Need a Feminist Reboot.* Cambridge, MA: MIT Press.

Latour, Bruno. 2004. "Why Has Critique Run Out of Steam? From Matters of Fact to Matters of Concern." *Critical Inquiry* 30(2): 225–248.

Lupton, Deborah. 2016. *The Quantified Self: A Sociology of Self-Tracking.* Malden, MA: Polity.

McStay, Andrew. 2018. *Emotional AI: The Rise of Empathic Media.* London: SAGE.

Martin, Aryn, Natasha Myers, and Ana Viseu. 2015. "The Politics of Care in Technoscience." *Social Studies of Science* 45(5): 625–641.

Merton, Robert K. 1942. "The Normative Structure of Science." Republished in *The Sociology of Science. Theoretical and Empirical Investigations,* edited by Norman W. Storer, 267–280. Chicago, IL: University of Chicago Press.

Merton, Robert K. 1973. *The Sociology of Science: Theoretical and Empirical Investigations.* Chicago, IL: University of Chicago Press.

Murphy, Michelle. 2015. "Unsettling Care: Troubling Transnational Itineraries of Care in Feminist Health Practices." *Social Studies of Science* 45(5): 717–737.

Orr, Julian E. 1996. *Talking about Machine*. Cornell: Cornell University Press.

Paro, "Paro Therapeutic Robot." http://www.parorobots.com/, accessed 20 September 2021.

Parsons, Talcott. 1951. *The Social System*. New York: Free Press.

Pellegrini, Ann, and Jasbir Puar. 2009. "Affect." *Social Text* 27(3(100)): 35–38.

Picard, Rosalind. 1997. *Affective Computing*. Cambridge, MA: MIT Press.

Papoulias, Constantina, and Felicity Callard. 2010. "Biology's Gift: Interrogating the Turn to Affect." *Body & Society* 16(1): 29–56.

Puig de la Bellacasa, María. 2011. "Matters of Care in Technoscience: Assembling Neglected Things." *Social Studies of Science* 41(1): 85–106.

Rhee, Jennifer. 2018. *The Robot Imaginary*. St. Paul, MN: University of Minnesota Press.

Robertson, Jennifer. 2010. "Gendering Humanoid Robots: Robo-Sexism in Japan." *Body & Society* 16(2): 1–36.

Robertson, Jennifer. 2017. *Robo Sapiens Japanicus: Robots, Gender, Family, and the Japanese Nation*. Oakland, CA: University of California Press.

Rose, Nikolas. 2007. *The Politics of Life Itself. Biomedicine, Power, and Subjectivity in the Twenty-First Century*. Princeton, NJ: Princeton University Press.

Salmela, Tarja, Anu Valtonen, and Deborah Lupton. 2019. "The Affective Circle of Harassment and Enchantment: Reflections on the ŌURA Ring as an Intimate Research Device." *Qualitative Inquiry* 25(3): 260–270.

Søraa, Roger. 2017. "Mechanical Genders: How Do Humans Gender Robots?" *Gender, Technology and Development* 21(1–2): 99–115.

Star, Susan Leigh. 1991. "Power, Technology and the Phenomenology of Conventions: On Being Allergic to Onions." *The Sociological Review* 38(1), 26–56.

Star, Susan Leigh. 1999. "The Ethnography of Infrastructure." *American Behavioral Scientist* 43(3): 377–391.

Stark, Luke. 2019. "Affect and Emotion in Digital STS." In *Digital STS: A Field Guide for Science & Technology Studies*, edited by David Ribes and Janet Vertesi, 117–135. Princeton, NJ: Princeton University Press.

Suchman, Lucy. A. 2007. *Human-Machine Reconfigurations: Plans and Situated Actions*, 2nd edn. Cambridge: Cambridge University Press.

Suchman, Lucy, Jeanette Blomberg, Julian E. Orr, and Randall Trigg. 1999. "Reconstructing Technologies as Social Practice." *American Behavioral Scientist* 43(3): 392–408.

Tronto, Joan C. 1993. *Moral Boundaries: A Political Argument for an Ethic of Care*. New York. Routledge.

Vallès-Peris, Núria, and Miquel Domènech. 2020. "Roboticists' Imaginaries of Robots for Care: The Radical Imaginary as a Tool for an Ethical Discussion." *Engineering Studies* 12(3): 157–176.

Winner, Langdon. 1980. "Do Artefacts Have Politics?" *Daedalus.* 109(1): 121–136.

Worx, ""Is Worx Landroid" the best robot mower money can buy?" https://eu.worx.com/landroid/en/, accessed 20 September 2021.

Maria Arnelid is a PhD student in Gender Studies, Linköping University, Sweden. She has a background in social anthropology and has previously been involved in research on the automation of decision making in the public sector. Her current research deals with the development of robots for health care, focusing on what sort of care is imagined in the development of such robots. She is affiliated to the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS) graduate school, and part of the larger project 'The ethics and social consequences of AI and caring robots. Learning trust, empathy and accountability', together with Ericka Johnson and Katherine Harrison.

Ericka Johnson is Professor of Gender and Society, Linköping University, Sweden. She has an interdisciplinary background in sociology, gender studies, and science & technology studies. She is the author of several monographs and anthologies, including *A Cultural Biography of the Prostate* (MIT Press, 2021) and *Refracting through Technology* (Routledge, 2019). Together with Dr Katherine Harrison, she is currently leading an interdisciplinary research project on the ethics and social consequences of AI and caring robots.

Katherine Harrison is Senior Lecturer in Gender Studies at Linköping University. Her research sits at the intersection of science & technology studies, media studies, and feminist theory, bringing critical perspectives on normativity and knowledge production to studies of different digital technologies. Together with Prof. Ericka Johnson, she is currently leading an interdisciplinary research project on the ethics and social consequences of AI and caring robots.

Valuation Studies 9(1) 2022: 109-139

Theme issue contribution

From Actuarial to Behavioural Valuation. The impact of telematics on motor insurance

Alberto Cevolini and Elena Esposito

Abstract

Algorithmic predictions are used in insurance to assess the risk exposure of potential customers. This article examines the impact of digital tools on the field of motor insurance, where telematics devices produce data about policyholders' driving styles. The individual's resulting behavioural score is combined with their actuarial score to determine the price of the policy or additional incentives. Current experimentation is moving in the direction of proactivity: instead of waiting for a claim to arise, insurance companies engage in coaching and other interventions to mitigate risk. The article explores the potential consequences of these practices on the social function of insurance, which makes risks bearable by socialising them over a pool of insured individuals. The introduction of behavioural variables and the corresponding idea of fairness could instead isolate individuals in their exposure to risk and affect their attitude towards future initiatives.

Keywords: adverse selection; behavioural valuation; telematics motor insurance; algorithmic prediction; subsidisation; risk transfer

Alberto Cevolini is Professor of Sociology at the University of Modena and Reggio Emilia.
Elena Esposito is Professor of Sociology at Bielefeld University and the University of Bologna.

© 2022 The authors [CC BY] *This work is licensed under a Creative Commons Attribution 4.0 International License.*
https://doi.org/10.3384/VS.2001-5992.2022.9.1.109-139
Hosted by Linköping University Electronic press
http://valuationstudies.liu.se

Introduction

Predictive algorithms are now so commonplace that, for many observers, they have become "an integral part of everyday life" (Kirkpatrick 2016: 16). These algorithms process data that refer to the past, but their purpose is to manage the uncertainty that refers to the future. That uncertainty becomes particularly significant when decisions have to be made that may cover such issues as the weather, stock prices or the movement of goods and people.

Referring to people, data mining algorithms are used to rate and score individuals on the basis of such questions as: who is more likely to regularly pay the instalments of a debt? Who is more likely to reoffend if he or she is paroled? Who is more likely to efficiently perform the tasks required by a firm? Much research has already demonstrated that predictive algorithms can be extremely useful in all such cases, as they offer decision makers a presumably fail-safe way to measure the risk of backing the wrong person (Provost and Fawcett 2013). That is why predictive algorithms are already used, for example, in hiring and recruiting decisions (Miller 2015; O'Neil 2016: ch. 6), in peer-to-peer lending (Biferali 2018) and in parole procedures (Harcourt 2006).[1]

Our aim in this article is to address another sector where the possibility of using predictive algorithms to evaluate individuals is attracting much attention and, in recent years, also major investments: insurance. The issue here has traditionally been one of how to evaluate individuals to be insured in such a way that any compensation payments that the company might have to make in future do not exceed premiums received in the past. Ironically, insurance companies would prefer to insure only those individuals who do *not* need insurance, i.e. those who will not claim. Because nobody can foresee the future, however, insurance companies strive to improve their selection of prospective policyholders so that the percentage of high-risk individuals in the pool is not disproportionate. When insurance companies fail to do this, they are confronted with the problem of adverse selection. The number of bad cases is too large compared to the number of good cases, and the loss ratio (i.e. the ratio of losses to premiums earned) deteriorates. Adverse selection is often a consequence of information asymmetry (Stiglitz 1983).[2] Policyholders

---

[1] Giving rise to problems of bias and discrimination, involving both the design of the algorithms and the data used. Both issues have been discussed at length in the relevant literature. See, for example, boyd and Crawford (2012); Gitelman (2013); Gillespie (2014); Pasquale (2015); O'Neil (2016).

[2] Literature on adverse selection is very great. An important contribution on this topic is Baker (2003), where adverse selection is investigated as a "dual problem" (selection can be adverse either to the insurance company or to the prospective policyholders) with pooling and de-pooling effects.

usually have more information about their actual exposure to danger than they are willing to disclose to an insurance company – either because they fear they would have to pay a higher premium, or because they fear the company could refuse coverage. This asymmetry also affects competitiveness: companies which better reduce information asymmetry, that is, whose percentage of high-risk individuals is not disproportionate, are better placed on the market. For insurance companies it is therefore extremely important to get as much information as possible, in order to improve their evaluation of prospective customers and to minimise adverse selection.

This is one of the reasons why they are adopting new forms of valuation, using algorithmic procedures to estimate the risk probability of individuals according to their actual behaviour. By means of monitoring and tracking devices, insurance companies can resort not only to more information but also to different information, namely, information about individual behaviour. This form of *behavioural valuation* promises to reverse the asymmetry: insurance companies might have more information than the customers themselves about their actual risk profile (Cevolini and Esposito 2020).

Recent techniques based on machine learning and using big data promise today to deliver reliable targeted predictions that refer to single individuals (Domingos 2015; Siegel 2016).[3] This possibility, if it could be implemented in practice, would be particularly well suited to guiding the valuation and selection routines of insurance companies. Digital technologies are expected to impact not only on pricing, but also on the entire value chain of the insurance industry (Eckert and Osterrieder 2020; Eling and Kraft 2020; Eling and Lehmann 2020), with potentially disruptive effects (Boobier 2016; Braun and Schreiber 2017; Albrecher et al. 2019). However, it has been observed that academic research on digitisation in the insurance industry is still quite scarce (Cather 2018; Eling and Lehmann 2020). Our investigation addresses current experimentation in this field, also with the goal of contributing to plugging the current gap in academic research.

We focus on the sector of third-party liability motor insurance because it is one where behaviour-based pricing is most advanced, in terms of both technological experimentation and concrete insurance practices. While giving due consideration to the huge and varied panorama of telematics third-party liability policies currently available worldwide, we focus in particular on the telematics policies offered by a selection of Italian insurance companies. The reason for this choice is that Italy was an early adopter of digital technologies for evaluating driving behaviour and is also still a cutting-edge country in developing

[3] For a critical analysis of the many controversial sides of these technologies, see, for example, Rona-Tas (2020).

motor insurance telematics from a legal and regulatory standpoint (Dang 2017).

The topic is very complex and, in part, also controversial (McFall et al. 2020). Some researchers have expressed doubts about what is often presented as one of the most revolutionary innovations in the digital valuation of insured parties, i.e. the possibility of customising policy premium. For Liz McFall and Liz Moor (2018: 205), it is too early to say that premiums are "tailored" to individual policyholders in connected insurance. This would mean that the feared hyper-individualisation of premiums (Billot et al. 2018) leading to a drastic revision of the traditional models of risk-sharing is still no more than a hypothesis, rather than reality in insurance practice. For Maiju Tanninen (2020: 8), empirical research shows that the utopian (or dystopian?) idea of personalised insurance "is not very easy to achieve", even though one cannot rule out that behaviour-based insurance is able to explore "alternative imaginings" for adapting policy premiums to individual risk exposure. In a recent study of telematics motor insurance, Laurence Barry and Arthur Charpentier (2020) reached the conclusion that, with regard to pricing and tarification practices based on assessment of the policyholder's exposure to the risk insured, "nothing revolutionary has happened yet" (7) and the expected disruption of the insurance industry "actually did not happen" (6).

Our research explores and questions these conclusions. Is there really nothing new in the use of digital technologies to assess risk exposure in motor insurance? Barry and Charpentier (2020: 7) themselves admit that the addition of behavioural variables to classical actuarial statistical variables "is itself *radically new*" (emphasis added). Our study further investigates this novelty. We use the current literature about telematics motor insurance as a starting point for empirical research in which we set out to understand what is really happening in insurance companies that sell telematics products,[4] with possible consequences also on the social role of insurance. In this article we present the first results of this research. We explore three different aspects of the impact of telematics-based techniques on insurance: challenges to business models and calculations of insurance companies (in the third section), challenges to the relationship between insurance providers and policyholders (fourth section), issues about fairness and discrimination (fifth section).

Looking inside an insurance company is notoriously difficult. To gather empirical material, we conducted semi-structured interviews

[4] Research conducted by Barry and Charpentier is based on a review of articles published about UBI and telematics motor insurance over the last decade, presented as "an exploratory analysis of documents" (6). The two researchers are aware of the fact that the articles they considered do not necessarily "reflect the actual practices of all insurers" (9).

using two questionnaires: one, more sociological, was aimed primarily at companies offering telematics third-party liability motor insurance; the other, more technical and actuarial, was aimed at the software providers, data scientists and mathematicians who analyse the data and the problems deriving from extracting information for predictive purposes. We interviewed three executives of Italian insurance companies, one actuarial mathematician of an Italian insurance company, three heads of data analytics of international software providers. An Italian executive with past experience as an actuarial mathematician answered both questionnaires.

The names of the interviewees and of the companies were anonymised by using capital letters for the companies and numbers for the interviewees (interview A.1, B.1 … and A.2, B.2 … when two or more people of the same company were interviewed). Because interviews were conducted between autumn 2020 and spring 2021, during the pandemic emergency caused by COVID-19, we opted for digital meetings. The interviews, which lasted 60 to 90 minutes, were conducted either in Italian or English by one of the authors, were recorded, transcribed and then carefully analysed and cross-commented by both authors.

Our interviews cover three of the four insurance companies which offer a telematics policy.[5] Respondents showed a striking uniformity in jargon, core issues, technical problems and digital solutions. The differences lie rather in the amount of data available and the purpose for which it is used. Apparently no one has the Coca Cola formula in its safe. Our interviewees also shared their expertise far beyond the concrete limits of their company practice, covering their knowledge of competitors and of the current experimentation in the field.

The results we have derived from our interviews so far confirm that the practice does not correspond to the often very emphatic narrative that accompanies digitalisation of the insurance industry.[6] Rather, our investigation of actual practices enabled us to identify some more subtle issues and important changes in the motor insurance sector, from which indications can be drawn for the evolution of insurance in general. The trends we observe might have an impact on the basic assumptions of insurance and on its social significance, thus introducing genuine though less emphatic novelties.

In this article, we present some of these issues and changes. At this stage we only explore the perspectives of insurers, software providers and data scientists, trying to get an insight into current innovations and their relevance to the field. Our knowledge of policyholders and

---

[5] Unfortunately, the company with the highest number of telematics policies declined to answer our questionnaires.

[6] Barry and Charpentier (2020: 2) talk about "myths […] associated to predictive analytics".

their attitude relies only on the way insurers observe subscribers of telematics third-party liability motor insurance policies. Further research might fill this gap, addressing directly a user's perspective.

In order to understand the impact of digital technologies on the assessment of policyholders' risk exposure, we first have to clarify how the insurance industry implements its risk management mechanism.[7] We present this preliminary clarification in the following section.

Subsidisation vs risk transfer

Insurance is a form of socialisation of risks, which generates specific financial solidarity between policyholders. The mechanism on which insurance business is based is usually described as "risk pooling and spreading". The underlying idea, which reproduces the primary insight of probabilistic calculation, is that while in terms of the individual the risk is in principle unpredictable, in terms of the group the aggregation of many individual risks generates reliable regularities that provide a basis for calculation. By pooling risks, in other terms, "*accidents* become *normal,* and in that sense not accidents at all" (Ericson et al. 2003: 47, emphasis added). So statistics offer the insurance industry a sort of "secondary normality" (Luhmann 1991: 1) that can be managed by means of calculation and the availability of sufficient data, giving the impression of controlling the unpredictability of the future.

For the insurance industry, risk pooling is an essential tool for transforming an unmanageable risk into a bearable one. From the viewpoint of an insurance company, the advantage of accepting a multiplicity of similar cases lies in the possibility of using good risks to compensate for bad ones: those who pay the premium but do not claim should offset[8] those who pay the premium but suffer damage (e.g. a car accident) for which they can then claim compensation. All these policyholders are members of a pool, prepared to pay an often fruitless premium in exchange for the certainty of compensation if they suffer the damage in question. That is how the financial risk is spread between all policyholders.

The risk is compensated by sharing it not only socially, between various policyholders, but also temporally, between various moments in time (Albrecht 1992; Farny 1995). As a result, premiums accumulated in the course of time are used to reimburse the unluckier cases when they occur. The uncertainty at play thus concerns not only *who* may be affected by misfortune, but also *when*. These two

---

[7] In this article we focus on ongoing insurance practice concerning actuarial valuation. Alternative uses of digital technologies in the insurance industry are of course possible.

[8] Or, "ristorano", as Benedetto Cotrugli already expressed it in the mid-fifteenth century (1602 [orig. ed. 1458]: 75).

dimensions of uncertainty – the social and the temporal – interact with each other, enormously increasing the complexity of the situation that the insurer has to tackle. The ingenious thing about the insurance mechanism is that it transforms this problem into a solution: pooling risks is a way of spreading risks between the members of the pool over the course of time, thus guaranteeing financial solidarity between all those who decide to take part by paying the premium. In this sense, "insurance is the paradigmatic risk-spreading institution" of modern society (Baker and Simon 2002b: 7). That said, however, solidarity between policyholders can take two different forms, with different social implications: *subsidiarity*, or *risk transfer*. We deal with them in this section of the paper.

If we follow the explanation of the insurance mechanism presented above, it must be admitted that policyholders who pay their insurance premiums are not paying it *for themselves*. Instead, by paying their premiums, they join what could be described as a "secondary collectivity", whose members are all those who have decided to pool their fate, in financial terms and with certain conditions. For the company, the only important thing is that, at the end of the predetermined period, total losses have been covered by the premiums collected from all policyholders who joined the group. In other words, there must be an "equivalence" between what the company collected in advance on the basis of its own estimates about future claims (in terms not just of frequency, but also of severity) and what the company will have actually paid in terms of compensation when the future will have become the past (Mahr 1951; Farny 1992). If the principle that governs calculation of the pure premium (i.e. the spreading of risk) is that of homogeneity, then the financial solidarity created among policyholders is termed *subsidiarity* (Thiery and Van Schoubroeck 2006; Lehtonen and Liukko 2011). What drives individuals to pool their fates under these conditions is essentially the uncertainty about who will be stricken by adverse fortune.

However, insurance companies know very well that not all individuals are exposed to the same degree of risk. In third-party liability motor insurance, for example, male drivers are statistically more exposed than females (whatever the reason may be) and those who have just passed their driving test are usually more exposed than mature drivers (presumably because they are less expert). It is therefore in the company's interest to introduce statistical differences, i.e. such variables as the driver's gender, age and place of residence, the type of car driven and so on, that enable segments to be created, with which a more accurate estimate of each segment's degree of risk exposure can be associated, so that the company can reach a more precise estimate of its corresponding expected loss. The company then employs this *segmentation* to elaborate a differentiated tarification that is updated continuously on the basis of historical claim data.

The most extreme form of solidarity emerging from *segmentation* has also been defined “chance solidarity”.[9] By that, scholars mean that all members of a particular segment share exactly the same probability of filing a claim, and who will be the misfortunate one is simply decided by chance. Risk is spread among members of the same segment but not between members of different segments. From the viewpoint of a segment, this type of solidarity can be regarded as fair. From the viewpoint of a subsidising society, this type of solidarity can instead be regarded as unwillingness to help the underprivileged (Lehtonen and Liukko 2011).

The underlying principle of segmentation favours a different conception of the insurance mechanism, closer to the concept of *risk transfer.* The basic idea here is that every policyholder (or every predefined group of policyholders) transfers a particular risk to the insurance company and that, as a consequence, it is right that every policyholder (or, every predefined group of policyholders) pays in proportion to the risk thus transferred.[10]

An insurance company that opts for subsidiary solidarity misses out on the opportunity of being more competitive. By drawing distinctions between the different extents to which members of the pool are exposed to risk, the company could create separate segments and calculate a premium corresponding to each of them. Those who are less exposed to the risk would pay less and would thus be more attracted by the policies offered by the company practising segmentation than by the one that opts for subsidiarity.[11] The

[9] See the seminal paper by De Wit and Van Eeghen (1984). Cf. also Lehtonen and Liukko (2011); Barry and Charpentier (2020).

[10] Both subsidiarity and risk transfer have pro- and contra moral justification (Baker 2003). Subsidiarity – simple risk pooling and spreading practised without the introduction of any particular differences – apparently has the advantage of eliminating all forms of direct discrimination: young drivers pay exactly the same as seniors, women pay the same as men, etc. But it also has many defects. It may be considered unfair by those policyholders who take more precautions, or who are normally more prudent, or in general believe that they are less exposed to risk. In addition to this, it is not in the company’s interest to neglect these differences, for reasons that are both technical and economic.

[11] In addition, the insurance company could reduce adverse selection and make more accurate estimates about expected losses in each of the individual segments. When segmentation is practised on the basis of differences correlated statistically to claims causing an actual loss for the company, the use of these differences is considered legitimate by Italian law on the basis of the argument that renouncing this correlation would undermine the actuarial structure of insurance business and, in extreme cases, would make coverage itself impossible.

terminology used to describe this kind of discrimination is "actuarial fairness".[12]

Sociological research has demonstrated the existence of a trend in the last century towards growing opposition to subsidisation and an increasing preference for segmentation. Around the middle of the twentieth century, this trend triggered a sort of "spiral of segmentation", inspired by the desire to adapt the premium as far as possible to individual hazard (Barry 2020: 175). Of two distinct forms of self-understanding of the insurance mechanism – insurance as a form of subsidisation, and insurance as a form of risk transfer – the latter gradually took hold, generating an increasingly marked subsidy-aversion argument. It could be said that the individual's point of view has taken over progressively from the point of view of the group. The development of behavioural tariffs based on telematic data in the third-party liability sector of motor insurance, which we shall discuss in the next section, can be seen as part of this trend toward individualisation. However, behavioural individualisation cannot be reduced to the principle of increasing segmentation leading to ever narrower groups. For the insurance company the matter is rather to identify which individuals belonging to a certain segment will probably perform better and which individuals will probably perform worse.

Digital valuation of the degree to which policyholders are exposed to the risk indeed offers a unique opportunity to boost the practice of insurance as a form of risk transfer, with a series of consequences that deserve investigation.[13] This approach is also supported intuitively by policyholders. No-one is keen on paying for other people's imprudence. And those who behave particularly prudently would like their prudence to be acknowledged (here, we could also say "rewarded") by the insurance company in the form of suitable reductions in their premiums. Until the end of the twentieth century, however, the technology necessary for effectively measuring the degree of the individual's exposure to risk was still lacking. That is why, at the end of the 1990s, the economic theory of insurance still considered

---

[12] We shall not go into the debate about "fairness" here, since it is extremely complex and varied and would call for a separate study. However, we will come back to the problem of discrimination in the final section of this article.

[13] There is actually nothing new about this idea from a legal standpoint. Early modern legal doctrine already hypothesised the contract of insurance as a very special form of aleatory contract in which the insurer accepts a financial risk (known as the *susceptio periculi)* in exchange for a premium. The premium guarantees that the pact is binding, but before accepting the obligation, the insurer obviously wants to know what risk they are running in reality. Jurists pointed out that the contract's equity depends on the relationship between price and risk, so that the premium should not vary *arithmetically,* but *geometrically,* in other words not in absolute terms, but in proportion to the risk to which the policyholder is actually exposed (Oñate 1654, Tract. 36, Disp. 131, Sect. II, n. 16: 677 f.).

irrelevant the question of whether the principle of subsidisation was generally fair, since the only way to make the insurance mechanism function on the *collective* plane was, as we have seen, to ensure the existence of an equivalence between the total amount of premiums received and the total losses expected by the insurance company (Farny 1995; Innami 1996). That the principle of equivalence led to a different conclusion on an *individual* plane obviously escaped nobody's notice, but since there was no way of directly measuring the risk transferred to the insurance company by each individual, the question was simply ignored (Farny 1995).

From the beginning of the new century, the use of telematics technologies (such as the installation of black boxes in vehicles), together with digital devices (such as mobile phones), has created unprecedented possibilities for monitoring individuals' driving behaviour and, as a consequence, modifying their insurance premiums to suit their real exposure to risk.[14] In our research, we investigate how this is done in practice and what opportunities and problems it generates in the relationship between insurance company and policyholder.

The introduction of behavioural variables into actuarial calculations

In order to understand how telemetry-based technologies impact on the valuation of policyholders, we interviewed insurance companies with extensive experience in this sector, because they were some of the first to sell third-party liability motor insurance based on the detection of drivers' behaviour. The telematics programs currently in use are typically formulated as pay-as-you-drive (PAYD) and pay-how-you-drive (PHYD) policies: in the first case, what is evaluated is the number of kilometres driven by policyholders over a certain period of time (e.g. a month), while in the second case it is their "driving style", for example whether they comply with speed limits, whether they drive by day or by night, whether they swerve or brake brusquely, whether they drive on city streets or country roads and so on. This information is fed back to the driver through an app on the policyholder's mobile phone. The telematics app conveys both granular and aggregated information. Every trip is detected and recorded, and criticalities (e.g.

[14] In the more innovative forms of mobile telematics, behavioural data are not produced by a black box installed in the vehicle (although this is still the practice among the Italian insurance companies that we interviewed), but by a mobile phone paired with a smart tag fixed to the vehicle's windscreen. While use of the mobile phone generates technical problems that do not apply in the case of the black box, the accuracy of the behavioural data it produces is now 90–95% compared to that of the behavioural data detected by a black box associated with the vehicle (Interview E.1, 30 March 2021). The behavioural variables detected are essentially the same in both cases.

excess of speeding, phone distraction, etc.) are visualised on the app after every trip. In addition, the policyholder receives an overall risk score which is usually updated monthly and measures the estimated dangerousness of the individual driving style. The telematics policies currently on offer on the Italian market combine both pay-as-you-drive and pay-how-you-drive features. In practice, this comprises the introduction of *behavioural variables* into actuarial calculations.

Professionals we interviewed confirmed that the first step they have to take to conduct this assessment is the creation of clusters of policyholders who are differentiated on the basis of classical actuarial variables: driver's gender and age, type of vehicle driven, driver's previous claim history (on which the established mechanism of the no-claims bonus is based) and place of residence. The historical data available enables a preliminary assessment to be made of the statistical probability that a given policyholder (e.g. a man aged 40 who lives in Milan, drives a Fiat Panda and has never had an accident) will file a claim in the following year. This operation is the crucial foundation for a first tarification of policyholders differentiated by segments according to standard actuarial methods.

The second step is to build on the traditional actuarial valuation by adding an assessment of risk exposure based on behavioural data. Our interviews reveal that behavioural data are essential for increasing the *variance* explained, i.e. for understanding why some of the individuals who belong to a group have a lower-than-average estimated probability, while others have a higher-than-average estimated probability. To go back to the example above: if the man aged 40 who lives in Milan, drives a Fiat Panda and has never had an accident has a 10% statistical probability of filing a claim in the year after he has bought the insurance policy, the behavioural valuation will help the company improve its estimate of probability associated with the single individuals belonging to the segment in question. In the opinion of one of our interviewees "one thing is certain: the delta risk factors[15] detected by the technology applied to insurance are objective and statistically measurable" (Interview C.2, 25 March 2021).

According to the metaphor used by another interviewee, we need to

> imagine this kind of exercise as something like a Rubrik's cube: on the first part of the face is the human world … I go to take a look at how [the policyholder] behaved last year from the point of view of my portfolio, from the point of view of my customer basket, so as to use probability to calculate how things will look next year. What we have here, then, is *a concept of pure*

[15] Delta risk factors are variables (in the case of motor insurance, e.g., harsh braking and sudden tailgating) which can affect the average risk exposure of individuals belonging to a certain pool. Aggressive driving style, for example, can cause a deviation from the average, making a difference in risk assessment that has high information value for the insurance company.

> *statistics*. On top, [you always have to] imagine then how the cube will look on the next side … [Here] I look at all the *parameters of the telematic data* and, as a consequence, all those data that I cannot see,[16] which are not the same as those I can download from an estimate, or that I can download from a databank that may be provided by the Ministry of Transport, or by the National Institute of Statistics (ISTAT), or again by the vehicle drivers' portal, so everything that I cannot see – and that's the boost that I get from telematics – that is to say that it manages to give me real data … The last side of the cube is everything that we end up creating with the *models of artificial intelligence*. (Interview F.1, 9 April 2021)[17]

This last phase is crucial. As we have seen, the company has a vital interest in guaranteeing that the total sum of the pure premiums received in advance is equivalent to the total sum of the compensation that the company will then have to pay out during the period stipulated as the policy's duration. This means that the claim is more relevant than the crash: that is because drivers do not file a claim for every crash, or because the severity of the crash is underneath their deductible/excess, or because drivers want to avoid being penalised by losing or worsening their no-claims bonus when the time comes to renew the policy.

So when a company has access to behavioural data, it starts from the filing of a claim in which its customer is at fault. This is the point at which data scientists step in. One of them explained to us that

> we take information from our own data to determine when a crash occurred to then get a view of what behaviour occurred immediately prior to a crash … It enables us to build *a view of risk* in a way that is *directly correlated back to those drivers' actual behaviours*. (Interview E.1, 30 March 2021)

The result is a predictive model: by monitoring driver behaviour for 91 days, the algorithmic data processing can "predict with a high degree of certainty what behaviours are likely to lead to a crash in the next 90 days" (Interview E.1, 30 March 2021). This predictive model refers to the actual driving behaviour and not only, as traditional actuarial models, to the claims that have been filed. One of the interviewees

---

[16] He cannot see the data referred to single individuals because the statistics "obliterate the individual" (Daston 1983: 23) and instead offer up "averages" derived from the aggregation of many sets of individual data.

[17] The meaning of Rubrik's cube metaphor is, in our opinion, that the addition of behavioural data to usual statistical variables should not be understood as a mere additive process. Like the real world, Rubrik's cube is a three-dimensional reality whose faces can be continuously recombined to obtain information that would not be available if the observer merely looked at one face of the cube at a time under predetermined conditions.

explained this difference by stating the need to distinguish between an "actuarial score" and a "behavioural score".

> An actuarial score predicts claim frequency, but a behavioural score is … basically looking at those risk behaviours that … contribute towards loss, but are not tied back to the insurance process. (Interview E.1, 30 March 2021)

The first element, the actuarial score, is the one traditionally used by insurance companies for the purpose of guaranteeing the profitability of their business. The second element, the behavioural score, measures an additional risk factor known as "driver aggressiveness", so the prudence or the lack of prudence practised when driving. This score is an aggregated result of telematics data processing and is usually layered into three profiles: low risk exposure (advanced driver), moderate risk exposure (normal driver), high risk exposure (reckless driver). An interviewee pointed out that, even if "there's a very big overlap between the two elements, because the elements that lead to safety and elements that create claims are … incredibly highly correlated" (Interview E.1, 30 March 2021), the actuarial score and the behavioural score are not identical and the difference is conceptually very important.

For insurance companies, the behavioural score provides a more accurate forecast of which customers are more likely to get into a crash irrespective of whether they will file a claim. The argument goes: if you can reward or penalise members of the same segment more selectively based on their actual behaviour, you can retain the better customers, while reducing the risk of churning and improving the loss ratio. This means that insurance companies can adopt commercial policies whose ultimate purpose is to practise "cream-skimming" (Cather 2018). From the insurance companies' standpoint, telematics thus constitutes an unprecedented opportunity for managing the classical problem of adverse selection. For one of our interviewees "the great advantage of behavioural tarification is that it provides an objective, rational and structural way to industrialise discounts that maximise customer retention" (Interview C.2, 25 March 2021). And for another interviewee "that is the real keystone to the entire system" (Interview F.1, 9 April 2021).

On the one hand, then, Barry and Charpentier are right: the aim of using digital technologies in third-party liability motor insurance is to "*refine* the existing segmentation thanks to new parameters" (Barry and Charpentier 2020: 8 emphasis in original).[18] Seen from another

---

[18] And without going to the extreme case of the "pool of one", which is often discussed in the literature with a degree of concern (Ramasastry 2012; Harrington 2017; McFall et al. 2020).

perspective, however, the detection of behavioural variables leads to a truly innovative evolution in the form of forecasting used by insurance companies. Digital valuation searches out "behavioural patterns" that enable it to explain why certain individuals within the average of a given segment can be said to perform better, while others perform worse. The aim of behavioural valuation is thus to help the insurance company identify those individuals within the given segment who for behavioural reasons are more likely to file a claim. One of our interviewees points out that "what we're trying to do in our industry … is identify those behaviours that are *causative* of risk, that are *controllabl*e … and are *predictive*" (Interview E.1, 30 March 2021).

The company rewards or penalises specific behaviours, adapting the tariff or discount when the policy is renewed. This paves the way for us to return to a point we mentioned in the introduction: the customisation of the premium. If by customisation we mean a tarification based entirely on each individual's personal data, then we can certainly say that digital insurance premiums are not customised. As we have seen, the basis for tarification still remains the classical actuarial model, based on variables *independent* of individual behaviour, such as age and gender.

If, on the other hand, we take customisation to mean an adaptation of the premium to take into account the individual's actual risk exposure, we can say that insurance companies have started experimenting with "tailor-made" premiums, based on monitoring the policyholders' behaviour. People belonging to the same segment may pay less or more (because they do or do not get a discount) according to their actual score. People belonging to the *same* segment, in other words, can pay *different* premiums.

Confirmation of this comes from the Italian insurance professionals we interviewed, who told us that companies are moving over from a logic of "discount upon renewal" (a simple commercial leverage that is applied to everyone when the telematics policy is renewed) to a more complex logic of "price upon renewal", in which behavioural data are "built into the actuarial architecture" (Interview C.2, 25 March 2021) to customise the premium according to specific predictions for single individuals.

The impact of telematics on communication with policyholders

Shifting our focus from insurance companies to individuals, we next asked insurance professionals how they perceive the response of policyholders to the change in method used to assess their exposure to risk, and what consequences they expect this change could have on the relationship between policyholders and insurance companies.

The interviews we conducted with Italian insurance companies illustrate a situation that is significantly different from the one that is normally portrayed, with a degree of concern being expressed in the literature, above all with regard to the issue of privacy. One company that decided to do away with telematics policies told us that the people who bought their telematics policies failed to download the app that was supposed to collect the data and then transmit them to the company (Interview D.1, 03 November 2020). The same problem was illustrated very clearly by another company that has been using telematics motor insurance policies since the beginning of the century. Speaking about the situation in Italy, one interviewee explained that those who paid for a telematics insurance policy did it "essentially because of the discount, rather than … for the opportunity to understand their [driving] behaviour" (Interview B.1, 24 March 2021). That is because Italian law obliges insurance companies that collect telematics data in their third-party liability motor insurance to give policyholders an entry discount, which is sometimes termed "welcome bonus", or "flat discount". When the policy is renewed in due course, a further assessment based on kilometres driven and customer's driving style is then taken into consideration to decide a possible future discount.

The same person (Interview B.1, 24 March 2021) told us that the flat discount can account for as much as 25% off the standard tariff, constituting a strong incentive for drivers who want to save on their car insurance. This is also confirmed by the fact that a preference for telematics policies was first encountered above all in those regions of Italy (in the centre and south of the country) where tariffs are higher because of the greater frequency of accidents. Telematics car insurance policies are thus said to have achieved a greater penetration in central and southern Italy for the perfectly predictable reason that, in those places where the premiums are higher, people have a greater incentive to save money. Yet, this also means that concern about the use of data is given less importance and remains marginal when compared to the economic advantage that can be achieved from a discount on the insurance premium. Our interviewee believes that the behavioural analysis

> was evidently not highlighted very much at the moment when the policy was sold, so we had a degree of … I don't want to say conflict, but anyway some resistance to understanding and managing the product when the time came for renewal, maybe because the customers who had bought the policy had done so essentially for the discount rather than for, shall we say, the possibility to get to know their own behaviour. When they found maybe that, on renewal, their discount was not so generous as the one they had enjoyed in the previous year, because they had driven more, or because they had been driving in ways that were not exactly safe, this generated a basic need for management also in the point of sale. (Interview B.1, 24 March 2021)

The message conveyed in this and other interviews seems to be that, when the first third-party liability motor insurance policies based on a behavioural valuation of the driver were introduced in Italy, there was none of what could be described as "policyholder education in telematics".

Another interviewee explained to us, in a very disenchanted tone, that when the insurance company first launched telematics third-party liability motor insurance policies at the beginning of the century, companies still had the habit of sending text messages to their customers' mobile telephones to update them about the status of their tariff. This practice was sometimes perceived as a form of intrusion and generated concerns about privacy. In 2021, on the other hand,

> to talk about problems of privacy because your car knows where you are is a bit silly, because these days everything, starting with your smartphone and continuing with your TV and your smart speakers … that is to say, everything knows where you are and what are doing, doesn't it? So for me that has now become an issue … *that's a false problem, it's more of a leftover from the past.* (Interview C.2, 25 March 2021)

That obviously does not mean that privacy is not a significant issue, both legally and ethically. But insurance executives perceive it as a false problem in practical terms because habituation to digital technologies gives the impression that privacy is downright impossible.

One unprecedented possibility that the digital valuation of the policyholder's driving behaviour offers today, on the other hand, is that of establishing a circular communication relationship between the insurance company and its customers. In this case, the combination of telematics technology with the mobile phone turns out to be crucial. The traditional insurance company would focus most of its dialogue with its customers on the phase leading up to the signature of the policy, employing questionnaires to gather information that it would then use to place the customer in a given actuarial class. Further communication was triggered later only by claims or at the renewal of the policy (which also included the risk of losing the costumer). In the more evolved version of motor telematics, on the other hand, policyholders receive targeted information about each single trip on an app, together with tips about any critical issues (e.g. when and where the customer broke the speed limit, swerved sharply or made a forbidden U-turn). At the same time, policyholders can call up their total score and the discount from which they can benefit if and when the time comes to renew their policies.

According to our interviewees, this technical possibility of feeding information back to the driver enables policyholders to know in real time how their driving style is evaluated by the insurance company. A computer scientist told us

> that's where we find the mobile program to be so unique … *mobile provides that feedback loop to drivers*, that's unique to that element and is temporal, and then you can layer in incentives like reward programmes, like changing premiums every 6 months or every 12 months, and you can show visually how individual driving behaviour leads to that. (Interview E.1, 30 March 2021)

From the insurance companies' and software providers' viewpoint, the purpose is to improve not only drivers actuarial scores, but also their behavioural scores – a significant step towards the ideal that underlies all road safety programmes, i.e. zero road accidents.

We wondered to what extent these ideals correspond to reality. The insurance companies we interviewed told us that historical data collected in the most recent telematics programs, ones that have only been active for a few years, do not yet enable statistically relevant conclusions to be drawn. One software provider realised that

> there is the Hawthorne effect that says that, you know, someone can change their behaviour for a certain given period of time while being monitored, but that Hawthorne effect fades over time. (Interview E.1, 30 March 2021)

Nevertheless, one insurance company detected a slight drop in the frequency of incidents, although not of their severity (Interview B.1, 24 March 2021). For that company, it was still early to say whether the telematics group can be distinguished in any statistically significant way from the non-telematics group. Another company, which classifies driving behaviour in three different buckets – more "evolved" drivers, "intermediate" drivers and "reckless" drivers – noticed a decrease in the level of risk. In the short and medium terms

> if we assign a value of 100 to the people who are in one segment rather than another, the great majority of customers in the higher-risk or medium-risk segments decrease towards medium-risk and low-risk classes. (Interview C.2, 25 March 2021)

This effect can be measured statistically, although it is hard to understand to what extent it is due to the data fed back to drivers. The interviewee himself actually pointed out that it is impossible to say with certainty

> whether the effect is related to education, to being afraid of paying more for the policy or to the fact that customers are annoyed when they find

> themselves assigned to the bad class; it is probably a combination of all these things. But the effect is visible. (Interview C.2, 25 March 2021)

In the opinion of many of our interviewees, in any case, "incentives are incredibly important" (Interview E.1, 30 March 2021).

> Just knowing how you drive, without maybe receiving some little gadget as a reward for virtuous behaviour, turns out to be, shall we say, a strategy that is rather incomplete and limping. (Interview B.1, 24 March 2021)

Many telematics programs that have been adopted in countries outside Italy do in fact make allowances not only for discounts weighted by parameters, but also for a series of incentives that range from little gadgets to cashback when filling the tank, vouchers for buying trips and so on. These incentives have been found – with some caution – to be of fundamental importance for reducing the incidence of road fatalities (Stevenson et al. 2018; Peer et al. 2020; Stevenson et al. 2021).

In any case, one significant element of this new relationship of communication between the insurance company and policyholders is that the insurance company can, for the first time, embark on "educational" or "coaching" actions that tend towards being *proactive:* instead of waiting for accidents to happen, as is traditionally the case, the insurance company adopts strategies that are designed to call attention to risks and as far as possible mitigate policyholders' exposure to danger. This would mean a profound transformation in the insurance company's mission and business model.

This already happens in mobile telematics, because drivers are well aware that allowing themselves to be distracted by their mobile telephones will be monitored by the app, something that does not happen with the black box. As being distracted by the telephone is the third cause of road accidents, dissuading the use of mobile telephones while driving is already a way of being proactive. But the intention is to intensify operations of this kind enormously in years to come. One of our interviewees told us that he and his team

> are grading drivers on their safety and providing information back to them. And our goal is that they *internalise that information* and become safer themselves; (Interview E.1, 30 March 2021)

while another interviewee stated: "This is our plan for 2021 in terms of products: *coaching*" (Interview H.1, 31 March 2021).

Discrimination and fairness in behavioural valuation

This last section of our article deals with a question of fundamental significance for the sociological analysis of insurance and for our investigation of the innovative potential of recent digital technologies: the impact that the valuation of customers with a telematics motor insurance policy could have on the risk of discrimination and on the fairness of the insurance mechanism in general. What emerges from the interviews we have conducted so far is that insurance companies are aware of the fact that the use of behavioural data to evaluate their customers can raise unprecedented questions about the social impact of their procedures. One of our interviewees told us about his uncertainty whether "it is not so much the insurance companies that are not ready, but more the consumer who is not ready" (Interview D.1, 3 November 2020).

Under the heading of "discrimination", the problem of the propriety of procedures of algorithmic evaluation is often tackled as part of the complex, sophisticated ethical and legal debate about algorithmic fairness, the idea of equality in insurance practices and corresponding expectations of algorithmic accountability. Twenty years ago Tom Baker (2003: 275) had already argued that "any particular individual is only one technological innovation away from losing his or her privileged status" of low-risk case, and that in principle new risk classification systems could penalise people previously ranked as "good customers".[19] Moreover, many are of the opinion that achieving a balance between fairness, equity and non-discriminatory practices in the insurance industry is an extremely hard challenge to meet. The problem is technical–actuarial, legal, ethical and political, all at the same time (Lehtonen and Liukko 2015). It is even impossible to provide an unequivocal answer to the question "How fair is actuarial fairness?", since equity does not always imply the absence of discrimination, nor can it be said that everything that is non-discriminatory is necessarily fair. The notion of fairness is complex and multidimensional (Minty 2021). Rather than providing more input to this debate, our intention here is to focus on a more circumscribed question: is the use of behavioural data discriminatory – or discriminatory in a way that is different from traditional insurance practices?

We start from a specific case. In 2004, the European Union issued a Directive, better known as the "Gender Directive" (2004/113/EC), to govern the terrain of unisex tariffs. The directive prohibited the use of gender to differentiate insurance tariffs, considering reference to the difference between males and females to be a discriminatory practice (Art. 5). Nevertheless, the directive contained a get-out clause that

[19] Baker referred to genetic testing in health and life insurance.

legitimised the use of gender difference when insurance companies could demonstrate that it was correlated statistically to the company's expected losses on the basis of a probabilistic assessment conducted on claims filed in the past. This exception was then challenged by the Belgian consumers' association Test-Achats. In 2011, the European Court of Justice accepted the challenge, prohibiting the use of gender difference in insurance tarification once and for all (C-236/09, Test-Achats).

The paradoxical result of this ruling has been that of indirectly creating new unfairness and discrimination. The Court of Justice referred to the general criterion (built on the Aristotelian idea of justice) whereby similar cases should be treated in like manner and different cases in a different manner, which is the basis of the risk transfer principle. In the case of third-party liability motor insurance, it has been known for some time that females are less exposed to risk of accident than males, but for the Court of Justice this depends not so much on gender in itself as on other variables, such as employment (men use their cars more for work), or excessive alcohol consumption (men drink more than women), although these are factors that are hard for insurance companies to monitor. The European Court of Justice thus seems to be saying that gender is being used as a proxy for other variables that are causally correlated to the different accident rates of males and females and, as such, is to be considered discriminatory (Cather 2020). In its judgement, the Court of Justice tried to find a tricky balance between ethical commitment (a certain idea of justice), technical problems (actuarial calculations) and legal prerequisites (nobody can be blamed for qualities not directly connected with the subject of blame).

But the introduction of unisex tariffs has obliged companies to reduce the premiums paid by men and increase those paid by women, despite the fact that men constitute more risk for insurance companies than women. In fact, if they are to comply with the principle of equivalence, insurance companies must guarantee that the pure premiums they receive are sufficient to cover expected losses. Unisex tarification eliminates one important factor of segmentation, obliging companies to opt for homogeneity when spreading expected financial risks. Let's suppose, for example, that the expected losses remain unchanged (let's say €160): while women previously paid a pure premium of €60 and men a pure premium of €100, with a unisex tariff both will pay a pure premium of €80. As a result, women end up having to *subsidise* the costs caused by men to insurance companies. If we then also consider that women on average earn less than men, the increase in the price of third-party liability motor insurance policies for women and the corresponding decrease in the price of third-party liability motor insurance policies for men have been considered by many observers to be an involuntarily iniquitous and discriminatory measure (Porrini 2011; Cipriani 2013; Fusco and Porrini 2020). Ultimately, then, it is paradoxical that unisex rating produces differences in insurance premiums for females and males, because when conditions remain unchanged, although they pay the same premium, women pay more than men for the risk the former really transfer to insurance companies (Fusco and Porrini 2020: 9).

There are at least two crucial points at stake in this exemplary matter. The first concerns the principle implicit in the ruling of the Court of Justice, which accepted the challenge lodged by the association Test-Achats, i.e. the fact that it is unfair to use proxy data even when they have a strong statistical correlation with the company's expected losses and thus comply with the principle of actuarial fairness.[20] The second point is that the entire issue that led up to unisex tariffs turns on a non-behavioural variable. But what changes if the company can have access to the values of behavioural variables? In other words, what changes if the company can observe how its policyholders really behave in practice?

One of our interviewees introduced an extremely interesting (and debatable) line of thinking about this, one that we believe deserves to be quoted here at length. In his opinion, an analysis based on the behavioural variables that telematics are capable of monitoring

> is *a very different analysis from any other way that insurance is used,* right? The rest of insurance is looking at *proxy variables* … that are *correlated* but not *causative*. What we're trying to do, specifically, is *identify those behaviours that are causative of risk* … These have turned out to be a really important element and *non-discriminatory*, because all those other elements, I mean … you yourself could move to the south of Italy and be paying more for insurance all of a sudden, but you're not a different driver, it just happens to be where you move to, your postcode changes, and the case of someone in the States is that if they get married or they get divorced or something happens to their credit, then all of a sudden they're paying more for insurance as a result of that, and none of that increases their risk, *it just happens to be the way that actuarial analysis looks at it,* looks at large groups of people and then tries to find patterns in laws.
>
> We're looking at *behavioural elements* and we're servicing that back to the driver. It's a very different approach from looking at risk. *I don't think there's a worry about penalising drivers in this case,* because we know that the drivers that have very high risk points are also drivers that get into

[20] Much of the discussion about algorithmic fairness concerns the difficult trade-off that ought to be achieved between predictive accuracy (as we know, proxy data often offer very strong correlations, which makes them useful for various different forms of business) and indirect discrimination. Cf. Loi and Christen (2021) for an overview of the debate from an ethical rather than legal standpoint.

> crashes and are also drivers that make claims. (Interview E.1, 30 March 2021)

In formal terms, this line of thinking corresponds exactly to the approach adopted by the European Court of Justice and described above, which considers the use of proxy data to be unfair. In traditional procedures used by insurance companies, individuals are not evaluated on the basis of their specific conditions and real behaviour, but because they are labelled as belonging to a group, such as that of individuals of female gender, or of divorcees, or of drivers who are resident in southern Italy. This kind of valuation is based on correlation, which notoriously does not necessarily correspond to any causal relationship. On the other hand, a valuation based on behavioural variables would not be unfair, so would also not be discriminatory, because it would be based neither on proxies, nor on statistical generalisations which, however well they may be segmented, place individuals in a group without taking their specific exposure to risk into account. The valuation made possible by telematics tools targets persons as individuals, taking their behaviour into account, without reference to any given groups, and so – we would add – without reference to any form of solidarity. Nobody would find themselves having to subsidise costs caused by others.

This substantial erosion of the principle of subsidiarity has stirred up a hornet's nest among analysts, as the customisation of premiums would lead to a drastic transformation of the principle of mutuality that has traditionally underpinned insurance. One unintentional effect of the predictive valuation of policyholders' behaviour would be the fact that the better we can predict future risk, "the less we'll be willing to share our fates with others" (Croll 2012). So the inevitable result would be the end of subsidisation.

Nevertheless, insurance practices that use behavioural data paint a different picture. To our way of thinking, the first thing is to draw a distinction between the principle of subsidisation and that of risk spreading. As we have seen, subsidisation is the issue at stake when members of one segment offset members of a different segment (younger drivers offset older drivers, women offset men and so on). The issue is more strictly one of risk spreading, on the other hand, when members of one and the same pool contribute to reimbursing the total losses expected within that pool. Segmentation operates in the opposite way to subsidisation, but does not do so without the classical mechanism of risk spreading, which remains the underlying foundation that makes the insurance industry function.

The behavioural valuation of policyholders for third-party liability motor insurance does not stop members of a pool from sharing a destiny that is still, to a considerable extent, predicted by non-behavioural factors. As we have seen, behavioural data are only complemented at a later stage in the actuarial architecture whose purpose is to identify the policyholders with whom a specific loss prediction is associated. This means that behavioural data drive the principle of risk transfer to extremes: those who have a better risk profile – either because they drive less, or because they drive more prudently – contribute less to reimbursing the expected losses within the pool than those who have a worse profile. Taken as a whole, though, the mechanism is still based essentially on risk spreading.

Behavioural data could nevertheless intensify the *tension* between the individual and the group. To our way of thinking, a real sea-change would only come about if the destiny (so the pure premium) were calculated *exclusively* on the basis of behavioural data. If that were the case, there would no longer be any pools, but in fact only "behavioural tribes", in which, as Cathy O'Neil puts it (2016, ch. 9) "those who act alike take on similar levels of risk". The underlying principle of risk pooling and spreading would then have to be fundamentally redesigned.

Conclusion

In concluding our exploration, we want to mention some considerations about possible social consequences of the use of digital technologies and algorithmic prediction in the field of insurance. Since Ewald (1986), research has shown that the impact of insurance on society depends on the form of solidarity in the management of future risk, which is closely related to available techniques to calculate and control them. Until now, these techniques were based fundamentally on statistical actuarial models. If we now look at the introduction of algorithmic techniques that follow a different logic, how does this affect the distribution of responsibility and risk management in society as a whole?

Our analysis shows that, at the moment, real innovation lies in the possibility of focusing on individual behaviour in a way that was previously unfeasible, which affects the relationship between insurance companies and policyholders and thereby also the respective business model. But what consequences does this have for the form of social solidarity and the willingness to take risks?

While it is true that risk often depends on individual behaviour, it is also true that behaviour does not always depend on the individual. Any connection between injury and behaviour raises tricky questions about the idea of responsibility and the criteria according to which a person should or should not be blamed for that injury. As Tom Baker (2003) points out with respect to moral justifications of risk classification, if someone deserves low-risk status, the moral claim to benefiting from that status is a strong one. For the same reason, if high-risk status is not deserved (as in the case of battered women), the moral claim to penalising it is weaker. Yet, the issue is that it is not always clear, as Lehtonen and Liukko (2015: 164) argue, "to what extent a person can be held responsible for his or her lifestyle, social [and cultural] milieu, or area of residence".

On the other hand, if risk is attributed to the individual, it does not necessarily follow that policyholders want to accept responsibility for all their behaviours and for possible consequences. Insurance valuations should take this eventuality into account. Valuations that refer to behavioural variables ought presumably to continue being combined with actuarial statistical considerations to keep the basic mechanism of insurance working, but maybe a case can be made for increasing customisation by exercising greater freedom in defining groups and how they are made up – including the group of those who do not want to accept the burden of being evaluated on the basis of their behaviour. This leads to our second consideration.

Historically speaking, insurance was introduced not to induce individuals to keep their exposure to risks under control, but to relieve individuals of their worries about possible future damages. As François Ewald has pointed out (1991), insurance was introduced as a "liberator of action", to enable individuals to undertake risky businesses in a relatively protected manner. The contingent financial certainty offered by insurance coverage allows policyholders to venture into activities whose future course, despite all the risks to which the enterprise exposes the insured, can be considered a "rival choosable" to the future course that would be realised if the insured were to give up the enterprise (Cevolini 2019b). The traditional purpose of insurance, therefore, has never been that of reducing risks, but it could be said to have been more that of multiplying them, guaranteeing the possibility of managing their consequences (Luhmann 1996). For policyholders, the possibility of falling ill, of having a car accident or that their house burns down is not reduced: if anything, it increases, as illustrated by the chronic problem of moral hazard (Stone 2002).

Moral hazard is, notoriously, a crucial issue of actuarial science (Baker 1996). The idea is that if I know that my insurance company will compensate against damage, I have less incentive to be careful and prevent an accident from happening.[21] The result can be an increase in risk exposure and consequently in prospective claims (Arrow 1971; Stiglitz 1983; Heimer 1985.). The paradox is, in the end, that "less loss

[21] In this respect, Heimer (1985) speaks of "reactive risks". Ferdinand Tönnies (1917) had already pointed out that this happens because the financial certainty provided by coverage makes the policyholder somehow unconcerned about the damage. On Tönnies's sociological investigation about insurance see Cevolini (2019a).

[for the insured] from a loss means more loss [for the insurer]" (Baker 1996: 270).[22]

Since the early 2000s, insurance companies have begun to address the problem of moral hazard emphatically in terms of prevention (Baker and Simon 2002a; Ericson et al. 2003). Scholars suggested that this trend was an effort to make people "more individually accountable for risk", that is, to let policyholders (at least in part) "embrace the risks" they wanted to insure against (Baker and Simon 2002b: 1).[23] From the insurance companies' standpoint, the question was how to "govern the insured", moving from the assumption that a safer environment not only means a better loss ratio for insurers but also materialises in lower premiums for insureds (Ericson et al. 2003). The intent was to explicitly turn insurance companies into "loss prevention companies" and thus to "stop claims before they happen" (Ericson et al. 2003: 271). However, how to reach this kind of self-education of policyholders was not clear and at the end insurance companies resorted to usual (and more practical) solutions such as spreading risks.

It now seems that the problem of moral hazard can be to some extent "technically controlled" by means of telematics technology (Van Hoyweghen et al. 2006: 1231). If 20 years ago the basic idea was to make drivers accountable for "being aware of crash risks and the consequences of miscalculating them" (Ericson et al. 2003: 272), yet without knowing how to detect such awareness, the feeling now is that by means of telematics data it is possible to *calculate* precisely *this miscalculation*. However, telemetry-based technology does not *control* the policyholders' behaviour. Their driving behaviour, instead, produces telematics data which is algorithmically processed to extract information. The latter is not only used by insurance companies to implement a more individualised risk assessment, but it is also fed back to drivers in order to give them the opportunity to control themselves. When this *self-control* is carried to the extreme, it can turn into a particular kind of inhibition. We wonder, then, if the introduction of behavioural valuations that burden the present with responsibility for the future could be, for policyholders, a source of anxiety that would eventually transform insurance into an "inhibitor of action" (Cevolini and Esposito 2020). Research will have to clarify

---

[22] Insurance companies have attempted to curb moral hazard by including clauses in policies whose purpose is to again increase the policyholder's incentive to be prudent. Loss-prevention measures such as rewards (e.g., no-claims bonus) and penalizing agreements (e.g., franchise and deductible) actually serve to encourage policyholders to keep their exposure to dangers under control and to assume at least some of the responsibility in case an accident occurs (Heimer 1985).

[23] Twenty years ago Baker and Simon (2002b: 3 ff.) spoke of a shift "from spreading risk to embracing risk", although they admitted that the latter does not remove the former.

whether, and to what extent, this may affect the social function of insurance and individuals' ability to enact the future, starting from available insurance coverages.

Acknowledgements: We are grateful to the anonymous reviewers and the editorial board of Valuation Studies for their comments, criticisms and suggestions. Our warm thanks as well to the participants of the workshop "The Future of Insurance. How Does Algorithmic Prediction Affect Insurance Practice", University of Bologna, 18–19 November 2021 for their contribution to a lively debate and an exciting in-depth discussion of the first draft of our article. This work was supported by the European Research Council (ERC) under Advanced Research Project PREDICT no. 833749.

References

Albrecher, Hansjörg, Antoine Bommier, Damir Filipovic, Pablo Koch-Medina, Stéphan Loisel, and Hato Schmeiser. 2019. "Insurance: Models, Digitalization, and Data Science." *European Actuarial Journal* 9(2): 349–360.

Albrecht, Peter. 1992. *Zur Risikotransformationstheorie der Versicherung: Grundlagen und ökonomische Konsequenzen*. Karlsruhe: Verlag VersicherungsWirtschaft.

Arrow, Kenneth. 1971. "Uncertainty and the Welfare Economics of Medical Care." In *Essays in the Theory of Risk-Bearing,* 177–211. Amsterdam and London: North-Holland.

Baker, Tom. 1996. "On the Genealogy of Moral Hazard." *Texas Law Review* 75(2): 237–292.

Baker, Tom. 2003. "Containing the Promise of Insurance: Adverse Selection and Risk Classification." In *Risk and Morality,* edited by Richard V. Ericson and Aaron Doyle, 258–283. Toronto, ON: University of Toronto Press.

Baker, Tom and Jonathan Simon (Eds.). 2002a. *Embracing Risk. The Changing Culture of Insurance and Responsibility.* Chicago, IL and London: The University of Chicago Press.

Baker, Tom and Jonathan Simon. 2002b. "Embracing Risk." In *Embracing Risk. The Changing Culture of Insurance and Responsibility,* edited by Tom Baker and Jonathan Simon, 1–25. Chicago, IL and London: The University of Chicago Press.

Barry, Laurence. 2020. "Insurance, Big Data and Changing Conceptions of Fairness." *European Journal of Sociology* 61(2): 159–184.

Barry, Laurence and Arthur Charpentier. 2020. "Personalization as a Promise: Can Big Data Change the Practice of Insurance?" *Big Data & Society* 7(1): 1–12.

Biferali, Giorgia. 2018. "Big data e valutazione del merito creditizio per l'accesso al peer to peer lending." *Il diritto dell'informazione e dell'informatica* 33(3): 487–509.

Billot, Romain, Cécile Bothorel, and Philippe Lenca. 2018. "Introduction to Big Data and Its Applications in Insurance." In *Big Data for Insurance Companies,* edited by Marine Corlosquet-Habart and Jacques Janssen, 1–25. London: John Wiley & Sons.

Boobier, Tony. 2016. *Analytics for Insurance. The Real Business of Big Data.* Chichester: John Wiley & Sons.

boyd, dana and Kate Crawford. 2012. "Critical Questions for Big Data. Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication and Society* 15(5): 662–679.

Braun, Alexander and Florian Schreiber. 2017. *The Current InsurTech Landscape: Business Models and Disruptive Potential.* St. Gallen: Institute of Insurance Economics, University of St. Gallen and Swiss Re Institut.

Cather, David. 2018. "Cream Skimming: Innovations in Insurance Risk Classification and Adverse Selection." *Risk Management and Insurance Review* 21(2): 335–366.

Cather, David. 2020. "Reconsidering Insurance Discrimination and Adverse Selection in an Era of Data Analytics." *The Geneva Papers on Risk and Insurance – Issues and Practice* 45: 426–456.

Cevolini, Alberto. 2019a. "Die Sicherung der Versicherung. Ferdinand Tönnies als Vorläufer einer Soziologie des Versicherungswesens." *Tönnies-Forum* 28(2): 22–33.

Cevolini, Alberto. 2019b. "Insurance as a Business of Imagination." *Sociologia e Politiche Sociali* 22(2): 105–125.

Cevolini, Alberto and Elena Esposito. 2020. "From Pool to Profile: Social Consequences of Algorithmic Prediction in Insurance." *Big Data & Society* 7(2): 1–11.

Cipriani, Giam Pietro. 2013. "Tariffe unisex, una stangata per giovani automobiliste." *lavoce.info,* 12 April. https://www.lavoce.info/archives/8487/tariffe-unisex-una-stangata-per-giovani-automobiliste/, accessed 5 May 2021.

Cotrugli, Benedetto. 1602. *Della mercatura et del mercante perfetto* [orig. ed. 1458]. Brescia: Alla Libraria del Bozzola.

Croll, Alistair. 2012. "New Ethics for a New World." *Radar O'Reilly*, 17 October. radar.oreilly.com/2012/10/new-ethics-for-a-new-world.html, accessed 5 May 2021.

Dang, Julia. 2017. *Unveiling the Full Potential of Telematics. How Connected Insurance Brings Value to Insurers and Consumers: An Italian Case Study.* Zurich: Swiss Re.

Daston, Lorraine. 1983. "Rational Individual Versus Laws of Society: From Probability to Statistics." In *Probability Since 1800. Interdisciplinary Studies of Scientific Development*, edited by Michael Heidelberger, Lorenz Krüger, and Rosemarie Rheinwald, 7–26. Bielefeld: B. K. Verlag.

De Wit, G. Willem and Jacob Van Eeghen. 1984. "Rate Making and Society's Sense of Fairness." *ASTIN Bulletin* 14(2): 151–164.

Domingos, Pedro. 2015. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York: Basic Books.

Eckert, Christian and Katrin Osterrieder. 2020. "How Digitalization Affects Insurance Companies: Overview and Cases of Digital Technologies." *Zeitschrift für die gesamte Versicherungswissenschaft* first online October 1: 1–28.

Eling, Martin and Mirko Kraft. 2020. "The Impact of Telematics on the Insurability of Risks." *Journal of Risk Finance* 21(2): 77–109.

Eling, Martin and Martin Lehmann. 2020. "The Impact of Digitalization on the Insurance Value Chain and the Insurability of Risks." *The Geneva Papers on Risk and Insurance – Issues and Practice* 43(3): 359–396.

Ericson, Richard, Aaron Doyle, and Dean Barry. 2003. *Insurance as Governance*. Toronto, ON: University of Toronto Press.

Ewald, François. 1986. *L'Etat providence*. Paris: Grasset.

Ewald, François. 1991. "Insurance and Risk." In *The Foucault Effect. Studies in Governmentality,* edited by Graham Burchell, Colin Gordon, and Peter Miller, 197–210. London: Harvester Wheatsheaf.

Farny, Dieter. 1992. *Buchführung und Periodenrechnung im Versicherungsunternehmen*. Wiesbaden: Gabler.

Farny, Dieter. 1995. *Versicherungsbetriebslehre*. Karlsruhe: Verlag VersicherungsWirtschaft.

Fusco, Giulio and Donatella Porrini. 2020. "Less Discrimination, More Gender Inequality: The Case of the Italian Motor-Vehicle Insurance." *The Journal of Risk Management and Insurance* 24(1): 1–11.

Gillespie, Tarleton. 2014. "The Relevance of Algorithms." In *Media Technologies. Essays on Communication, Materiality, and Society*, edited by Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot, 167–194. Cambridge, MA: The MIT Press.

Gitelman, Lisa (Ed.) 2013. *"Raw Data" is an Oxymoron*. Cambridge, MA: The MIT Press.

Harcourt, Bernard. 2006. *Against Prediction: Profiling, Policing and Punishing in an Actuarial Age*. Chicago, IL: University of Chicago Press.

Harrington, Joseph. 2017. "A Risk Pool of One: How Technology Could Make Insurance Obsolete." *Insurance Journal,* 15 May.

Heimer, Carol. 1985. *Reactive Risk and Rational Action. Managing Moral Hazard in Insurance Contracts*. Berkeley, CA: University of California Press.

Innami, Hirokichi. 1996. "Das Äquivalenzprinzip in der Versicherungswirtschaft." *Zeitschrift für die gesamte Versicherungswissenschaft* 55: 17–30.

Kirkpatrick, Keith. 2016. "Battling Algorithmic Bias: How Do We Ensure Algorithms Treat Us Fairly?" *Communications of the ACM* 59(10): 16–17.

Lehtonen, Turo-Kimmo and Jyri Liukko. 2011. “The Forms and Limits of Insurance Solidarity.” *Journal of Business Ethics* 103(1): 33–44.

Lehtonen, Turo-Kimmo and Jyri Liukko. 2015. “Producing Solidarity, Inequality and Exclusion through Insurance.” *Res Publica* 21: 155–169.

Loi, Michele and Markus Christen. 2021. “Choosing How to Discriminate: Navigating Ethical Trade-Offs in Fair Algorithmic Design for the Insurance Sector.” *Philosophy and Technology* first online March 13: 1–26.

Luhmann, Niklas. 1991. *Soziologie des Risikos*. Berlin and New York: Walter de Gruyter.

Luhmann, Niklas. 1996. “Das Risiko der Versicherung gegen Gefahren.” *Soziale Welt* 43(3): 273–284.

McFall, Liz and Liz Moor. 2018. “Who, or What, Is InsurTech Personalizing? Persons, Prices and the Historical Classifications of Risks.” *Distinktion: Journal of Social Theory* 19(2): 193–213.

McFall, Liz, Gert Meyers, and Ine Van Hoyweghen. 2020. “Editorial: The Personalisation of Insurance: Data, Behaviour and Innovation.” *Big Data & Society* 7(2): 1–11.

Mahr, Alexander. 1951. *Einführung in die Versicherungswirtschaft. Allgemeine Versicherungslehre*. Berlin: Duncker & Humblot.

Miller, Claire Cain. 2015. “Can an Algorithm Hire Better Than a Human?” *The New York Times,* 25 June. https://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html, accessed 5 May 2021.

Minty, Duncan. 2021. “Why Equality of Fairness Will Shape the Future of Insurance.” *Ethics and Insurance,* 3 March. https://ethicsandinsurance.info/2021/03/03/equality-of-fairness/, accessed 5 May 2021.

Oñate, Petrus. 1654. *De contractibus onerosis tomus tertius*. Romae: Typis Angeli Bernabò haeredis Manelphij.

O’Neil, Cathy. 2016. *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy.* New York: Broadway Books.

Pasquale, Frank. 2015. *The Black Box Society. The Secret Algorithms That Control Money and Information.* Cambridge, MA and London: Harvard University Press.

Peer, Stefanie, Alexander Mauermann, and Katharina Sallinger. 2020. “App-Based Feedback on Safety to Novice Drivers: Learning and Monetary Incentives.” *Transportation Research Part F* 71: 198–219.

Porrini, Donatella. 2011. “Chi ha paura delle polizze rosa?” *lavoce.info,* 25 March. https://www.lavoce.info/archives/26830/chi-ha-paura-delle-polizze-rosa/, accessed 5 May 2021.

Provost, Foster and Tom Fawcett. 2013. *Data Science for Business. What You Need to Know about Data Mining and Data-Analytics Thinking*. Beijing: O’Reilly.

Ramasastry, Anita. 2012. “Progressive Car Insurance’s ‘Snapshot’ Experiment: Should Consumers Be Wary of Being Individually Tracked

While Driving?" *Verdict,* 14 August. https://verdict.justia.com/2012/08/14/progressive-car-insurances-snapshot-experiment, accessed 5 May 2021.

Rona-Tas, Akos. 2020. "Predicting the Future: Art and Algorithms." *Socio-Economic Review* 18(3): 893–911.

Siegel, Eric. 2016. *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. Hoboken, NJ: John Wiley.

Stevenson, Mark, Anthony Harris, Duncan Mortimer, Jasper S. Wijnands, Alan Tapp, Frank Peppard, and Samantha Buckis. 2018. "The Effects of Feedback and Incentives-Based Insurance on Driving Behaviours: Study Approach and Protocols." *Injury Prevention* 21(1): 89–93.

Stevenson, Mark, Anthony Harris, Jasper S. Wijnands, and Duncan Mortimer. 2021. "The Effects of Telematic Based Feedback and Financial Incentives on Driving Behaviour: A Randomised Trial." *Accident Analysis and Prevention* 159: 1–8.

Stiglitz, Joseph. 1983. "Incentives and Insurance: The Pure Theory of Moral Hazard." *The Geneva Papers on Risk and Insurance* 8(26): 4–33.

Stone, Deborah. 2002. "Beyond Moral Hazard: Insurance as Moral Opportunity." In *Embracing Risk. The Changing Culture of Insurance and Responsibility,* edited by Tom Baker and Jonathan Simon, 52–79. Chicago, IL and London: The University of Chicago Press.

Tanninen, Maiju. 2020. "Contested Technology: Social Scientific Perspectives of Behaviour-Based Insurance." *Big Data & Society* 7(2): 1–14.

Thiery, Yves and Caroline Van Schoubroeck. 2006. "Fairness and Equality in Insurance Classification." *The Geneva Papers on Risk and Insurance: Issues and Practice* 31(2): 190–211.

Tönnies, Ferdinand. 1917. "Das Versicherungswesen in soziologischer Betrachtung." *Zeitschrift für die gesamte Versicherungswissenschaft* 17: 603–624.

Van Hoyweghen, Ine, Klasien Horstman, and Rita Schepers. 2006. "Making the Normal Deviant: The Introduction of Predictive Medicine in Life Insurance." *Social Science & Medicine* 63: 1225–1235.

Alberto Cevolini is Professor of Sociology at the University of Modena and Reggio Emilia and a senior researcher in the project "Personalized Insurance" at the University of Bologna as part of the main research project "The Future of Prediction: The Social Consequences of Algorithmic Forecast in Insurance, Medicine and Policing" (ERC-2018-ADG, no. 833749, PI Prof. Elena Esposito). He has published extensively on social system theory and the sociology of insurance.

Elena Esposito is Professor of Sociology at the University Bielefeld and the University of Bologna. She has published extensively on the theory

of society, media theory, memory theory and the sociology of financial markets. Her current research on algorithmic prediction is supported by a five-year Advanced Grant from the European Research Council (ERC). Her latest book is *Artificial Communication: How Algorithms Produce Social Intelligence,* MIT Press, 2022.

Valuation Studies 9(1) 2022: 141-170

Dynamics of Standardised Quality. Long-term shifts in organic product qualification

Nadine Arnold and Simon Dombrowski

Abstract

The qualities of standardised products are often perceived as naturally stable. This article scrutinises this perceived stability and investigates which aspects of standardised quality remain stable, and which change in the longer term. Our conceptual framework, anchored in the literature on standards and valuation studies, suggests that while standardised qualities appear to be stable over time and space, it is in these spatial and temporal dimensions of qualification that controversies and changes are expected. Empirically, we investigate the organic quality which has been maintained in the German mass market since the 1970s by the standard-setter Bioland. Searching our archival data for disruption that refers to events, which were interpreted by Bioland as reasons for adjusting the qualification, the data show that Bioland reacted swiftly to manifold disruption triggered by actors located along the production and distribution chain as well as outside it. Pooling Bioland's responses, we identify four shifts in terms of the (1) meaning, (2) focus, (3) organisation, and (4) relationships of quality. Due to these long-term shifts, little except the name of the standardised quality remained stable. Thus, the article concludes that standardised qualification must be dynamic and changeable if it is to be stably relevant in markets.

Keywords: standards; market; organisation; stability; diversity; archival research

Nadine Arnold, PhD in Sociology, is Assistant Professor of Organisational Theory at the Vrije Universiteit Amsterdam.
Simon Dombrowski, PhD in Sociology, is a science manager at the Hamburg University of Technology.

© 2022 The authors This work is licensed under a *Creative Commons Attribution 4.0 International License*.
https://doi.org/10.3384/VS.2001-5992.2022.9.1.141-170
Hosted by Linköping University Electronic press
http://valuationstudies.liu.se

Introduction

For a standardised product to be exchanged in a market, it must go through a qualification process, at the end of which the product will have obtained its qualities (Callon et al. 2002; Beckert and Aspers 2011; Helgesson and Kjellberg 2013). Every day, countless standardised products, hardly exciting due to their seemingly stable qualities, are exchanged in markets across the world. This applies—but not exclusively—to food sold in supermarkets. For example, think of bananas, apples, flour, and pasta. Such products are qualified in a standard form, which means that their characteristics are presented in an analytic way that allows for mass production and long-distance trade (Boltanski and Esquerre 2020). By the time a standardised product is exchanged in a market, its qualities appear unambiguous and uncontested (Bessy and Chauvin 2013). The reason for this is that standardised products' qualification, which involves classification (assessing the broader category of the product and what it is) and evaluation (considering how good or bad the specimen is) (Kuipers and Franssen 2020), precedes the products' exchange in the market (Eymard-Duvernay 1989; Musselin and Paradeise 2005; Beckert and Musselin 2013; Arnold and Hasse 2016). Looking at the actual act of exchange in the market, one gets the impression that standards stabilise the quality of products. This general impression has been substantiated by studies that distinguish standardised, reproducible qualities from singular and exceptional ones (Karpik 2010; Boltanski and Esquerre 2020) or explain that standard forms can achieve an "exceptional stability and universality" (Thévenot 1984: 11).

In contrast, the literature on standards emphasises the dynamics of standards and standardisation (Brunsson and Jacobsson 2000; Timmermans and Epstein 2010; Higgins and Larner 2010; Busch 2011; Brunsson et al. 2012; Loconto and Demortain 2017). Under the thesis that we live in "a world of standards but not a standard world" (Timmermans and Epstein 2010: 69), scholars highlight the proliferating multiplicity of standards (Djelic and den Hond 2013; Arnold and Loconto 2021), the use of standards to create differentiation and diversity (Busch 2011; Loconto and Demortain 2017), or the manifold tensions in the setting and following of standards (e.g. Brunsson and Jacobsson 2000; Higgins and Larner 2010). In this context, Brunsson et al. (2012: 627) highlight that "while standards might aim at the creation of stability and sameness, standardisation itself is a highly dynamic phenomenon". From here, this article aims to shed light on the dynamics of a long-term standardised product qualification process by asking which aspects of standardised quality remain stable, and which change in the longer term.

The impression that standardised qualities are stable is based on the fact that standardisation is closely linked to organisation (Bowker and

Star 1999: 37; Brunsson et al. 2012, Gustafsson 2020). The term organisation can be understood to refer to two phenomena. First, it refers to the formal standardisation organisations that take decisions relevant to qualification (e.g. International Organisation of Standardisation (ISO), Fairtrade International, safety agencies). We call these formal organisations standard-setters. Second, organisation refers to the standards and other organisational elements (e.g. controls, sanctions, rankings) decided upon to influence the development and maintenance of quality (Ahrne and Brunsson 2011, 2019). Both organisational standard-setters and elements can be considered market intermediaries, as they are not usually part of either the supply or demand side of the market but help to reduce the uncertainty of market exchanges by defining the relationships between buyers and sellers (Eymard-Duvernay 1989; Musselin and Paradeise 2005; Beckert and Musselin 2013; Bessy and Chauvin 2013; Ahrne et al. 2015). In doing so, they establish relationships of trust between producers and consumers and play a pivotal role in the successful qualification of market objects (Varga 2019; Wilde 2020).

While standardised product qualities maintained by formal standard-setters appear to be stable over time and space, our conceptual framework, anchored in the literature on standards and valuation studies, will suggest that it is in these spatial and temporal dimensions of qualification that tensions and changes are expected. Empirically, we investigate organic quality, a standardised product quality that has been introduced and adapted in the German mass market since the 1970s, among others, by the standard-setter Bioland. The advanced age of this specific quality will bring us the unique opportunity to examine and identify long-term shifts in a standardised qualification that concerns not only the meaning and focus of the qualification but also its organisation and the relationships behind it. These empirical findings will allow us to argue that lively dynamics are a prerequisite for maintaining quality in an ostensibly stable manner.

In the remainder of this article, we first develop our conceptual framework for studying dynamics in standardised qualification. Then, we introduce our empirical case study and provide information about our methodology. Thereafter we present our empirical findings, identify a broad variety of actors who disrupted the qualification, and describe how Bioland responded to qualification disruption while maintaining its product quality. We discuss our findings in the fourth section by identifying four major shifts that allowed organic quality to be maintained in a growing space and to endure over time. We conclude with a brief reflection on what it means when standardised quality goes hand in hand with dynamic processes.

Standardised qualification: Temporal and spatial processes between stability and diversity

Qualification is a process that should gain stability through the use of standards because that is standardisation's aims (Timmermans and Epstein 2010: 84). Given that the notion of stability refers to the idea that something has the "strength to stand and endure,"[1] a stable qualification is perceived as one that neither differs depending on the setting nor changes over time. This understanding manifests when we approach standards as investments in form, following Thévenot (1984).[2] He argued that investments in form vary in stability and universality depending on their lifespan and area of validity. While the latter (area of validity) depends on where a certain investment is applied, a long lifespan gives "the right to reproduce a particular form […] over a certain period of time" (Thévenot 1984: 11). In addition to patents and government regulations, standards are an illustrative example of a form-giving investment with a long life and a high degree of validity (Thévenot 2015). This means that standards are expected to stabilise in space and time; but paradoxically, it is precisely in these dimensions (spatial and temporal) that the triggers for change must be suspected. Let us interrogate both dimensions, one after the other.

Spatial dimension

The spatial dimension of product qualification refers to the space in which, at a given point in time, actors bring different definitions of quality and bargain over them (Musselin and Paradeise 2005). The actors who define the qualification space are, of course, the buyers and sellers who exchange the focal product in a specific locale. This becomes clear, for example, in Garcia-Parpet's (2007) description of the strawberry auction in Fontaines-en-Sologne. At the auction, sellers display the strawberries for sale to buyers in the auction hall. When determining the price during the auction process, market participants are physically present. This seems to be a commonplace procedure for standardised products, and Dobeson and Kohl (2020: 45) write that standardised non-durable products "are usually evaluated and traded 'on the spot' […] between multiple buyers and sellers." However, the qualification space for standardised products is not limited to places of

---

[1] *Merriam-Webster*, s.v. "stability," accessed 22 April 2021. https://www.merriam-webster.com/dictionary/stability.

[2] Thévenot's work also indicates that stabilising attempts of standardisation are closely linked to organisational phenomena. Close reading reveals that Thévenot's (1984) conceptualised investments in forms is based on a reinterpretation of Taylor's seminal *Principles of Scientific Management* (1911 [2016]), which are known to unfold in and from organisational contexts. In other words, investment in form typically results from organisational effort.

exchange between buyers and sellers (e.g. auction halls and consumer markets). The exchange of things is only one stage in their lives (Appadurai 1986), and qualification "occurs across social contexts, and certainly not exclusively in economic settings" (Kuipers and Franssen 2020: 144).

Indeed, Garcia-Parpet (2007) shows that actors other than those who exchange strawberries influence strawberries' qualification. For example, the Comite Economique du Val de Loire (a formal standard-setter) shapes the qualification process via its quality criteria. In fact, due to the involvement of actors other than sellers and buyers, the strawberries' qualification process is distributed among many locations and instances (e.g. the auction hall, sessions of the economic committee that set quality criteria, and laboratories of the seed manufactures). Therefore, qualification processes are shaped not just by the criteria applied by people in a single social space (Boltanski and Thévenot 1983); rather, multiple actors from various spaces are involved in the qualification process. Supporting this position, Callon et al. (2002) claimed that products are qualified in "hybrid forums" of experts from different disciplines (e.g. economics, law, food science, media, and consumer protection) who are not necessarily visible during the actual market exchange. In the case of standardised qualities, the actors who set and enforce the standards play a particularly relevant role.

It is this diversity of actors that whirls up qualification, provoking shifts and dynamics. This is also true in the case of standardised qualification, as Loconto and Demortain (2017) argue. They put the thesis forward that standardisation is the result of a dynamic interaction of three spaces in which standards are made, followed, and circulate. This means, first, diversity emerges when actors decide about standards in conflict-and power-laden processes, which typically take place within formal standard-setters (e.g. Hallström and Boström 2010). Second, diversity results when actors follow and translate standards in locally contingent ways (e.g. Higgins and Larner 2010; Arnold and Loconto 2021), or third, when standards circulate between competing standard-setters and other relevant third parties, such as government agencies, social movements, or control authorities (e.g. Busch 2011; Gustafsson 2020; Arnold 2022). Consequently, standards do indeed attempt to stabilise in different spatial arenas and can achieve a high degree of diffusion, but it is this diffusion that makes standardisation dynamic.

Temporal dimension

By the term temporal, we refer to the long-term dimension of qualification, while others, when interrogating the temporal dimension of qualification and valuation, typically focus on "moments of valuation" (Antal et al. 2015). In these temporally restricted moments, products are requalified due to different orders of worth (Boltanski and Thévenot 2006), competitors seek to detach consumers from competing products (Callon et al. 2002), and/or buyers mobilise different judgement devices for evaluating products (Karpik 2010). Examining such moments is useful for understanding the qualification of things with uncertain qualities, such as singular goods (Karpik 2010) or counterfeits (Bessy and Chateauraynaud 2014). In the case of standardised products, however, quality appears momentarily stable, whereas we can expect dynamics in the longer term. When referring to long-term product qualities, we do not refer to the individual product's material durability. Other researchers have dealt with the distinction between durable and non-durable products (Dobeson and Kohl 2020). Rather, we mean that a specific quality is attributed to many different products over several years, or even decades.

Standards are a key tool for attributing and evaluating qualities in a stable way over a long period, but this longevity is also a cause of change. While long-term developments in qualification have received little scholarly attention (Musselin and Paradeise 2005: 26–32), we know that time is a risk of standardising (Bowker and Star 1999: 193). In this vein, Timmermans and Epstein wrote that "standards can stabilise some action in a moving world, but when the world around the standards changes, the standard quickly becomes outdated or altered" (Timmermans and Epstein 2010: 84). To keep up with societal changes, standards that underpin the qualification process transform with a tendency to accumulate, and they are complemented by other organisational elements (e.g. controls and sanctions) to legitimise and enforce adoption of the standards (e.g. Loconto 2017; Gustafsson 2020; Arnold 2022).

Over time, however, it is not only the organisation of standardised quality that may change; it is equally possible that the quality itself may change. For example, for a long time, fair trade quality coffee stood for solidarity and an inferior, bitter taste, while today fair trade is also associated with exquisite roast aroma and flavour (Arnold 2017). Interestingly, Boltanski and Esquerre (2020) claim that the meaning of quality itself has changed. Following them, quality nowadays refers to something exceptional and special, whereas quality used to mean primarily something that is standardised and uniform. However, just as the meaning of quality or a specific quality can transform when detached from concrete objects, the quality of certain durable things can also shift. An illustrative example is rubbish. In this case, the value of an object declines until it is classified as rubbish, but this rubbish can then regain value through the attribution of new qualities, such as antiquity (Thompson 1979).

In sum, we know that standardised product qualification is destabilised by the actors involved in the making, following, and circulating of standards. If these disruptions are responded to in a quality-assuring way, long-term shifts of the qualification process and its quality are to be expected.

Case and method

Organic food qualification led by Bioland

Food is well suited for examining standardised product qualities. While economic sociologists have been intrigued by the study of singular foods whose qualification is controversial and highly ambiguous, such as foie gras (DeSoucey 2018) or wine (Garcia-Parpet 2011), most of the everyday food that we consume is highly standardised (Busch 2011). Over the last few decades, these rather mundane everyday food products (e.g. potatoes, apples, sausages) that are sold in ordinary supermarkets have been increasingly attributed to new qualities, such as being CO2-free, dolphin-safe, organic, fair trade, GMO-free, or environment-friendly. Rural sociologists have used the notion of "quality turn" to summarise this trend, referring to the shift from production-based qualities to qualities that emphasise nature and local embeddedness (Allaire and Sylvander 1997; Murdoch et al. 2000).

The specific quality that we are dealing with is organic quality, which has been attributed to a growing number of agricultural products for many years and has become the most well-known and most studied food quality (FAO 2014). Germany is one of the countries where the organic qualification process started early, after the organic–dynamic movement started advocating for more self-sustaining agricultural production methods in the 1920s (Conford 2001). Seeing that this movement has resulted in many formal standard-setters specialising in organic qualification (Biokreis e. V., Biopark e. V., Bio-Initiative GmbH, Bioland, Demeter e. V., Ecoland e. V., Ecovin e. V., Gäa e. V., Verbund Ökohöfe e. V., Naurland e. V.), Germany exemplifies the fact that the construction of standardised product qualities is fundamentally linked to the creation of organisations (Brunsson et al. 2012).

We chose to investigate Bioland's qualification process because it is particularly extended in terms of space and time. Before describing this extension in more detail, we would like to note that Bioland is a non-profit organisation that can be characterised as a hybrid between a commercial and critical intermediary (Karpik 2010: 100–101). As a critical intermediary, it is an association of organic food producers that does not engage in the exchange of organic products. Simultaneously, it is a commercial intermediary because its members, who represent half of the organised organic farmers in Germany, produce, and sell organic food.[3] Given that the producers of organic Bioland foods are also members of Bioland, we will refer to them as producer-members.

The qualification space in which Bioland operates is extensive, and Bioland is the largest organic food standard-setter in Germany in terms of producer-members (8.154) and cultivated area (475.068 ha).[4] Consequently, the Bioland label is widely recognised by more than half the German population.[5] In addition to this spatial spread, Bioland was a good match for our research purposes because it was founded in 1971 and has successfully maintained its organic qualification for 50 years. During this period, the meaning of organic quality changed significantly. We will now briefly describe Bioland's qualification set-up and what it stands for today.

Bioland's organic food quality emerged in the 1970s, when farmers from southern Germany adopted the concept of "organic farming" from Swiss farmers. At that time, the scope of this quality was particularly narrow and focused on small-scale local production chains. Only a few family farms in southern Germany specialising in grain harvesting and dairy farming practised organic culture. The producers who formed Bioland defined organic quality using the following six principles:

> life creates life;
> improve health;
> reduce costs;
> increase performance;
> do everything that promotes soil fertility; omit everything that destroys soil fertility;
> do not buy soil fertility, but build it yourself.
> (Authors' translation)[6]

The principles show that the idea of organic farming was originally meant to reduce agricultural production costs (i.e. the cost of chemical fertilisers and plant-protection agents) by naturally fostering soil fertility. Farms were conceptualised as circular economies that maintained soil fertility using the manure produced by livestock, while animals, in turn, were to be fed with the harvest from the fields. The

---

[3] Bund Ökologische Lebensmittelwirtschaft e.V. 2021. *Branchenreport 2021. Ökologische Lebensmittelwirtschaft.* Berlin.

[4] Bund Ökologische Lebensmittelwirtschaft e.V. 2021. *Branchenreport 2021. Ökologische Lebensmittelwirtschaft.* Berlin.

[5] Max Rubner Institut 2008. *Nationale Verzehrstudie II. Ergebnisbericht, Teil 1.* Karlsruhe.

[6] Siegfried Kuhlendahl 1996. "Auf dem Weg zum organisch-biologischen Landbau," *Bio-land* (1): 10–12.

resulting lower yield was to be offset by price premiums, which were justified by the fact that organic products were healthier and free of pesticides. Dairy, meat, and processed products (e.g. cheese, cookies, yoghurt, French fries) were not qualified by Bioland, and animal welfare has not yet been taken into account. Furthermore, qualified products were sold at farmers' markets or in local alternative food stores, while government regulations concerning organic agriculture did not yet exist.

Today, 50 years later, the description of Bioland's organic quality is different, and Bioland's modified principles now focus on animal welfare, the production of nutritious food, and environmental protection:

> operating farms using a circular production process;
> promoting soil fertility;
> keeping animals in a humane way;
> producing valuable food;
> promoting biological diversity;
> preserving natural resources;
> securing a future worth living for people.
> (Authors' translation)[7]

These principles are followed by many farms and horticultural organisations, while almost all German harvested foods and a broad variety of processed foods are qualified as organic by Bioland. Moreover, Bioland products are sold from various outlets, ranging from alternative farmers' markets to mass discounters. Finally, following the increasing popularity of organic product qualities, governments around the world have introduced regulations for organic production. In Germany, EU directives regulate the use of labels such as "organic," "bio," or "biological," meaning that only foods from certified farmers and food manufacturers can obtain these labels. However, it is important to note that Bioland's principles go beyond EU regulations in some cases, especially in the area of animal welfare.

Data collection and analysis

To explore which aspects of standardised organic quality have remained stable and which have changed in the longer term, we collected rich archival data produced by and about the case organisation (Bioland) (Ventresca and Mohr 2002). Table 1 provides an overview of these data. We read the journals (*bio gemüse Rundbrief* and *bio-land*) published by Bioland from 1974 to 2012, which were especially valuable for our study because, in addition to reporting on

---

[7] Bioland, Die sieben Prinzipien. https://www.bioland.de/sieben-prinzipien, accessed 22 April 2021.

current news, best practice, and agricultural research results, they systematically included documents such as annual reports and reports on Bioland's biannual general meetings. Journals published by competing organic standard-setters provided information about Bioland's qualification process from an outside perspective. These journals included ones by Biokreis e.V. (*Bio-Nachrichten*), Demeter e.V. (*Lebendige Erde* and *Demeter-Blätter*), and Naturland e.V. (*Naturland Magazin* and *Naturland Nachrichten*) Finally, we also read the newsletter of the umbrella organisation of German organic standard-setters published by the Arbeitsgemeinschaft ökologischer Landbau e.V. (AGÖL-Info) and two independent journals for scientists, activists, and practitioners concerned with organic farming (*IFOAM-Bulletin* and *Ökologie & Landbau*) published by the foundation Stiftung Ökologie & Landbau. Given that both insider and outsider journals were published periodically throughout our chosen research period, we had access to comprehensive insights from multiple perspectives, which enabled validation.

Description	Publisher	Official journal title (in German)	Time period	Number of pages read
Journals published by the focal standard-setter	Bioland e. V.	*bio gemüse Rundbrief*, from 1980 onwards *bio-land*	1974–2012	approx. 9,500
Journals published by competing standard-setter	Biokreis e. V.	*Bio-Nachrichten*	1983–2012	approx. 2,000
	Demeter e. V.	*Lebendige Erde* *Demeter-Blätter*	1946–2012 1962–1995	approx. 4,000
	Naturland e. V.	*Naturland Magazin*, from 2003 onwards *Naturland Nachrichten*	1994–2012	approx. 3,200
Journal published by the umbrella organisation of German organic standard-setters	Arbeitsgemeinschaft ökologischer Landbau e. V.	*AGÖL-Info*	1995–2000	approx. 100
Independent journal concerned with organic farming in Germany	Stiftung Ökologie & Landbau	*IFOAM-Bulletin* (1977–1988) *Ökologie & Landbau*	1977–2002	approx. 6,300
	Total pages read			approx. **25,200**

Table 1. Overview of Data Collection.

Source: Adapted by permission from Springer Nature Customer Service Center GmbH: Springer, Qualitaetsdarstellungen und ihre Stoerungen by Dombrowski and Arnold 2021.

By triangulating these data, we first reconstructed Bioland's historical development since the emergence of organic product quality in Germany in 1971, developing a detailed narrative (Dombrowski 2019). The narrative laid the groundwork for further analysis aimed at identifying shifts in the qualification process and their causes. To explore shifts in Bioland's standardised quality, we relied on the insight that the qualification process can best be observed in dynamic situations in which actors contest the meaning of a quality or are otherwise engaged in adapting, extending, or altering its meaning (Boltanski and Thévenot 2006; Helgesson and Kjellberg 2013; Bessy and Chateauraynaud 2014; Antal et al. 2015). Following this insight, we searched our data for quality-related disruption that refers to events, which were interpreted by Bioland as reasons for adjusting the qualification. This means that Bioland has always reacted in case of a disruption and taken measures to counter it. We did not consider disruptive events that were ignored by Bioland.

Using this approach, we identified ten major events in our narrative in which the qualification was disrupted, and as a result, Bioland decided to take active measures to address the disruption. Our analysis of the disruption was theoretically pre-informed (Baur 2009) because we assumed that the stability and dynamics of qualification were best examined with a focus on its spatial and temporal dimensions. Thus, after specifying the subject of disruption, we used our data to determine its origin and trigger. Specifically, we examined our data to specify the actors that disrupted the quality and its underlying process. In doing so, we caught a broad variety of actors who endangered quality and its underlying process, noting that disruption varied spatially. While much disruption originated with actors who were involved in the supply and distribution chain by buying and selling products, other disruption stemmed from actors who influenced the qualification process without exchanging products (e.g. social movement, media, and policymakers). As you will read further on, we use this distinction to systematically present our empirical results. However, once we had identified the actors and were as familiar as possible with our data, we examined in detail how Bioland responded to disruption in order to ward it off and maintain quality. By pooling these responses, we could identify long-term shifts in the organic qualification process that were necessary for organic quality to be maintained over time.

Dynamics of Bioland's organic quality, 1970–2012

We highlight the disruptions that were most discussed in the journals without presenting them chronologically. This means we first deal with the start of the qualification process, which began with the creation of Bioland. Then, we direct our attention to disruption triggered by actors in the production and distribution chains. For each disruption, we highlight the disruptive actors in advance, which were producer-members, food manufacturers, and supermarkets. Afterwards, we focus on disruption that stemmed from the actors who disturbed the qualification process without exchanging products, such as social movements, research institutes and policymakers, media, and competing standard-setters.

Initiating the qualification

In West Germany, the number of farms halved from 1.6 million in 1949 to 662,000 in 1986, while the average area under cultivation per farm more than doubled from 8 ha in 1949 to 18 ha in 1986 (Henning 1988). In the course of this transformation, German farmers aimed at specialised and efficient production, increasingly relying on costly inputs (e.g. farm machinery, chemically produced fertilisers, pesticides) to improve yields and animal production (Uekötter 2012).

In contrast, the producer-members who founded Bioland in 1971 ran small, unspecialised farms engaged in crop farming and dairy farming and could not compete economically with larger, specialised, and more efficient operations due to a lack of resources for costly inputs. The founding producer-members adopted and further developed the idea of organic farming, which was originally invented by Swiss farmers. The similarity between German and Swiss farms and the geographical proximity of the two countries facilitated the adoption and diffusion of organic quality, which merged the following two ideas: (1) protecting farmers' independence from the agrochemical industry and the state; and (2) adopting a business concept based on reducing production costs (by using agricultural techniques that do not need costly inputs) and selling grains and vegetables at an organic price premium. In practice, Bioland bundled its members' supply, which mostly consisted of grain, to sell it in larger quantities to grain mills and larger bakeries, while milk was sold conventionally to dairies without price premiums.

The foundation of the formal organisation, Bioland, was meant to fight structural changes in agriculture and promote organic agriculture in Germany, initiating the organic qualification process. At this time, Bioland primarily oriented itself to its producer-members. Until 1985, there were few formalised standards for defining the techniques that constituted Bioland quality, but producer-members were trained through workshops, regional groups, and publications on organic agriculture. Interestingly, Bioland's focus was on organic farming as a method of cost reduction rather than something that resulted in health benefits and enabled the sale of premium quality products to consumers, with the lack of residue from chemical fertilisers and pesticides highlighted as the main component of quality. The spatial reach of organic quality was quite narrow, as it was restricted to a limited number of small farms engaged in dairy and crop framing in southern Germany. The limited size of the qualification space was also reflected on the sales side, as producer-members sold their harvest directly at local farmers' markets and later founded specialised food stores. For consumers, the products were qualified by knowledgeable farmers and store clerks acting as "personal judgement devices" (Karpik 2010). At this early stage, products hardly ever had a standard form, and Bioland mainly invested in relationships, which is why organically qualified products had not yet reached the masses.

Disruptions triggered by actors in production and distribution chains

Producer-members. As Figure 1 shows, the popularity of Bioland's product quality grew tremendously among agricultural producers between 1978 and 2012. This membership growth disturbed the qualification in two ways. First, the growing number of producer-members implied a growing variety of farms, thus challenging the idea of the kinds of farms that were allowed to participate in the organic qualification process. By applying for membership, new producers whose farms did not structurally resemble those of the founding producer-members (due to higher levels of specialisation or cultivation of larger areas) challenged the qualification. In 1985, to respond to new applications, Bioland set new standards for defining the kinds of producers that were eligible (or not) for membership. In a speech introducing the new standard, the Bioland managing director at the time stated the following:

> They [the new rules prescribed in the standard] make clear what environmentally sustainable contemporary agriculture looks like, and they make it clear that this method cannot be bent to the needs of every farm, no matter how "structurally degraded" it may be. (Speech by Bioland managing director, 1985, authors' translation)[8]

More specifically, Bioland introduced new rules that made membership impossible for highly specialised factory farms, which hindered the latter from establishing collaborative relationships with Bioland and its product quality.

---

[8] P. Grosch. 1986. "Entwicklungsbericht Bundesversammlung 1986," *Bio-land* (2): 3.

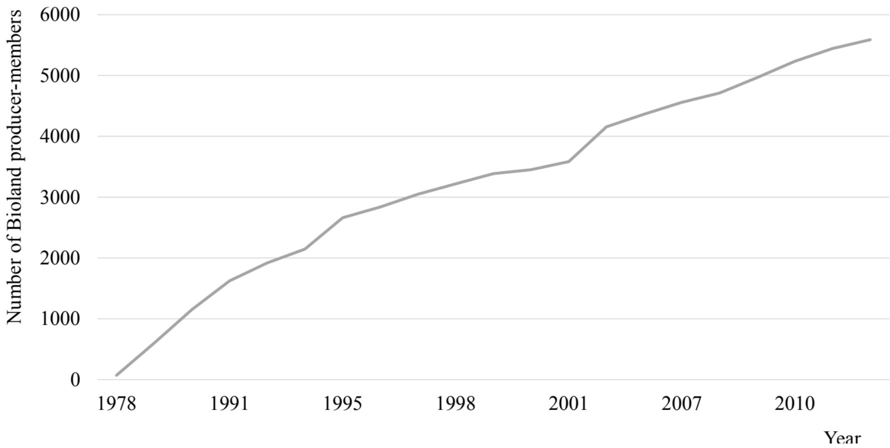

Figure 1. Bioland's membership development figures 1978-2012.
Source: Authors' own data collected from membership figures in various issues of the journals listed in Table 1 and the annual statistics of the Bundesverband Oekologische Landwirtschaft.

Second, growing membership was accompanied by demands to extend the organic quality and its underlying process to new products, which caused far-reaching disruption – for example, in cases of gardening and beekeeping. Unlike for harvesting grains, Bioland had not established criteria for gardening or producing honey organically. In the case of gardening, the horticultural operations that disrupted the qualification were run by producer-members. Given that they rarely engaged in animal husbandry, horticultural farms could not produce the manure necessary for fertilising their patches and needed to buy fertilisers from external sources. By doing so, they were not conforming to the main principle of Bioland farming, namely circular farming. Moreover, they used plastic foil to protect plants and heat greenhouses, which was not compatible with Bioland's ideal of environmentally friendly production. Therefore, to enable the qualification of products from horticultural farms, Bioland created a new intra-organisational division, a standard development commission, to adapt Bioland's quality ideals to horticulture. The commission, composed of gardeners and Bioland's staff, defined which practices constituted gardening according to Bioland principles. For example, horticultural farms were allowed to buy fertiliser only if they paid special attention to the quality of the manure (avoiding manure from factory farms), recycled plastic foil, and only achieved a moderate extension of the cultivation time through heating.[9]

In the case of honey production, the qualification was disrupted by beekeepers. They could not completely avoid chemical fertilisers and

[9] Bio-land. 1988. "Bioland-Erzeugungsrichtlinien. Bestimmungen für den Gemüsebau," *Bio-land* (1): 35–36.

pesticides (as required by Bioland) because bees collected pollen from fields on adjacent farmlands that could have been treated using both. Therefore, the honey could not be qualified as organic, which is why Bioland's beekeepers started debating what Bioland's organic beekeeping should entail. Finally, at Bioland's general assembly, they proposed an annexe to the Bioland standard that prohibited the treatment of beehives with specific chemical preservatives and defined the practices of natural beekeeping that emphasised animal welfare as a major component.[10] Overall, Bioland reacted to disruption triggered by producer-members by forming new intra-organisational divisions and setting new rules, which stabilised and expanded the qualification space at different points in time. Similar developments took place regarding viticulture and pig farming, and we will see that setting rules and creating organisational divisions were important changes for maintaining standardised product quality.

Producer-members and food manufacturers. Besides demands by food manufacturers, the growing number and variety of producer-members led to other disruption. Beginning in the late 1980s, dairy farmers who could not rely on direct sales demanded new marketing opportunities for processed products to attain additional sales channels, expecting an increase in the price of organic milk. At the same time, food manufacturers wanted to expand their organic qualification to new products, such as fruit yoghurt, liquors, or wheat rolls. However, as Bioland defined its product quality as healthy and natural, it was a highly contested question within the standard-setter whether products that required extensive industrial manipulation (e.g. white flour, white sugar) could be qualified as organic by Bioland. A board member described the controversy as follows:

> Dairy farmers are demanding that UHT [ultra-high-temperature processing] milk should be approved, while others are threatening to leave the organisation if this is done. There are similar discussions for almost every product. (Statement, board member, Bioland, 1991, authors' translation)[11]

After intense internal conflict, Bioland opted to link its product quality to the well-established concept of whole food nutrition, which claimed that food is healthier if it is less processed (Koerber 2012). Producer-members advocating for formal expansion of the qualification space to include whole food products received external support. More specifically, the association of whole food manufacturers and

[10] R. Geist. 1991. "Bioland-Bienenhaltung," Bio-land (2): 36–37; U. Schumacher 1993. "Bioland muss ein Zeichen setzen," *Bio-land* (1): 39–41.

[11] Christoph Ziechaus-Hartelt. 1991. "Bioland – Ein Verband entwickelt sich," *Bio-land* (2), 13–14.

distributors claimed an expansion, as they aimed to differentiate their offerings by qualifying them as healthier than other organic products sold in supermarkets. Accordingly, highly processed ingredients and products (e.g. white sugar, UHT milk) were not allowed according to Bioland's rules; however, besides raw agricultural products, Bioland's qualification included products processed according to the concept of whole-food nutrition. By adopting the concept of whole-food nutrition, Bioland could contribute to stabilising its product quality while extending its qualification to processed food.

Producer-members and supermarkets. Producer-members disrupted the qualification not only in terms of who may produce organically, and which products may be qualified as organic (as discussion of previous disruption has shown), but also in terms of where qualified products could be sold. Our data showed that supermarkets played an important role in this case, although initially the disruption was triggered by producer-members from remote areas.

Originally, qualification relied on personal judgement devices, such as the trustworthiness and knowledge of producer-members at local markets and farm stores. Given that this qualification was not financially viable for the producer-members in remote regions, Bioland sought to develop marketing structures that would be more beneficial to its producer-members. More specifically, Bioland wanted to develop impersonal, generalised judgement devices that could extend the qualification to more anonymous retail locations (supermarkets) by informing consumers remotely about organic quality. To accomplish this, Bioland developed a label (a visual symbol that signified compliance with Bioland standards), and in so doing mobilised a tool that, while responding to producer-members' wishes, built new relationships with supermarkets, where products are primarily qualified by their packaging and display on the shelves (Cochoy 2007). Given that German supermarkets had been developing their own organic brands (visual symbols that were not linked to compliance with independent standards) since the early 1980s, the foods qualified by Bioland were displayed next to supermarkets' "pseudo-organic" products. Some of these pseudo-organic products had astonishingly similar labels – for example, in one extreme case, products bore the label BIOLAN (compared with the word Bioland, only the letter "d" was missing). In court, Bioland's actions against this imitation were unsuccessful.

Consequently, expansion of Bioland's qualification space for supermarkets resulted in another disruption, namely, competition, and confusion with pseudo-organic products. Given supermarkets' power in food chains and the definition of quality standards (Ponte and Gibbon 2005), maintaining organic quality was challenging for Bioland at this time. At first, Bioland intended to supply only those supermarkets that were willing to undertake additional qualification action, such as special training for shop clerks and marketing coordination with Bioland. However, few supermarkets accepted these conditions, which is why Bioland attempted to ensure the organic quality of its products by restricting sales to selected supermarket chains regarded as premium food retailers, such as the Edeka cooperative. Consequently, Bioland invested in the appearance of its label, one that did not require the support of or collaboration with supermarkets. Every couple of years, marketing specialists would redesign the Bioland label and develop additional marketing tools, such as brochures and leaflets.[12] These recurring activities were intended to visibly distinguish Bioland products from other (pseudo-)organic products and to adapt the label's design to changing marketing trends.

Disruptions from outside the production and distribution chain

Social movements. In line with the insight that social movement actors constitute and challenge markets (Weber et al. 2008; King and Pearce 2010), our data showed that actors from the animal welfare movement disrupted Bioland's qualification. The movement, largely organised by agricultural scientists and veterinarians, disrupted Bioland's qualification by problematising ignorance regarding animal welfare in organic agriculture. Until the mid-1980s, Bioland's producer-members and staff rarely considered animal welfare in their internal debates and discourse, and animal-unfriendly practices, such as indoor dairy farming that involved animals being tied down in the stables, were widespread. Against this background, various groups from the animal welfare movement lobbied for the adoption of the kind of animal husbandry that would respect the natural needs of farm animals. In particular, the leading agricultural scientist of the Naturland Association defended the more natural forms of animal husbandry, especially in relation to cows. His claim was supported by scientists who developed measurable animal welfare criteria. A leading document described the animal welfare situation as follows:

> It is necessary to take all measures that support animal-friendly, environmentally compatible, farmer-oriented, but also quality-oriented livestock farming and that do justice not only to the economic importance of the animal but also to the ethical responsibility of humans. How can these requirements be met? Organisations should tighten guidelines for animal

---

[12] o.V. 1992. "Bioland – Marke mit Zukunft," *Bio-land* (4): 12–23; R. Langerbein 1994. "Bioland stellt Weichen für zukünftige Markenpolitik," *Bio-land* (4): 32–33; age. 2004. "Überzeugende Arbeit," *Bio-land* (3): 46; o. V. 2010. "Neues Bioland-Zeichen," *Bio-land* (10): 39.

> husbandry and optimise controls. (Contribution by an employee of the German Animal Welfare Association, 1993, authors' translation)[13]

To some extent, this disruption by the animal welfare movement was positively received by Bioland. In particular, producer-members who saw animal welfare as offering new marketing and differentiation opportunities, as well as possibilities for premium pricing, welcomed the challenge. As a result, Bioland gradually included rules for animal husbandry in its standards, extending the meaning of Bioland's quality even further. However, to avoid losing producer-members who could not meet the new criteria, Bioland introduced transitional rules and exemptions.[14] Today, animal welfare is one of the core elements of the organic qualification led by Bioland.

Research institutes and policymakers. The disruption brought about by actors from the field of German agricultural policy have changed fundamentally over the course of the studied period. Two episodes of disruption were particularly important. In the first, in 1983, research institutes of the German Länder agriculture administration attacked organic product quality so vehemently that Bioland was almost wiped out. During this period, the German agriculture policy field was structured by close ties between the state, the federal agricultural administration, state-funded agricultural research organisations, and the dominant German farmers' association (Rieger 2007). These actors shared strong beliefs in increasing the productivity of farms via the chemicalisation, mechanisation, and specialisation of farms, meaning that alternative production methods, such as organic agriculture, were viewed very negatively. In this context, research institutes of the German Länder agriculture administration challenged the legality of organic qualification in Germany. Specifically, research institutes questioned the alleged "pesticide-free" nature of organic food. Their accusations were based on a scientific study that found that organically produced foods contained approximately the same level of residues of chemical pesticides as conventional agricultural products. These findings could be explained by the drift of pesticides and the fact that the study included a large number of "pseudo-organic" products in its sample.[15] However, based on these findings, research institutes suggested that the organic qualification violated the German food law of 1974, which allowed the pesticide-free designation to be attributed

---

[13] S. Hencke. 1993. "Im Visier: Tierhaltung auf dem Bio-Hof," *Bio-land* (3): 26–27.

[14] Bioland. 1989. "Die neuen Bioland-Richtlinien," *Bio-land* (4): 36; S. Braun and H. Hinrich 1989. "Leserbrief zu den neuen Bioland-Richtlinien und Antwort," *Bio-land* (4): 41–42.

[15] U. Ahrenhöfer. 1984. "Auszüge aus der kritischen Stellungnahme von Vertretern des ökologischen Landbaus zur VDLUFAS-Studie," *IFOAM-Bulletin* (1): 13–14.

only to products that had no traces of pesticides whatsoever.[16] Given that it was impossible for a product to be completely "pesticide-free" due to drift from adjacent fields that had been treated with pesticides, Bioland (together with other organic standard-setters) responded to this attack by jointly setting a common meta-standard to defend organic quality. In the preamble, the joint document stated the following:

> Consumers' understanding of general environmental pollution is negatively affected by the idea that , 1984, authors' translation)[17]

The quotation points to an interesting change in the qualification away from the material product and towards the production process that underlies the product. Previously, the qualification process focused on the product. A food item was designated as pesticide-free, with health and naturalness being important elements in the construction of organic quality. Due to the attack by research institutes, focus shifted towards production methods. Consequently, the production process was now designated pesticide-free, with the environment developing into an important element in the qualification:

> Organic agriculture and horticulture are land cultivation methods that aim to sustainably and consistently care for the natural resources entrusted in accordance with the interrelationships and interdependencies of the natural order of life. (Bioland standard, 1985, authors' translation)[18]

This shift from product to process, which was necessary to maintaining organic product quality, required the introduction of new organisational forms. In the 1980s, driven by a general trend towards independent third-party certification (Loconto and Busch 2010; Arnold and Hasse 2016; Gustafsson 2020), Bioland started to enforce its qualification via new forms of control by creating a new formal

---

[16] H. Vetter, W. Kampe, and K. Ranfft. 1983. "Qualität pflanzlicher Nahrungsmittel. Ergebnisse einer 3jährigen Vergleichsuntersuchung an Gemüse, Obst und Brot des modernen und alternativen Warenangebots," *VDLUFA-Schriftenreihe* (7).

[17] Stiftung ökologischer Landbau; Arbeitsgemeinschaft naturnaher Obst-, Gemüse- und Feldfruchtanbau e.V. (ANOG), Biokreis Ostbayern e.V.; Fördergemeinschaft organisch-biologischer Landbau e.V. (Bioland); Forschungsring biologisch-dynamische Wirtschaftsweise (Demeter); Verband für naturgemäßen Landbau (Naturland) 1984. "Rahmenrichtlinie für die Erzeugung landwirtschaftlicher Produkte aus ökologischem Landbau in der Bundesrepublik Deutschland," *IFOAM-Bulletin* (2): 15–20.

[18] Fördergemeinschaft organisch-biologischer Land- und Gartenbau. 1985. "Erzeugungsrichtlinien der Fördergemeinschaft organisch-biologischer Land- und Gartenbau e. V. 3. Lesung," *Bio-land* (5): 5.

organisational unit, notably an independent certification commission (Anerkennungskommission). While control was previously maintained informally by producer-members who visited each other's farms, organising controls were now the prerogative of the certification commission, whose procedures were formally accredited by the association of German organic organisations (Arbeitsgemeinschaft Ökologischer Landbau, AGÖL). This disruption underlines Bioland's tremendous organisational effort (the creation of new organisational units and the introduction of organisational elements, such as independent certification and accreditation procedures) in addition to shifts in focus (from product to process) to maintain its organic product quality. These organisational efforts were supported by the European Community (EC), which, in 1991, decided to grant additional subsidies to organic farms (to reduce agricultural overproduction) and started regulating the use of labels such as "organic" (Lampkin et al. 1999). However, growing acceptance of organic quality did not protect Bioland from further disruption, as Bioland was suddenly confronted with new policies promoting organic agriculture and legally protecting organic qualification. This brings us to the second disruption that originated in the field of agriculture policy.

In 1991, the EC adopted "Regulation (EEC) No. 2092/91 on organic production of agricultural products and indications referring thereto to agricultural products and foodstuff." While the regulation outlawed "pseudo-organic products," it transferred organic product quality to what Thévenot (1984) described as the "state form," which could be used by anyone. By developing this regulation related to organic product quality, policymakers disrupted Bioland's qualification because it was no longer clear what Bioland's quality actually stood for and what distinguished it from the qualification led by the EC. This disruption intensified in 2001, when the German government introduced a voluntary state label for organic foods (named *Biosiegel*) that met the requirements of EC regulations. From then on, supermarkets, and other distributors could rely on well-known regulations for organic qualification without having to cooperate with Bioland. In short, Bioland was in danger of becoming obsolete:

> Since the EC Regulation "Organic Farming" has come into force, competition in the organic market has intensified. The range of organic producers [...] has grown considerably, and food retailers are entering the market with their own brands [...] It is our task to emphasise Bioland as a trademark with special qualities in the future and to assert it on the market. (Report by Bioland's executive on the new label policy, 1994, authors' translation)[19]

[19] R. Langerbein. 1994. "Bioland stellt die Weichen für die zukünftige Markenpolitik," *Bio-land* (4): 33.

Once again, Bioland responded by setting new rules, although, unlike during previous disruption, these rules were not aimed at expanding the qualification space by establishing new relationships. This time, the rules served as a means of distinguishing Bioland quality from the state form and the qualities of other organic standard-setters. More specifically, Bioland wanted to make its standards stricter and for its product quality to be "more" organic than its competitors' organic qualities. For example, Bioland set ambitious rules for animal husbandry, giving animals more attention than the official organic regulations. In so doing, Bioland invested in a wide variety of promotional materials (e.g. brochures, websites, posters) to discursively demonstrate the superiority of its organic product quality.

Media. The media disrupted Bioland's qualification by informing the public of malpractice among German organic food producers. The most prominent disruption was the so-called nitrofen scandal. In 2002, state authorities discovered residue of the illegal pesticide nitrofen in animal feed used on organic farms, prompting media headlines such as "Trust Gambled Away,"[20] "Innocence Lost,"[21] or "Organic Poultry Picked Contaminated Grains."[22] In the articles, journalists scandalised the nitrofen detection, which fundamentally threatened the legitimacy of organic product quality. Although scientific investigations later revealed that the contamination occurred in a storage facility where not only organic products were stored and that Bioland-qualified products were not affected, Bioland reacted immediately. Bioland invested in transparency by establishing a strict separation of Bioland production chains from other (conventional and organic) food production chains and developing a commodity traceability system. These organisational endeavours were intended to limit the risk posed by future legitimacy threats and resulted in further expansion of the meaning, which now also included food safety.

Other organic standard-setters. While Bioland cooperated with other standard-setters in defending organic product quality from attacks by agricultural research institutes, these other formal organisations also acted as sources of disruption. We know that organic standard-setters are in competition with one another (Reinecke et al. 2012; Fouilleux and Loconto 2017), and it was this competition for producer-members that disturbed Bioland's qualification in the late 1990s. Until 1999,

---

[20] C. Merey. 2002. "Vertrauen verspielt." *Frankfurter Neuer Presse*, 27 May 2002.

[21] Nordkurier. 2002. "Verlorene Unschuld." 28 May 2002.

[22] Hamburger Morgenpost. 2002. "Bio-Geflügel pickte verseuchte Körner." 25 May 2002.

Bioland had not allowed the use of copper in potato production, while other organic standard-setters (except for Demeter) all allowed copper as "natural" protection against fungal diseases. Bioland took the position that copper was harmful to other microorganisms in the soil and should therefore be banned. However, when a fungal disease caused severe decline in a potato harvest, the producer-members threatened to leave Bioland and become members of competing organic standard-setters that allowed the use of copper. To safeguard its relationship with producer-members, Bioland decided to reclassify copper:

> With the extreme weather conditions [like last year's summer], many would have asked themselves, "Do I use copper, or do I move away from Bioland?" In addition, there is the high dependency of many farms on potatoes and the competition with other associations [i.e. standard-setters], which, with the exception of Demeter, allow copper to be used. (Report from Bioland General Assembly, 1999, authors' translation)[23]

This disruption illustrates that Bioland's product quality was affected by other relevant standard-setters, especially when producer-members (on which Bioland was depending) used the available alternatives as leverage. However, Bioland found the means to stabilise its product quality by once again revising its rules.

Long-term shifts in organic qualification

A summary look at our empirical results shows that Bioland reacted to disruptions triggered by actors located along the production and distribution chain (producer-members, processors, supermarkets) as well as outside it (media, social movements, competitors, policymakers, and research institutes). In so doing, Bioland reacted swiftly to disruptions caused by media, social movements, or research institutes and did not wait until actors who were imperative to the production and trade of qualified products (e.g. producers, distributors) challenged the process. Bioland as a standard-setter thus operated in a manner like producers who observe the signals of other actors in the field because they are unable to know in advance how consumers will react to their production decisions (White 1981).

Given that more and more actors became involved in qualification, the qualification space constantly expanded. All the stabilising efforts that we identified were pragmatic responses by Bioland to fix momentary disruption. At the same time, responses shaped the unfolding of that process by creating new relationships (e.g. Bioland's relationships with beekeepers or supermarkets), eliciting new decisions about the standardisation system (e.g. revising and adding rules,

---

[23] Age. 1999. "Pragmatisch, aber nicht anpasslerisch," *Bio-land* (1): 42.

creating new forms of control), or initiating shifts in meaning (e.g. inclusion of food safety, emphasis on animal welfare). In other words, singular disruption to which Bioland has reacted has affected the quality and its underlying process in a long-lasting way. For example, the nitrofen scandal had a lasting impact on Bioland's qualification. As we have shown, Bioland reacted to the scandal by strictly separating Bioland production chains from other conventional and organic food production chains and by developing a traceability system. Implementation of this latter instrument was a pragmatic response taken by Bioland at a particular point in time, but it made food safety a central part of further qualification, an outcome that was highly appreciated by producers and supermarkets. By emphasising food safety, Bioland was able to successfully distinguish its organic quality from those of its competitors. The empirical study thus shows that standardisation is a continuous, dynamic process that does not lead to closure (Loconto and Demortain 2017) while at the same time unveiling four main shifts in the unclosed organic qualification process. We discuss these shifts separately, although they are intertwined and together make up the dynamics of standardised quality.

The first dynamic is reflected in a shift in meaning of organic quality from economic self-help to an environmental project that emphasises the value of animal welfare and safety. This shift is relevant for those who are interested in the specific content of organic quality, such as rural sociologists or food policymakers. It highlights the contingency of the meaning of a supposedly stable quality – a contingency of which the actors involved in momentary qualification are seldom aware. Rhetorically asked, which consumer, producer, or supermarket today knows the former meanings of organic quality and/or would assume that organic quality has something to do with self-help? Identifying long-term shifts in the meaning of quality adds to the literature on qualification as it sheds light on the little-studied long-term development of quality (Musselin and Paradeise 2005), demonstrating that quality is not only contested and changeable at certain moments but also over time – even when it is supported by standards.

In contrast to the first, organic-specific change, the second is more abstract, referring to the shift in focus of the qualification from product to process. This shift was exemplified by the value "pesticide-free," which Bioland first attributed to food products but later to the production process. This change in standard-setting is the result of a criterion bias, as the standard-setter qualified what is easier to measure and evaluate (Singer 1996: 212 f.) – and these are processes rather than the material properties of food. Consequently, this article and the "quality turn" (Allaire and Sylvander 1997; Murdoch et al. 2000) are not only about new product quality, but above all about new process quality. This pioritisation of standardising processes (and not products) explains why consumers and the media are often surprised when they discover that organic food is not necessarily healthier, as standards focus on processes rather than on the food products themselves (EatSmarter! 2018). For scholars of valuation and qualification this shift is of interest because it suggests that product quality is justified in the longer term by processes rather than by material properties.

The focus on process quality manifests in the fact that during long-term qualification, Bioland increasingly organised production and trade processes, deploying rules, standards, or control and traceability systems to ward off disruption. Even though organic qualification started with the creation of formal organisation (Bioland) and new specialised organisational divisions were added later, Bioland stabilised quality continuously with the use of various organisational elements (rules, standards, and controls). As a result, we can observe a third shift from formal organisation to an accumulation of organisational elements that operate outside the boundaries of Bioland between the standard-setter and the producer-members, the supermarkets, retailers, and other involved parties. This shift discovered in the context of organic qualification reflects the accumulation of standards and control (e.g. Djelic and den Hond 2013; Gustafsson 2020) as well as fundamental change in the organisational world, where new, less bureaucratic and more flexible forms of organisations, distinct from the rather classic, formal organisation, are gaining societal relevance (e.g. Ahrne and Brunsson 2011, 2019; Bartley et al. 2019). These new organisational elements that come into play outside formal organisations require attention if we are to study and better understand the nexus between organisation, on the one hand, and valuation and qualification, on the other (Hauge 2016; Meier and Peetz 2021).

Finally, our empirical study shows that Bioland has invested not only in its organisation but also in relationships to maintain its quality. While Bioland initially focused on its relationship with producer-members over time, the standard-setter has increasingly responded to disruption triggered by actors operating outside the production and commodity chain (e.g. media, state, and social movement), building meaningful relationships with them. This shift towards multifaceted relationships has been central to the stabilisation (and paradoxically also the dynamics) of standardised quality. Thus far, research on food qualification has attributed the importance of relationality exclusively to singular niche products. This bias is evident in the study by Varga (2019), who argued that alternative food networks – characterised by strong civil society embeddedness – qualify food based on relations between farmers, their suppliers, and workers and customers. In contrast, and following his argumentation, standard-setters, such as Bioland do not rely on relationships but instead use product-oriented standards to realise their large-scale qualification in conventional markets. This dichotomous distinction between standards and relations may result in the misleading conclusion that relations are irrelevant to the construction of standardised quality. Contrary to this dichotomous understanding, this article provides empirical evidence that the key insight of valuation studies, namely that processes of quality and value construction are always relational (Heinich 2020; Kuipers and Franssen 2020), also applies to standardised quality and products. However, to discover that controversies and relationships matter in standardised qualification, the analytical focus must be expanded spatially and temporally.

Conclusion

If we take the diversity and extent of the identified changes seriously (shifts in meaning, qualification focus, organisation, and relationships), we must ask what has not shifted during qualification. This brings us back to our original question, which asked both what has changed and what has stayed the same. The placative, pointed answer would probably be that apart from the terms "organic" and "Bioland," hardly anything has remained the same. In this sense, we confirm the thesis that one should understand the "stability of standards [...] as the result of underlying dynamic processes" (Brunsson et al. 2012: 627). This means that the multiple shifts we discovered in qualification are a necessary condition for organic quality having existed in German food markets for around 30 years. Against this background, we conclude that not only is every good category a living one (Bowker and Star 1999), but also standardised qualification must be dynamic and changeable if it is to be stably relevant in markets.

Acknowledgements: We thank members of our working group "Organisation and Valuation", Andrea Mennicken and two anonymous reviewers for their very helpful and constructive comments. We have benefited from the Spring Workshop of the Section of the Sociology of Organizations of the German Sociological Association 2018 in Bremen, the 39th Congress of the German Sociological Association 2018 in Göttingen, as well as the Economic Sociology Conference of the European Sociological Association 2018 in Konstanz.

References

Ahrne, Göran, and Nils Brunsson. 2011. "Organization Outside Organizations: The Significance of Partial Organization." *Organization* 18(1): 83–104.

Ahrne, Göran, and Nils Brunsson (Eds.). 2019. *Organization Outside Organization: The Abundance of Partial Organization in Social Life*. Cambridge: Cambridge University Press.

Ahrne, Göran, Patrik Aspers, and Nils Brunsson. 2015. "The Organization of Markets." *Organization Studies 36*(1): 7–27.

Allaire, Gilles, and Bertil Sylvander. 1997. "Qualité spécifique et systèmes d'innovation territoriale." *Cahiers d'économie et de sociologie rurales (CESR)* 44: 29–59.

Antal, Ariane Berthoin, Michael Hutter, and David Stark. 2015. *Moments of Valuation: Exploring Sites of Dissonance*. Oxford: Oxford University Press.

Appadurai, Arjun (Ed.). 1986. *The Social Life of Things: Commodities in Cultural Perspective*. Cambridge: Cambridge University Press.

Arnold, Nadine. 2017. *Standardisierungsdynamiken im Fairen Handel. Die Entwicklung des Schweizer Fair Trade Feldes und dessen Standards*. Wiesbaden: Springer VS.

Arnold, Nadine. 2022. "Accountability in Transnational Governance: The Partial Organization of Voluntary Sustainability Standards in Long-Term Account-Giving." *Regulation & Governance* 16(2): 375–391.

Arnold, Nadine, and Raimund Hasse. 2016. "Die Organisation von Wertzuschreibungen: Zur Bedeutung von Drittparteien für die Signalisierung moralischer Qualitäten in Märkten." *Berliner Journal für Soziologie* 26(3–4): 329–351.

Arnold, Nadine, and Allison Loconto. 2021. "Serving Magically Perfect Fruit Globally: Local Nesting in Translating Multiple Standards." *Organization Studies* 42(2): 327–349.

Bartley, Tim, Matthew Soener, and Carl Gershenson. 2019. "Power at a Distance: Organizational Power across Boundaries." *Sociology Compass* 13(10): e12737.

Baur, Nina. 2009. "Problems of Linking Theory and Data in Historical Sociology and Longitudinal Research." *Historical Social Research [Historische Sozialforschung]* 34(1): 7–21.

Beckert, Jens, and Patrik Aspers (Eds.). 2011. *The Worth of Goods: Valuation and Pricing in the Economy.* Oxford: Oxford University Press.

Beckert, Jens, and Christine Musselin. 2013. "Introduction." In *Constructing Quality: The Classification of Goods in the Economy,* edited by Jens Beckert and Christine Musselin, 1–30. Oxford: Oxford University Press.

Bessy, Christian, and Francis Chateauraynaud. 2014. *Experts et faussaires : Pour une sociologie de la perception*. Paris: Pétra.

Bessy, Christian, and Pierre-Marie Chauvin. 2013. "The Power of Market Intermediaries: From Information to Valuation Processes." *Valuation Studies* 1(1): 83–117.

Boltanski, Luc, and Arnaud Esquerre. 2020. *Enrichment: A Critique of Commodities*. Cambridge: Polity Press.

Boltanski, Luc, and Laurent Thévenot. 1983. "Finding One's Way in Social Space: A Study Based on Games." *Social Science Information* 22(4–5): 631–680.

Boltanski, Luc, and Laurent Thévenot. 2006. *On Justification: Economies of Worth*. Princeton, NJ: Princeton University Press.

Bowker, Geoffrey C., and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

Brunsson, Nils, and Bengt Jacobsson (Eds.). 2000. *A World of Standards*. Oxford: Oxford University Press.

Brunsson, Nils, Andreas Rasche, and David Seidl. 2012. "The Dynamics of Standardization: Three Perspectives on Standards in Organization Studies." *Organization Studies* 33(5–6): 613–632.

Busch, Lawrence. 2011. *Standards: Recipes for Reality.* Cambridge, MA: MIT Press.

Callon, Michel, Cécile Méadel, and Vololona Rabeharisoa. 2002. "The Economy of Qualities." *Economy and Society* 31(2): 194–217.

Cochoy, Franck. 2007. "A Sociology of Market-Things: On Tending the Garden of Choices in Mass Retailing." *The Sociological Review* 55(2_suppl): 109–129.

Conford, Philipp. 2001. *The Origins of the Organic Movement*. Edinburgh: Floris Books.

DeSoucey, Michaela. 2018. *Contested Tastes: Foie Gras and the Politics of Food*. Princeton, NJ: Princeton University Press.

Djelic, Marie-Laure, and Frank den Hond. 2013. "Introduction: Multiplicity and Plurality in the World of Standards." *Business and Politics* 16(1): 67–77.

Dobeson, Alexander, and Sebastian Kohl. 2020. "Why Durability Matters? Towards a Comparative Economic Sociology of Market Organization." *European Journal of Sociology [Archives Européennes de Sociologie]* 61(1): 33–64.

Dombrowski, Simon. 2019. *Die organisierte Hand des Marktes. Verbandliche Koordination auf den deutschen Märkten für biologische Lebensmittel.* Wiesbaden: Springer VS.

EatSmarter!. 2018. "Neue Test-Ergebnisse: Warum BIO Nicht Immer Besser Ist." https://eatsmarter.de/gesund-leben/gesundheit/neue-test-ergebnisse-warum-bio-besser, accessed 24 December 2021.

Eymard-Duvernay, François. 1989. "Conventions de qualité et formes de coordination." *Revue économique* 40(2): 329–360.

FAO (Food and Agriculture Organisation). 2014. "Impact of International Voluntary Standards on Smallholders' Market Participation in Developing Countries: A Review of the Literature." Rome: Food and Agriculture Organization of the United Nations.

Fouilleux, Eve, and Allison Loconto. 2017. "Dans les coulisses des labels: régulation tripartite et marchés imbriqués." *Revue française de sociologie* 58(3): 501–531.

Garcia-Parpet, Marie-France. 2007. "The Social Construction of a Perfect Market: The Strawberry Auction at Fontaines-en-Sologne." In *Do Economists Make Markets?*, edited by Donald MacKenzie, Fabian Muniesa, and Lucia Siu, 20–53. Princeton, NJ: Princeton University Press.

Garcia-Parpet, Marie-France. 2011. "Symbolic Value and the Establishment of Prices: Globalization of the Wine Market." In *The Worth of Goods: Valuation and Pricing in the Economy,* edited by Jens Beckert and Patrik Aspers, 131–154. Oxford: Oxford University Press.

Gustafsson, Ingrid. 2020. *How Standards Rule the World: The Construction of a Global Control Regime.* Cheltenham: Edward Elgar.

Hallström, Kristina Tamm, and Magnus Boström. 2010. *Transnational Multi-Stakeholder Standardization: Organizing Fragile Non-State Authority.* Cheltenham; Northampton, MA: Edward Elgar.

Hauge, Amalie Martinus. 2016. "The Organizational Valuation of Valuation Devices: Putting Lean Whiteboard Management to Work in a Hospital Department." *Valuation Studies* 4(2): 125–151.

Heinich, Nathalie. 2020. "A Pragmatic Redefinition of Value(s): Toward a General Model of Valuation." *Theory, Culture & Society* 37(5): 75–94.

Helgesson, Claes-Fredrik, and Hans Kjellberg. 2013. "Introduction: Values and Valuations in Market Practice." *Journal of Cultural Economy* 6(4): 361–369.

Henning, Friedrich Wilhelm. 1988. *Landwirtschaft und Ländliche Gesellschaft in Deutschland.* Paderborn: Schöningh.

Higgins, Vaughan, and Wendy Larner. 2010. *Calculating the Social: Standards and the Reconfiguration of Governing.* Basingstoke: Palgrave Macmillan.

Karpik, Lucien. 2010. *Valuing the Unique: The Economics of Singularities.* Princeton, NJ: Princeton University Press.

King, Brayden G., and Nicholas A. Pearce. 2010. "The Contentiousness of Markets: Politics, Social Movements, and Institutional Change in Markets." *Annual Review of Sociology* 36(1): 249–267.

Koerber, Karl W. 2012. *Vollwert-Ernährung Konzeption einer zeitgemässen und nachhaltigen Ernährung.* Stuttgart: Karl FHaug.

Kuipers, Giselinde, and Thomas Franssen. 2020. "Qualification." In *Pragmatic Inquiry. Critical Concepts for Social Sciences,* edited by John R. Bowen, Nik Dodier, Jan Willem Duyvendak, and Anita Hardon, 143–168. London: Routledge.

Lampkin, Nicolas, Carolyn Foster, Susanne Padel, and Peter Midmore. 1999. *The Policy and Regulatory Environment for Organic Farming in Europe.* Stuttgart: Universität Hohenheim, Institut für Landwirtschaftliche Betriebslehre.

Loconto, Allison. 2017. "Models of Assurance: Diversity and Standardization of Modes of Intermediation." *The ANNALS of the American Academy of Political and Social Science* 670(1): 112–132.

Loconto, Allison, and Lawrence Busch. 2010. "Standards, Techno-Economic Networks, and Playing Fields: Performing the Global Market Economy." *Review of International Political Economy* 17(3): 507–536.

Loconto, Allison, and David Demortain. 2017. "Standardization as Spaces of Diversity." *Engaging Science, Technology, and Society* 3: 382–392.

Meier, Frank, and Thorsten Peetz (Eds.). 2021. *Organisation und Bewertung*. Organisationssoziologie. Wiesbaden: VS Verlag für Sozialwissenschaften.

Murdoch, Jonathan, Terry Marsden, and Jo Banks. 2000. "Quality, Nature, and Embeddedness: Some Theoretical Considerations in the Context of the Food Sector*." *Economic Geography* 76(2): 107–125.

Musselin, Christine, and Catherine Paradeise. 2005. "Quality: A Debate." *Sociologie du travail* 47: 89–123.

Ponte, Stefano, and Peter Gibbon. 2005. "Quality Standards, Conventions and the Governance of Global Value Chains." *Economy and Society* 34(1): 1–31.

Reinecke, Juliane, Stephan Manning, and Oliver von Hagen. 2012. "The Emergence of a Standards Market: Multiplicity of Sustainability Standards in the Global Coffee Industry." *Organization Studies* 33(5–6): 791–814.

Rieger, Elmar. 2007. "Bauernverbände. Agrarische Interessenpolitik, institutionelle Ordnung und politischer Wettbewerb." In *Interessenverbände in Deutschland,* edited by Thomas von Winter and Ulrich Willems, 294–315. Wiesbaden: VS Verlag für Sozialwissenschaften.

Singer, Benjamin D. 1996. "Towards a Sociology of Standards: Problems of a Criterial Society." *The Canadian Journal of Sociology [Cahiers Canadiens de Sociologie]* 21(2): 203–221.

Taylor, Frederick. 1911 [2016]). *The Principles of Scientific Management*. New York: Cosimo Classics.

Thévenot, Laurent. 1984. "Rules and Implements: Investment in Forms." *Social Science Information* 23(1): 1–45.

Thévenot, Laurent. 2015. "Certifying the World: Power Infrastructures and Practices in Economies of Conventional Forms." In *Re-Imagining Economic Sociology,* edited by Patrik Aspers and Nigel Dodd, 195–223. Oxford: Oxford University Press.

Thompson, Michael. 1979. *Rubbish Theory: The Creation and Destruction of Value*. Oxford University Press.

Timmermans, Stefan, and Steven Epstein. 2010. "A World of Standards but Not a Standard World: Toward a Sociology of Standards and Standardization." *Annual Review of Sociology* 36(1): 69–89.

Uekötter, Frank. 2012. *Die Wahrheit Ist Auf Dem Feld: Eine Wissensgeschichte Der DeutschenLandwirtschaft*. Göttingen: Vandenhoeck & Ruprecht.

Varga, Mihai. 2019. "From the Qualities of Products to the Qualities of Relations: Value Conventions in the Solidarity Economy in Sicily." *Valuation Studies* 6(1): 63–86.

Ventresca, Marc J., and John W. Mohr. 2002. "Archival Research Methods." In *Blackwell Companion to Organizations,* edited by Joel A.C. Baum, 805–828. Malden, MA: Blackwell Publishing.

Weber, Klaus, Kathryn L. Heinze, and Michaela DeSoucey. 2008. "Forage for Thought: Mobilizing Codes in the Movement for Grass-Fed Meat and Dairy Products." *Administrative Science Quarterly* 53(3): 529–567.

White, Harrison C. 1981. "Where Do Markets Come From?" *American Journal of Sociology* 87(3): 517–547.

Wilde, Mandy de. 2020. "A Care-Infused Market Tale: On (Not) Maintaining Relationships of Trust in Energy Retrofit Products." *Journal of Cultural Economy* 13(5): 561–578.

Nadine Arnold, PhD in Sociology, is an Assistant Professor of Organisational Theory at the Vrije Universiteit Amsterdam. Her research focuses on the nexus between organisation and society, specifically on new forms of organisation and how these shape and are formed by society. She pays particular attention to the organisation of responsibility and how the phenomena of organising and valuing interrelate in the context of sustainability. Theoretically, Nadine mobilises insights from institutional and organisational theory as well as economic sociology. Empirically, she works on fair trade, sustainability standards, and (food) waste.

Simon Dombrowski, PhD in Sociology, is a science manager at the Hamburg University of Technology. Previously, he conducted research at the interface of economic and organisational sociology at the Department of Social Sciences at the University of Hamburg. His research focuses on the historical dynamics of the organisation of markets and quality.