Article

Philosophy of the Social Sciences
2024, Vol. 54(1) 69–92

DOI: 10.1177/00483931231188825
journals.sagepub.com/home/pos
Sage

We Have Big Data, But Do We Need Big Theory? Review-Based Remarks on an Emerging Problem in the Social Sciences

Hermann Astleitner[1]

Abstract

Big data represents a significant challenge for the social sciences. From a philosophy-of-science perspective, it is important to reflect on related theories and processes for developing them. In this paper, we start by examining different views on the role of theories in big data-related social research. Then, we try to show how big data is related to standards for evaluating theories. We also outline how big data affects theory- and data-based research approaches and the process of theory building. Discussions include a summary of lines of reasoning, limitations, and a proposal for future developmental steps.

Keywords

big data, theory development, social research, philosophy of science, artificial intelligence

Received 5 October 2022

[1] Paris Lodron University of Salzburg, Austria

Corresponding Author:
Hermann Astleitner, Paris Lodron University of Salzburg, Erzabt-Klotz-Strasse 1, Salzburg 5020, Austria.
Email: Hermann.Astleitner@Plus.Ac.At

1. Introduction

"Big data" has undoubtedly reached the social sciences (e.g., Foster et al. 2016 [2020]; Jemielniak 2020). In general, the term describes "the collection, processing, analysis and visualization associated with very large data sets" (Emmanuell and Stanier 2016, 1). In particular, big data is characterized by facets such as volume (covering enormous quantities of data), velocity (creating evidence in real time), variety (integrating different structuring approaches), exhaustivity (capturing a full sample), resolution and identification (differing in granularity and addressing multiple entities and processes), relationality (linking multiple data sets), extensionality and scalability (adding new data and expanding in size), veracity (containing uncertainty and errors), value (extracting many different insights), and variability (shifting meaning in relation to the context) (Kitchin and McArdle 2016). In the field of social sciences, big data has numerous implications for exploration techniques, sample design or creation, measurement and data collection, data integration and combination, data validation and quality assessment, error frameworks, statistical analyses, and interpretation as well as automated reporting and publishing.[1] These and similar implications raise many critical questions and unsolved problems in the methodology or methods of the social sciences related to statistics, measurement and testing, or computer applications such as machine learning and artificial intelligence.[2] The potential of big data is also apparent in the fact that new subdisciplines are currently emerging in the social sciences that are based on big data, including "data science methods" (Foster et al. 2016 [2020]), "computational social science" (McLevey 2022), and "digital humanities" (Drucker 2021). From these significant developments emerges the problem of whether changes in the methodology and methods of a science also produce changes in other fundamentals of this science.

Objectives of the Study and Research Questions

Based on a comprehensive review of literature in the field of social sciences, our objective is to collect, structure, and reflect ideas and concepts on the role of theories and theory building in relation to big data. Theories are propositional systems about social phenomena, and their validity is tested by confronting them with evidence and relevant data. First, we start by asking

[1] For a helpful comprehensive, detailed, and critical introduction to innovative methods, see Hill et al. (2020).

[2] Rubin (2022) argues that we should be wary of such tools. The data sources used are not always of good quality and the predictions made are often not exactly verifiable. This leads to serious problems and opens questions of ethics and accountability.

whether there is a need, in the field of philosophy of science, to focus on big data and theories. Second, we examine whether big data has implications for the design or formulation of theories. Third, we investigate whether there are problems related to standards for evaluating theories based on big data. Finally, we seek to determine whether big data has significant potential in the discovery and building of theories. Since big data is not yet a well-developed research field in the social sciences, especially when considering a theoretical perspective, we take an exploratory problem-sensitive approach to uncover open questions or unsolved problems for future investigations.

2. Big Data, Philosophy of the Social Sciences, and Big Theory

From a philosophy-of-social-science perspective on big data, some researchers have discussed a paradigmatic shift toward "new empiricism" (based on a stronger focus on data evidence; Arbia 2021) or "digital positivism" (related to computer-generated evidence about the world; Fuchs 2017). More specifically, Chin-Yee and Upshur (2019) have identified three major philosophical problems associated with big data: a) the epistemological-ontological problem arising from the theory-ladenness of big data, b) the epistemological-logical problem resulting from the limitations of algorithms as well as reliability and interpretability issues, and c) the phenomenological problem of the irreducibility of human experiences to data. Recently, Pietsch (2021) has discussed the methodological dichotomy between inductivism and hypothetico-deductivism and argued that successful big data algorithms are based on variational induction in the tradition of Mill's methods.[3] Cabrera (2021) has stressed that using correlations based on big data for predictive purposes is dangerous unless some plausible causal-nomological connection undergirds these correlations. Others have dealt with big data ethics and found problems regarding privacy (big data traces life and makes it transparent), propensity (big data increases the likelihood of random findings based on incidental co-occurrences, which might have negative consequences for people), and research ethics (big data collects data without informed consent) (Zwitter 2014). All of this work essentially confirms the relevance of big data to the philosophy of social science.

As methods and evidence play a larger role in social sciences through the development of big data, the question inevitably arises of whether this impacts another essential fundamental of research and philosophy of science, namely, theories. At present, there are different assumptions on the role of theories in

[3] Pietsch addresses the essential questions of whether correlations are replacing causality, whether theories are in jeopardy, and whether big data will lead to an entirely new science.

relation to big data. For example, Mazzocchi (2015, 1250) has wondered whether big data could be the end of theory in science: "Analyzing vast volumes of data will yield novel and often surprising correlations, patterns and rules. Inasmuch as the latter emerge through a bottom-up process based on inductive processes and statistical manipulation, no theory is apparently required." However, others like Wise and Shaffer (2015, 9) are convinced that theory matters more than ever in the age of big data because it provides guidance about a) which variables to include in a model on data structures, b) which confounds, subgroups, or covariates influence the data, c) which results to attend to, d) how to interpret complex results, e) how to apply results, and finally f) how to generalize results to other settings. From a natural science perspective, others have gone so far as to assert that "big data need big theory too" (Coveney, Dougherty, and Highfield 2016), that is, a scientific theory that describes and explains phenomena related to big data.

Following these different assumptions, the question is what big theories or big data-related theories could be and whether such concepts would be viable in the social sciences. Are big data-related theories just new theories or other types of theories like, for example, open system approaches? Are these theories extended core theories or changes in the peripheral areas of theories? Perhaps big data means that the social sciences should pay less attention to theory confrontation or replacement and more to unified approaches in science and theory integration (or even completion) to create more complex theories that can deal with complex big data. Another approach could be to focus more on process theories because big data can often be collected over a longer period. If big data depicts many layers within social entities such as individuals, groups, institutions, or societies, then something like meta-theories would be required that create a connection between (hierarchically organized) micro and macro areas. Alternatively, it may also be relevant to push pluralistic approaches that, for example, consider alternative explanations or competitive theories as well. Existing approaches on data or evidence-based theory building in the social sciences do not give sufficient answers to these and similar questions; therefore, they will also be important for future investigations (e.g., Shoemaker, Tankard, and Lasorsa 2004; Reynolds 2007; Swanson and Chermack 2013; Jaccard and Jacoby 2009 [2020]).[4]

A first preliminary answer to these questions could be that big data-related theories might be "system theories" concerning a large set of hypotheses on complex social systems (e.g., Dabbaghian and Mago 2014). The social sciences have produced multiple complex system theories, for instance, on

[4] However, it must be mentioned that the frameworks for theory construction proposed by Jaccard and Jacoby (2009 [2020]) also contain approaches on mathematical modeling and simulation that are relevant to theory-building processes regarding big data.

human development (Bronfenbrenner 1994), health (e.g., Sturmberg and Martin 2013), or work (Alter 2013). A comprehensive overview of new developments in the field of system theory related to “complexity theory” (as an array of overlapping and interacting areas such as system theories, complex adaptive systems, and chaos) in social sciences is given by Turner and Baker (2019). However, a philosophy of the social sciences perspective has revealed that such theories “have limited use in the study of society and that social processes are too complex and particular to be rigorously modeled in complexity terms. Theories of social complexity are shown to be inadequately developed, and typical weaknesses” (Stewart 2001, 323) include the misuse “of complexity theory in alliance with systems theory as a general and dominant metatheory for the social sciences,” “the myth that all or most social processes can usefully be quantified in mathematical terms,” the uncritical use of “metabiological, organicist (and organismic) model of social systems,” and the problem that “theories of social complexity are parented by a limited range of social philosophies that are each subject to ongoing social debate” (Stewart 2001, 329–30). The notion that we need larger or more complex theories to deal with big data can also be questioned in another way. Additional evidence from big data creates more pressure on theories. From an analytical perspective, it can be argued that big data is not about the size of a theory but about the extent to which the data allows us to reject alternative theories. In this sense, big data does not imply a larger theory but rather a smaller set of competing theories that fit the available data.[5] From a social science data-based perspective, this assumption can be supported because big data delivers additional information on rival independent and intervening variables, control variables, and antecedent and consequent variables (see the elaboration model of theory-based data analysis by Aneshensel 2002 [2013], 12). In this sense, big data could contribute to the monopolization of theories in the social sciences and lead to a reduction of theoretical plurality.

Overall, the relationship between theories and big data seems unclear and controversial and requires further analysis. It is likely that big data does not produce a disruptive paradigm change for theories in the social sciences but rather incremental gradual shifts.

3. General Big Data Implications for Theories

At first sight, it could be concluded that big data has no significant implications for theories in the social sciences because it only means more of the same (i.e., more data). This is evidenced by the fact that significant theoretical perspectives are missing from important contributions on big data and social

[5] I would like to thank the anonymous reviewer for this important point of view.

research (e.g., Woo, Tay, and Proctor 2020). However, merely having more data for developing and testing theories does not seem to be the only general implication of big data for theories within the social sciences. Maass et al. (2018) have stressed the necessity to distinguish between (more inductive-orientated) data-driven and (more deductive-orientated) theory-driven research when using big data. Luo et al. (2019) have pointed out that, on the one hand, big data could stimulate theory building by producing new insights on phenomena through data mining (i.e., finding patterns in data by using statistical methods, database systems, or machine learning), and, on the other hand, theory could help to structuralize big data sets by selecting or combining data segments based on a theoretical perspective. Among other things, Haslbeck et al. (2021) have emphasized the role of structured “data models” as an essential link between data and (formalized) theories.

Considering and integrating important elements of these authors’ arguments and assumptions, we depict in Figure 1 the relationship between big data and theory in the social sciences. Big data allows a more data-driven research process. In such a process, a real phenomenon produces a large amount of data. In this data, patterns of relationships between data segments are identified, for example, by applying statistical methods such as correlational analysis. Structuring this data creates a data model that offers an organized representation of data. This representation can partly or as a whole be converted into hypotheses that are embedded in a theory. Big data can also play a role in a theory-driven research process. Here, the starting point is a theory, its hypotheses, and related operationalized variables. Based on these variables, new data is collected or existing data is mapped. A test is then performed in which a theory-based model is fitted with the data and related

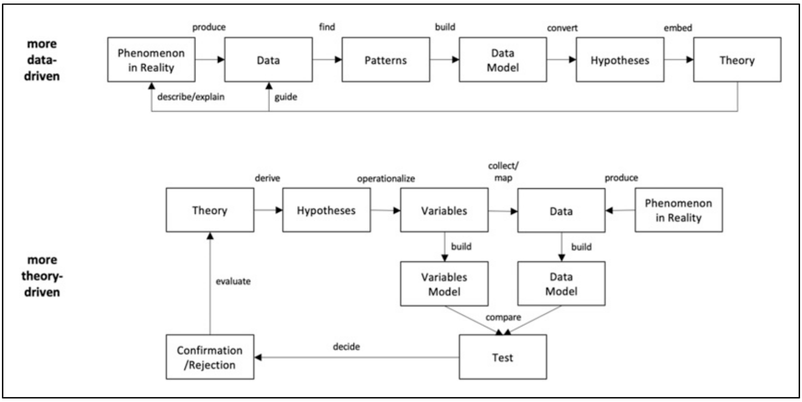

Figure 1. The relationship between big data and theory related to Maass et al. (2018), Luo et al. (2019), and Haslbeck et al. (2021).

data models. When this comparison is positive, the related theory is positively evaluated or confirmed. If the theory does not fit the data, it is rejected.

However, such data- and theory-driven approaches are not new in social research. Using a data-driven approach and finding (correlational) structures in data to stimulate theory building is found in, for example, traditional exploratory and confirmatory factor analysis (Reio and Shuck 2015). Theory-driven approaches are applied, for instance, in traditional structural equation modeling (Schumacker and Lomax 1996 [2010]).[6] It seems that more data does not lead to revolutionary research approaches but rather provides fundamentals for gradually expanding and optimizing already existing research processes in social research and related theory-building activities.

Others see big data not as a slowly appearing phenomenon but as a threat to advances in the social sciences and associated theories. More data could also lead to more bad data: having more data may generally contribute positively to the progress in the social sciences as theories can be tested more often (with multiple evidence) and be more comprehensive (with greater depth and breadth) but only insofar as it is good (or valid) data. Scholz (2017, 27–34) pointed out big data pitfalls, noting that "big data will not translate into useful observations by means of big data" and that "any type of data […] is subjective and big data especially are often collected with a certain intent and often repurposed for different goals." He also highlighted the complexity barrier, which "describes the situation wherein the effort of gaining scientific insight from a research field rises exponentially when approaching a certain threshold of complexity" and argued that "taken out of context, big data lose their meaning." Following these arguments, bad big data may also lead to bad new theories. Of course, the problem of bad theories existed before big data, and good theories can make bad predictions.[7]

To avoid bad big data and its negative consequences for theories in the social sciences, existing and newly developed theories must be able to provide support in data structuring, data selection, and related decision-making. In this sense, a reciprocal relationship between theories and big data is given, which also means that data- and theory-driven approaches to big data do not exist independently of each other. Thus, big data may seem especially relevant to research activities in which data- and theory-driven stages are systematically combined or integrated. Such approaches exist in the social sciences, for example, as "design-based research" (iterative design-and-test cycles for

[6] Structural equation modeling contains multivariate statistical analysis techniques, which enable the analysis of highly complex structural relationships between measured and latent variables. Such methods also allow, for example, the analysis of multiple samples, multiple groups, as well as higher-order, dynamic, or multimethod models.

[7] For an insightful take on when good theories can make bad predictions, see, for example, Batitsky and Domotor (2007).

developing interventions to address problems) or “mixed method research” (a combination of exploratory qualitative and testing-related quantitative research) (e.g., Ryu 2020). For example, Reimann (2016) has tried to link big data-related learning analytics to design-based research. However, except for rather individual suggestions, the role of big data and related theories in such integrative or unified research approaches is still vague. Above all, it is important to combine unified or integrated approaches in the philosophy of science with similar research approaches. Such integration of research approaches also existed long before big data (Edmonds and Kennedy 2013; Schurz 2013).

In addition, to avoid bad big data and its negative implications for theories in the social sciences, big data has to be evaluated in terms of its importance and quality by establishing data-evaluation and decision-making activities concerning the certainty of the evidence, the magnitude of benefits and harms, considerations of resource use, feasibility, acceptability, and equity as well as values and preferences (Djulbegovic and Guyatt 2017). Such a decision-making process regarding big data is influenced by numerous factors such as contractual governance, big data tool availability and capabilities, collaboration and knowledge exchange, or process integration and standardization (Janssen, van der Voort, and Wahyudi 2017). This situation might also lead to an expansion of traditional data-cleaning activities and primary and secondary traditional standards on data quality (objectivity, reliability, validity, fairness, etc.) in the social sciences to other standards like availability (accessibility, timeliness), usability (credibility), or relevance (fitness) (Cai and Zhu 2015; Ilyas and Chu 2019). Whether considering such additional standards of data quality would have a positive or a negative effect on theories in the social sciences is still an open question. For example, having more data and additional or stricter standards could contribute to solving problems with the discovery and testing of theories in the social sciences, such as the “replication crisis” (concerns about the credibility of research findings) or “publication bias” (when the outcome of a study biases the decision for publishing) as more (also high-quality) evidence is available that could be a repetition of existing findings or could not be ignored for publishing (Wagner 2022). Here, too, there are already other existing proposals in the social sciences about the replication and publication bias problems that make it possible to handle the problem without big data, such as standards for publication (Tincani and Travers 2019). These and other similar standards are related to data and how to handle it, but this leaves open the question of whether changes in standards apply not only to methods but also to theories.

4. Big Data and Standards for Evaluating Theories

Traditionally, we use the following standards for evaluating theories in social sciences (Prochaska, Wright, and Velicer 2008, 565): Clarity (the theory has well-defined terms), consistency (the theory's components do not contradict each other), parsimony (the theory explains the phenomenon in the least complex manner), testability (the theory's propositions generate empirical evidence and can be falsified), empirical adequacy (the theory explains why something happened and when something will happen), productivity (the theory reveals new phenomenon), generalizability (the theory generalizes to other situations, places, times), integration (the theory combines a set of constructs in systematic and meaningful patterns), utility (the theory provides service), practicability (a theory-based intervention demonstrates significant efficacy), and impact (the theory has effects in a combination by reach x efficacy x number of behaviors changed).

However, these traditional standards must be re-evaluated if we consider the fact that there are different types of big data-related theories that have to meet different requirements. For example, when focusing a complex system on big data-related theories, additional general constituent properties must be kept in mind, such as holistic systemness, the integration of structural and functional hierarchies, the coordination of dynamics on multiple time scales, various kinds and levels of complexity, encoding and sending/receiving information, or the implementation of regulation subsystems to achieve stability (Mobus and Kalton 2015). We can treat such constituent properties of complex system theories as standards for their evaluation because the absence of these properties will lower the quality of the theory. Alternatively, with big data-related theories as (machine-)learning systems, we must consider other standards for evaluating theories, for example, interpretability (the degree to which the model can be interpreted by humans), robustness (the ability of a model to handle missing data and still make good predictions), or effectiveness (the degree to which the algorithm detects context changes) (Siebert et al. 2022). When considering computational models, standard-based evaluation activities must concern, for instance, model verification (internal consistency of the model), input validation (representation of the target in a model), process validation (representation of real-world mechanisms within a model), and output validation (fit of the model or its ability to predict future states of a target) (Gräbner 2018). The existence of multiple relevant standards for big data-related theories in the social sciences creates consequential problems for theory building (see Figure 2).

First, there is the problem of standard incompatibility. For example, according to traditional standards for the evaluation of theories, parsimony (of elements and relationships) is important. However, theories on complex systems are highly elaborate and not at all parsimonious as they cover many

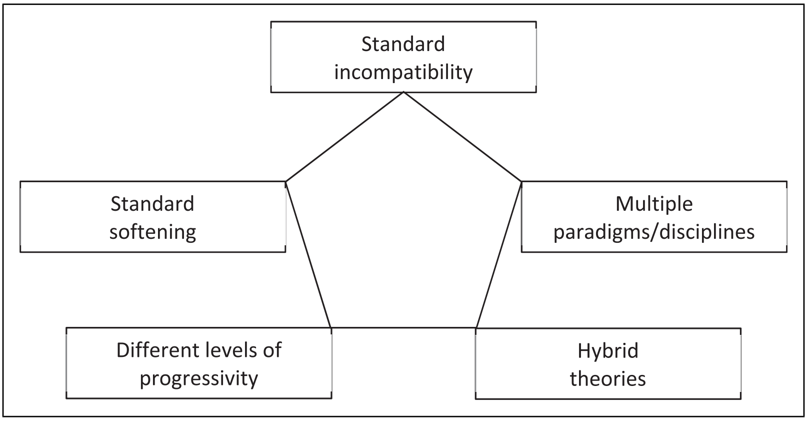
Standard incompatibility
Standard softening
Multiple paradigms/disciplines
Different levels of progressivity
Hybrid theories

Figure 2. Big data-related problems in theory building.

elements and relationships to produce a holistic or system-like picture of a social phenomenon. To deal with the problem of standard incompatibility, the social sciences could rely on meta-standards that relate to general and superior criteria that can be included in the evaluation of theories and theory development processes. When the standards are incompatible, one such meta-standard could be that theories should be built and tested from simple to complex versions in a step-by-step procedure. Thus, standard incompatibility is eliminated as simple (more parsimonious) models are not incompatible with complex (less parsimonious) models because complex models arise from simpler models. In this example, a meta-standard could be that theories should be developed and tested from simple (parsimonious) to complex (non-parsimonious) versions that are built on each other. An example of simple and complex versions of a theory within the social sciences (not focused on big data) is the Cognitive-Affective-Social Theory of Learning in Digital Environments (CASTLE) model developed by Schneider et al. (2022), which is based on a simpler version by Moreno (2006).[8] Overall, and despite these examples, the role of meta-standards in handling standard incompatibilities in big data-related theory development and testing still requires further investigation.

Second, there is the problem of standard softening when multiple big data-related theories have to be integrated at different stages of theory development. For example, when combining or integrating theories from different

[8] The original model describes and explains how instructional technology affects sensory, working, and long-term memory. The extended model integrates social cues, verbal and non-verbal information, and social processes, as well as motivation, emotion, and metacognition.

subject areas as a necessity in a big data situation, some theories are likely strong theories with significant evidence whereas others are weak and without sufficient evidence. The question is now what happens to a complex big data-related theory when some parts have sufficient evidence and others do not. A possible solution to the problem of differently developed theories and related standard softening might be the use of Bayesian approaches (Gill 2002 [2014]). They enable the use of probability as a measure of uncertainty, which makes it possible to integrate variables and relationships with strong or weak empirical evidence (e.g., Uusitalo 2007). Bayesian approaches are probabilistic graphical models containing a set of variables and their conditional dependencies. For example, Humphreys and Jacobs (2015) have presented a Bayesian framework for the integration of quantitative and qualitative data at different levels or weights of evidence. A recent example of the use of the Bayesian approach in the field of social sciences was provided by Alvarez-Galvez (2016) regarding the relationship between socioeconomic status and health. However, whether Bayesian approaches represent a sustainable helpful tool for theory building and testing is still discussed (e.g., Jones and Love 2011).

Third, the different levels of progressivity are problematic when new and established phenomena and new and established theoretical assumptions are addressed simultaneously. In such a situation, we can handle established phenomena with established or new theories or handle new phenomena with established or new theories (Zahra 2007). In these cases, we need a standard for whether to be more or less progressive in big data-related theory building. A possible solution to this problem could be to focus on comparing multiple models when developing and testing theories. Multiple models could then contain old and new theories as well as old and new phenomena. These multiple model tests and other similar ones are established in many fields of social sciences (e.g., Jang 2008). Another solution to make a decision concerning progressivity in theory building is related to the variability of empirical findings. Yang (2002) has suggested examining the variability of effect sizes (i.e., a measure of the magnitude of an effect or relationship) for the variables of interest. If there is no significant variability, the focus should be on an old theory. If there is significant variability, moderators or interaction effects should be identified. If they are significant, a new phenomenon should be included in an old theory; if they are not significant, theory building should be guided by relying on new variables and new theories. However, to the best of our knowledge, neither multiple model tests nor a focus on effect variability have been applied in the context of big data in the social sciences.

Fourth, there is the problem of the existence of different types of theories when building big data-related theories. Big data is often based on hybrid data from different temporal, local, institutional, and other sources. What is true for data might also be true for theory and theory building: some big data-related theories might be "hybrid theories" in which different types of theories are integrated. According to Svejvig (2021), a "hybrid theory" is a combination of two or more types of theories. These include descriptive theories (about how something is), predictive theories (about what will happen if something occurs or an explanation for why something occurs), prescriptive theories (about what to do to achieve a certain goal), and practice theories (about locally situated actions). The problem is that these different types of theories are associated with different goals and different standards for evaluating them. For example, Miller (2010) identified five standards for evaluating practical (evaluation) theories (operational specificity, range of application, feasibility in practice, discernable impact, and reproducibility) that are not compatible with standards for evaluating other types of theories, such as descriptive or predictive theories. There might be different solutions to this problem. Other types of theories may be used as a trigger for generating new ideas for theory development without changing the original type of the theory in focus. For example, in the field of social sciences, Astleitner (2018) used descriptive and explanatory motivational theories to build a prescriptive instructional design theory on multidimensional engagement in learning.[9] Alternatively, theories can be translated into other types of theories. For instance, Funnell and Rogers (2011) presented the concept of practical "program theories" (about how an intervention results in outcomes) in which different descriptive, prescriptive, and practical theories are integrated. In addition, it may be possible to include descriptive and prescriptive elements in the same theory. For example, Renkl (2014) proposed a theory of example-based learning containing both descriptive and prescriptive elements. However, overall, the problems encountered by hybrid theories in the context of big data have not yet been analyzed in full detail in social sciences.

Fifth, the existence of multiple paradigms and multiple disciplines in big data-related theories based on different areas of society as well as scientific disciplines poses a problem (e.g., Hu and Zhang 2018). Traditionally, in social science, many researchers assume paradigm incommensurability (Kuhn 1962) due to differences in ontology, epistemology, methodology, or views on human nature that would not allow deriving theories on data from different sources embedded in different paradigms. Nonetheless, theories from different paradigms can arguably be integrated into a unified approach based on a multi-paradigm perspective without integration or serve as a new domain for developing a new theory (Schurz 2013; Wang and Segal 2014; Zahra and

[9] Prescriptive theories in the field of instructional design contain principles for enhancing learning (and development) through instructional activities. For example, according to Astleitner (2018), such a principle can be as follows: In order to stimulate intrinsic motivation in the classroom, promote fantasy and curiosity by establishing playful activities and discovery learning.

Newey 2009). A similar situation is observed from an interdisciplinary perspective in social sciences (e.g., Opp 2009). Approaches to solving the problems resulting from multiparadigm or multidisciplinary big data rely on the assumption that different theories can have a multiparadigmatic or multidiscipline perspective. At least for multidiscipline perspectives, such theories can be found in social research, notably an interdisciplinary theory of coordination (Malone and Crowston 1991), an interdisciplinary model of school absenteeism (Kearney 2008), an interdisciplinary model of consumer trust (Love, Mackert, and Silk 2013), and an interdisciplinary privacy and communication model (Bräunlich et al. 2021). Such theories could be used as models to develop interdisciplinary theories based on big data. Another approach might be to apply methods of theory building that make it possible to build multiparadigm and interdisciplinary theories. Such approaches exist in metaparadigm perspectives (Gioia and Pitre 1990), multiparadigm simulation methods (e.g., Mykoniatis and Angelopoulou 2020), and the development of interdisciplinary theoretical frameworks (e.g., CohenMiller and Pate 2019). These approaches have not yet significantly been used and tested in the context of big data and social sciences.

In summary, it must be emphasized that many of the abovementioned problems faced in evaluating theories (standard incompatibility, standard softening, levels of progressivity, hybrid theories, and multiple paradigms/ disciplines) are not new in the social sciences. These problems all existed before big data and have been dealt with accordingly, although without an explicit big-data perspective (e.g., Christian et al. 2019; Magnani and Bertolotti 2017; Repko and Szostak 2008 [2021]).

5. Big Data as a Source for Building Big Theories Based on Data Mining

Although these questions remain open and many themes on big data have long been discussed in the social sciences without big data, more evidence may enable the evaluation of theories in a more sophisticated way, especially using systematically evaluated and structured big data. In this way, big data is an important source for the evaluation of theories that is generally related to the building of theories in social sciences: a negatively evaluated theory is redesigned or eliminated to allow new theories to appear. Nevertheless, there is also a more specific way that big data can change theory building in the social sciences.

Big data allows for the identification of social phenomena that would not be discovered by human observation or creativity alone, such as complex population dynamics (e.g., Turchin 2003). Especially, more or less intelligent data mining plays a central role as a process of transforming raw data into useful information or knowledge. Kar and Dwivedi (2020) have shown that text mining (e.g., topic modeling or summarizing opinions), network mining (e.g., flagging noticeable deviant events or behaviors), and image mining (e.g., extracting patterns from groups of images) can identify patterns that deliver fundamentals for theory development. In social sciences, open-source tools for data mining are available, such as R (https://www.r-project.org), WEKA (https://cs.waikato.ac.nz/ml/weka/), and ORANGE (https://orangedatamining.com). For example, Tomasevic, Gvozdenovic, and Vranes (2020) have used data mining with large datasets and machine-learning techniques from artificial intelligence in the field of performance prediction. Safron and DeYoung (2022) have focused on complex approaches for modeling persons that might be used in computational psychiatry. They have based their model on probabilistic descriptions of patterns of emotion, motivation, cognition, and behavior. A sophisticated approach to using big data tools for theory building in social sciences was proposed by Shrestha et al. (2021). They used and tested an induction approach supported by a machine learning-based algorithm comprised of different stages, including the division of data randomly into two samples, the identification (via data-mining techniques) of comprehensible and robust associations within a first sample, the building of a theoretical explanation, and the testing of hypotheses in the second sample. This procedure corresponds to traditional methods of cross-validation and relates to the philosophy-of-science problem that data used to build a model cannot play a role in confirming the model's predictions (De Rooij and Weeda 2020; Steele and Werndl 2020). Overall, however, a conclusive test about whether such approaches are less, equally, or more effective in theory building than traditional approaches is still missing.

There are also counter-arguments against using such big data tools for theory building in the social sciences. Foster and Stine (2004) have questioned the need for modern data-mining techniques in theory building. They found that traditional statistical regressions resulted in better or equally predictive models of bankruptcy than recently developed data-mining tools. Others have argued that focusing on big data is not necessary because big data can also be broken down into smaller datasets that can be handled with conventional analysis techniques. For example, traditional statistical cluster analyses enable the reduction of large samples by identifying similar groups of individuals, and statistical factor analyses support the reduction of numerous observed variables to a few underlying constructs. The results produced by reduced datasets can then be combined in meta-analyses (i.e., delivering statistics resulting from the combination of multiple scientific studies) (Cheung and Jak 2019). Such meta-analyses have been shown to support the development of scientific knowledge and theories as they increase the precision of a research paradigm by clarifying constructs and their relationships, finding mediators and moderators, and supporting consensus building (Chan and Arvey 2012). However, what might be clear to social scientists is not to philosophers of social science. For example, many social scientists assume that statistical methods such as cluster or factor analyses produce (results as evidence or) data. Others think that in such a case, an abductive operation must be considered that distinguishes in more detail between "selective abductions" (selecting an optimal candidate from possible explanations) to "creative abductions" (introducing new theoretical concepts and models) (Schurz 2016).

Another possible limitation concerns the integration of complex micro- and macro-data into one unifying theoretical approach in a clean relationship based on micro-micro links. For example, Raub, Buskens, and Van Assen (2011) have discussed how micro-macro links have to be conceptualized and used to model micro and macro models in the field of sociology. From a theory-building perspective, these authors have also examined the problem that theoretical approaches often focus exclusively on one hierarchical level of a phenomenon (e.g., the micro level) without having valid assumptions for other levels (e.g., the macro level). Additionally, including micro theories in micro-macro models produces a problem for theory testing as it "leaves the question open [as to] how much priority should be given to empirical tests and to the empirical corroboration of micro-assumptions when the focus is on macro-outcomes" (17). Others, like Van den Berg and Van der Klaauw (2001), have combined micro and macro unemployment data using complex mathematical modeling alone. However, innovative and significant theoretical perspectives on important social or individual factors based on decades of social research on unemployment are missing (Wanberg 2012). This lack or scarcity of theory research corresponds to traditional exploratory research approaches in which no theory is available (e.g., Brashear et al. 2006).

Overall, the situation of theory building in the social sciences is heterogeneous, and big data and related intelligent data mining tools a) are not at all used for theory building, b) are not necessary because alternatives are available for theory building, c) are used without significant benefit, or d) are used without demonstrating superiority over traditional methods of theory building. However, data mining seems to be a very promising approach for the social sciences given the rapid development of artificial intelligence tools (e.g., Luan et al. 2020).

6. Discussions

In this paper, we discussed different potentials, problems, and open questions regarding big data in theories and theory building in the social sciences based on a comprehensive literature review. We found highly heterogeneous assumptions and perspectives resulting in five controversial lines of argumentation. First, on the one hand, the development and use of complex system theories could be the appropriate reaction to big data; on the other hand, social phenomena may not be validly mapped using complex formal systems. Second, some scholars postulate no implications for theories in social sciences as big data only means more of the same. Still, others see gradual changes in the development and testing of theories based on a combination of data- and theory-driven research processes. Third, some authors point out that bad big data could lead to bad theories. Therefore, big data should undergo an advanced evaluation process that must be based on additional data-related criteria but also on existing theories. Fourth, big data-related theories have to meet extended standards for theory development and testing based on the specific type of theoretical framework. However, such extended standards are associated with problems of incompatibility, softening, progressivity, hybrid approaches, and multiple paradigms and disciplines. Fifth, big data might enable the identification (via data mining) of social phenomena that would not be discovered by human imagination. Yet, a conclusive test to determine whether such big data-related tools and procedures are effective or even more effective than more traditional approaches to theory building is still missing.

Our paper is subject to some limitations. We have collected arguments and counter-arguments without offering sound solutions to controversies. We limited our contribution to identifying and relating arguments. We have not developed a meta- or similar theory or approach that would make any type of data and related theory embeddable. Moreover, we have not captured all the important implications of big data for theory building in social sciences. For example, we did not discuss what the creation of big data in real time means for theory building and testing or how the expansion of theories and the scalability of big data approaches are related. However, our collection of problems and open questions can be seen as a starting point for research programs or projects in the social sciences. In this respect, our review not only presents a summary of lines of argumentation but also stimulates future research in the philosophy of social sciences as well as in many other fields of social sciences.

Despite the open questions and limitations, one thing seems to be clear for future social sciences research using big data: the integration of big data into processes of theory development and testing in the social sciences arguably requires an expansion of researchers' skills in the areas of formal logic, mathematics, and computer programming (e.g., Brooker 2020). On the one hand, this implies the need for greater integration of these fields into study programs to be able to train and prepare young social scientists. Additional courses will have to be offered or existing method training will have to be adapted. For example, the software package R could be used not only for statistical analyses in the social sciences but also as a programming tool for complex modeling (e.g., Kennedy and Waggoner 2021). On the other hand, this also makes it necessary for social scientists to expand interdisciplinary cooperation with philosophers of science, logicians, mathematicians, and computer scientists. Social scientists will have to understand how computer programs (analyzing big data) think and that results are not only data but also patterns of abductive inference that require not less but more theoretical work (Schurz 2017).

7. Conclusions

The role of big data in the social sciences is dominated by a focus on quantitative research methods. However, in this paper, we have reflected on big data and previously underestimated topics related to theories and theory development. We found that the relationship between big data and theories is controversial and changing, requiring continued attention, holistic analysis, and critical appraisal. Dealing with big data will likely generate more complex theories, which will create a number of problems in theory development and associated standards. These problems range from standard incompatibility, standard softening, progressiveness, and hybrid theories to the integration of multiple paradigms or disciplines. We also pointed out that artificial intelligence applications are potentially important tools for theory development using big data. The matter is dynamic, but the philosophy of social sciences should keep an eye on these developments while making more efforts toward interdisciplinarity.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Hermann Astleitner https://orcid.org/0000-0002-2934-7126

References

Alter, Steven. 2013. “Work System Theory: Overview of Core Concepts, Extensions, and Challenges for the Future.” *Journal of the Association for Information Systems* 14 (2): 1. DOI: 10.17705/1jais.00323

Alvarez-Galvez, Javier. 2016. “Discovering Complex Interrelationships Between Socioeconomic Status and Health in Europe: A Case Study Applying Bayesian Networks.” *Social Science Research* 56: 133-143. DOI: 10.1016/j.ssresearch.2015.12.011

Aneshensel, Carol S. 2002 (2013). *Theory-Based Data Analysis for the Social Sciences*. 2nd ed. Los Angeles: Sage.

Arbia, Giuseppe. 2021. *Statistics, New Empiricism and Society in the Era of Big Data*. Cham, Switzerland: Springer.

Astleitner, Hermann. 2018. "Multidimensional Engagement in Learning—An Integrated Instructional Design Approach." *Journal of Instructional Research* 7: 6-32. DOI: 10.9743/JIR.2018.1

Batitsky, Vadim and Zoltan Domotor. 2007. "When Good Theories Make Bad Predictions." *Synthese* 157 (1): 79-103. DOI: 10.1007/s11229-006-9033-0

Brashear, Thomas G., Danny N. Bellenger, James S. Boles, and Hiram. C. Barksdale Jr. 2006. "An Exploratory Study of the Relative Effectiveness of Different Types of Sales Force Mentors." *Journal of Personal Selling & Sales Management* 26 (1): 7-18. DOI: 10.2753/PSS0885-3134260101

Bräunlich, Katharina, Tobias Dienlin, Johannes Eichenhofer, Paula Helm, Sabine Trepte, Rüdiger Grimm, Sandra Seubert, and Christoph Gusy. 2021. "Linking Loose Ends: An interdisciplinary Privacy and Communication Model." *New Media & Society* 23 (6): 1443-1464. DOI: 10.1177/1461444820905045

Bronfenbrenner, Urie. 1994. "Ecological Models of Human Development." In *International Encyclopedia of Education*, edited by Torsten Husén and T. Neville Postlethwaite, vol. 3, 1643-1647. New York: Elsevier Science.

Brooker, Philip D. 2020. *Programming with Python for Social Scientists*. Los Angeles: Sage.

Cabrera, Frank. 2021. "The Fate of Explanatory Reasoning in the Age of Big Data." *Philosophy & Technology* 34 (4): 645-665. DOI: 10.1007/s13347-020-00420-9

Cai, Li and Yangyong Zhu. 2015. "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era." *Data Science Journal* 14: 2. DOI: 10.5334/dsj-2015-002.

Chan, MeowLan Evelyn and Richard D Arvey. 2012. "Meta-Analysis and the Development of Knowledge." *Perspectives on Psychological Science* 7 (1): 79-92. DOI: 10.1177/1745691611429355

Cheung, Mike W.-L. and Suzanne Jak. 2019. "Challenges of Big Data Analyses and Applications in Psychology." *Zeitschrift für Psychologie* 226 (4): 209-211. DOI: 10.1027/2151-2604/a000348

Chin-Yee, Benjamin and Ross Upshur. 2019. "Three Problems with Big Data and Artificial Intelligence in Medicine." *Perspectives in Biology and Medicine* 62 (2): 237-256. DOI: 10.1353/pbm.2019.0012

Christian, Alexander, David Hommen, Nina Retzlaff, and Gerhard Schurz, eds. 2019. *Philosophy of Science: Between the Natural Sciences, the Social Sciences, and the Humanities*. Cham, Switzerland: Springer.

CohenMiller, A. S. and Elizabeth Pate. 2019. "A Model for Developing Interdisciplinary Research Theoretical Frameworks." *Qualitative Report* 24 (6): 1211-1226. Retrieved from: https://nsuworks.nova.edu/tqr/vol24/iss6/2

Coveney, Peter V., Edward R. Dougherty, and Roger R. Highfield 2016. "Big Data Need Big Theory Too." *Philosophical Transactions of the Royal Society A:*

Mathematical, Physical and Engineering Sciences 374 (2080): 20160153. DOI: 10.1098/rsta.2016.0153

Dabbahian, Vahid and Vijay Kumar Mago, eds. 2014. *Theories and Simulations of Complex Social Systems*. Berlin: Springer.

De Rooij, Mark and Wouter Weeda. 2020. "Cross-Validation: A Method Every Psychologist Should Know." *Advances in Methods and Practices in Psychological Science* 3 (2): 248-263. DOI: 10.1177/2515245919898466

Djulbegovic, Benjamin and H. Guyatt Guyatt. 2017. "Progress in Evidence-Based Medicine: A Quarter Century on." *Lancet* 390 (10092): 415-423. DOI: 10.1016/S0140-6736(16)31592-6

Drucker, Johanna. 2021. *The Digital Humanities Coursebook. An Introduction to Digital Methods for Research and Scholarship*. London: Routledge.

Edmonds, W. Alex and Thomas D. Kennedy. 2013. *An Applied Reference Guide to Research Designs. Quantitative, Qualitative, and Mixed Methods*. Los Angeles: Sage.

Emmanuel, Isitor and Clare Stanier. 2016. "Defining Big Data." In BDAW'16: Proceedings of the International Conference on Big Data and Advanced Wireless Technologies, edited by Djallel Eddine Boubiche, Hani Hamdan, and Ahcéne Bournceur, 1-6. New York: Association for Computing Machinery. https://dl.acm.org/doi/10.1145/3010089.3010090

Foster, Dean P. and Robert A. Stine. 2004. "Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy." *Journal of the American Statistical Association* 99 (466): 303-313. DOI: 10.1198/016214504000000287

Foster, Ian, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, and Julia Lane, eds. 2016 (2020). *Big Data and Social Science: Data Science Methods and Tools for Research and Practice*. 2nd ed. Boca Raton, FL: CRC Press.

Fuchs, Christian. 2017. "From Digital Positivism and Administrative Big Data Analytics Towards Critical Digital and Social Media Research." *European Journal of Communication* 32 (1): 37-49. DOI: 10.1177/0267323116682804

Funnell, Sue C. and Patricia J. Rogers. 2011. *Purposeful Program Theory. Effective Use of Theories of Change and Logic Models*. San Francisco: Jossey-Bass.

Gill, Jeff. 2002 (2014). *Bayesian Methods. A Social and Behavioral Sciences Approach*. 3rd ed. Boca Raton, FL: CRC Press.

Gioia, Dennis A. and Evelyn Pitre. 1990. "Multiparadigm Perspectives on Theory Building." *Academy of Management Review* 15 (4): 584-602. DOI: 10.5465/amr.1990.4310758

Gräbner, Claudius. 2018. "How to Relate Models to Reality? An Epistemological Framework for the Validation and Verification of Computational Models." *Journal of Artificial Societies and Social Simulation* 21 (3): 8. DOI: 10.18564/jasss.3772

Haslbeck, Jonas M. B., Oisín Ryan, Donald J. Robinaugh, Lourens J. Waldorp, and Denny Borsboom. 2021. "Modeling Psychopathology: From Data Models to

Formal Theories." *Psychological Methods* 27 (6): 930-957. DOI: 10.1037/met0000303

Hill, Craig A., Paul P. Biemer, Trent D. Buskirk, Lilli Japec, Antje Kirchner, Stas Kolenikov, and Lars E. Lyberg, eds. 2020. *Big Data Meets Survey Science: A Collection of Innovative Methods*. Hoboken, NJ: John Wiley & Sons.

Hu, Jiming and Yin Zhang. 2018. "Measuring the Interdisciplinarity of Big Data Research: A Longitudinal Study." *Online Information Review* 42 (5): 681-696. DOI: 10.1108/OIR-12-2016-0361

Humphreys, Macartan and Alan M. Jacobs. 2015. "Mixing Methods: A Bayesian Approach." *American Political Science Review* 109 (4): 653-673. DOI: 10.1017/S0003055415000453

Ilyas, Ihab F. and Xu Chu. 2019. *Data Cleaning*. New York: Association for Computing Machinery.

Jaccard, James and Jacob Jacoby. 2009 (2020). *Theory Construction and Model-Building Skills: A Practical Guide for Social Scientists*. 2nd ed. New York: Guilford Press.

Jang, Hyungshim. 2008. "Supporting Students' Motivation, Engagement, and Learning During an Uninteresting Activity." *Journal of Educational Psychology* 100 (4): 798-811. DOI: 10.1037/a0012841

Janssen, Marijn, Haiko van der Voort, and Agung Wahyudi. 2017. "Factors Influencing Big Data Decision-Making Quality." *Journal of Business Research* 70: 338-345. DOI: 10.1016/j.jbusres.2016.08.007

Jemilniak, Dariusz. 2020. *Thick Big Data. Doing Digital Social Sciences*. Oxford: Oxford University Press.

Jones, Matt and Bradley C. Love. 2011. "Bayesian Fundamentalism or Enlightenment? On the Explanatory Status and Theoretical Contributions of Bayesian Models of Cognition." *Behavioral and Brain Sciences* 34 (4): 169-188. DOI: 10.1017/S0140525X10003134

Kar, Arpan Kumar and Yogesh K. Dwivedi. 2020. "Theory Building with Big Data-driven Research—Moving Away from the 'What' Towards the 'Why.'" *International Journal of Information Management* 54: 102205. DOI: 10.1016/j.ijinfomgt.2020.102205

Kearney, Christopher A. 2008. "An Interdisciplinary Model of School Absenteeism in Youth to Inform Professional Practice and Public Policy." *Educational Psychology Review* 20 (3): 257-282. DOI: 10.1007/s10648-008-9078-3

Kennedy, Ryan and Philip D. Waggoner. 2021. *Introduction to R for Social Scientists: A Tidy Programming Approach*. Boca Raton, FL: CRC Press.

Kitchin, Rob and Gavin McArdle. 2016. "What Makes Big Data, Big Data? Exploring The Ontological Characteristics of 26 Datasets." *Big Data & Society* 3 (1): 1-10. DOI: 10.1177/2053951716631130

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Love, Brad, Michael Mackert, and Kami Silk. 2013. “Consumer Trust in Information Sources: Testing an Interdisciplinary Model.” *Sage Open* 3 (2): 2158244013492782. DOI: 10.1177/2158244013492782

Luan, Hui, Peter Geczy, Hollis Lai, Janice Gobert, Stephen J. Yang, Hiroaki Ogata, Jacky Baltes, Rodrigo Guerra, Ping Li, and Chin-Chung Tsai. 2020. “Challenges and Future Directions of Big Data and Artificial Intelligence in Education.” *Frontiers in Psychology* 11: 580820. DOI: 10.3389/fpsyg.2020.580820

Luo, Jar-Der., Jifan Liu, Kunhao Yang, and Xiaoming Fu. 2019. “Big Data Research Guided by Sociological Theory: A Triadic Dialogue among Big Data Analysis, Theory, and Predictive Models.” *Journal of Chinese Sociology* 6: 11. DOI: 10.1186/s40711-019-0102-4

Maass, Wolfgang, Jeffrey Parsons, Sandeep Purao, Veda C. Storey, and Carson Woo. 2018. “Data-driven Meets Theory-Driven Research in the Era of Big Data: Opportunities and Challenges for Information Systems Research.” *Journal of the Association for Information Systems* 19 (12): 1253-1273. DOI: 10.17705/1jais.00526

Magnani, Lorenzo and Tommaso Bertolotti, eds. 2017. *Springer Handbook of Model-Based Science*. Cham, Switzerland: Springer.

Malone, Thomas W. and Kevin Crowston. 1991. *Toward an Interdisciplinary Theory of Coordination* (Technical Report no. 120). Cambridge, MA: Center for Coordination Science, Massachusetts Institute of Technology. https://dspace.mit.edu/bitstream/handle/1721.1/2356/SWP-3294-23946943-CCS-TR-120.pdf?sequence=1

Mazzocchi, Fulvio. 2015. “Could Big Data Be the End of Theory in Science? A Few Remarks on the Epistemology of Data-Driven Science.” *EMBO Reports* 16 (10): 1250-1255. DOI: 10.15252/embr.201541001

McLevey, John. 2022. *Doing Computational Social Science. A Practical Introduction*. London: Sage.

Miller, Robin Lin. 2010. “Developing Standards for Empirical Examinations of Evaluation Theory.” *American Journal of Evaluation* 31 (3): 390-399. DOI: 10.1177/1098214010371819

Mobus, George E. and Michael C. Kalton. 2015. *Principles of Systems Science*. New York: Springer.

Moreno, Roxana. 2006. “Learning in High-Tech and Multimedia Environments.” *Current Directions in Psychological Science* 15 (2): 63-67. DOI: 10.1111/j.0963-7214.2006.00408.x

Mykoniatis, Konstantinos and Anastasia Angelopoulou. 2020. “A Modeling Framework for the Application of Multi-Paradigm Simulation Methods.” *Simulation* 96 (1): 55-73. DOI: 10.1177/0037549719843339

Opp, Karl-Dieter. 2009. *Theories of Political Protest and Social Movements: A Multidisciplinary Introduction, Critique, and Synthesis*. London: Routledge.

Pietsch, Wolfgang. 2021. *Big Data*. Cambridge: Cambridge University Press.

Prochaska, James O., Julie A. Wright, and Wayne F. Velicer. 2008. “Evaluating Theories of Health Behavior Change: A Hierarchy of Criteria Applied to the Transtheoretical Model.” *Applied Psychology* 57 (4): 561-588. DOI: 10.1111/j.1464-0597.2008.00345.x

Raub, Werner, Vincent Buskens, and Marcel A. L. M. Van Assen. 2011. “Micro-Macro Links and Microfoundations in Sociology.” *Journal of Mathematical Sociology* 35 (1–3): 1-25. DOI: 10.1080/0022250X.2010.532263

Reimann, Peter. 2016. “Connecting Learning Analytics with Learning Research: The Role of Design-Based Research.” *Learning* 2 (2): 130-142. DOI: 10.1080/23735082.2016.1210198

Reio, Thomas G. Jr. and Brad Shuck. 2015. “Exploratory Factor Analysis: Implications for Theory, Research, and Practice.” *Advances in Developing Human Resources* 17 (1): 12-25. DOI: 10.1177/1523422314559804

Renkl, Alexander. 2014. “Toward an Instructionally Oriented Theory of Example-Based Learning.” *Cognitive Science* 38 (1): 1-37. DOI: 10.1111/cogs.12086

Repko, Allen F. and Rick Szostak. 2008 (2021). *Interdisciplinary Research. Process and Theory*. 4th ed. Los Angeles: Sage.

Reynolds, Paul Davidson. 2007. *Primer in Theory Construction*. Boston: Pearson.

Rudin, Cynthia. 2022. “Why Black Box Machine Learning Should be Avoided for High-stakes Decisions, in brief.” *Nature Reviews Methods Primers* 2 (1): 81. DOI: 10.1038/s43586-022-00172-0

Ryu, Suna. 2020. “The Role of Mixed Methods in Conducting Design-Based Research.” *Educational Psychologist* 55 (4): 232-243. DOI: 10.1080/00461520.2020.1794871

Safron, Adam and Colin G. DeYoung. 2022. “Integrating Cybernetic Big Five Theory with the Free Energy Principle: A New Strategy for Modeling Personalities as Complex Systems.” In *Measuring and Modeling Persons and Situations*, edited by Dustin Wood, Stephen J. Read, P. D. Harms, and Andrew Slaughter, 617-649. London: Academic Press.

Schneider, Sascha, Maik Beege, Steve Nebel, Lenka Schnaubert, and Günter Daniel Rey. 2022. “The Cognitive-Affective-Social Theory of Learning in Digital Environments (CASTLE).” *Educational Psychology Review* 34 (1): 1-38. DOI: 10.1007/s10648-021-09626-5

Scholz, Tobias M. 2017. *Big Data in Organizations and the Role of Human Resource Management: A Complex Systems Theory-Based Conceptualization*. Frankfurt am Main, Germany: Peter Lang.

Schumacker, Randall E. and Richard G. Lomax. 1996 (2010). *A Beginner’s Guide to Structural Equation Modeling*. 3rd ed. New York: Routledge.

Schurz, Gerhard. 2013. *Philosophy of Science. A Unified Approach*. New York: Routledge.

Schurz, Gerhard. 2016. “Common Cause Abduction: The Formation of Theoretical Concepts and Models in Science.” *Logic Journal of the IGPL* 24 (4): 494-509. DOI: 10.1093/jigpal/jzw029

Schurz, Gerhard. 2017. “Patterns of Abductive Inference.” In *Springer Handbook of Model-Based Science*, edited by Lorenzo Magnani and Tommaso Bertolotti, 151-173. Dordrecht, the Netherlands: Springer.

Shoemaker, Pamela J., James William Tankard Jr., and Dominic L. Lasorsa. 2004. *How to Build Social Science Theories*. Thousand Oaks, CA: Sage.

Shrestha, Yash Raj, Vivianna Fang He, Phanish Puranam, and Georg von Krogh. 2021. “Algorithm Supported Induction for Building Theory: How Can We Use Prediction Models to Theorize?” *Organization Science* 32 (3): 856-880. DOI: 10.1287/orsc.2020.1382

Siebert, Julien, Lisa Joeckel, Jens Heidrich, Adam Trendowicz, Koji Nakamichi, Kyoko Ohashi, Isao Namba, Rieko Yamamoto, and Mikio Aoyama. 2022. “Construction of a Quality Model for Machine Learning Systems.” *Software Quality Journal* 30 (2): 307-335. DOI: 10.1007/s11219-021-09557-y

Steele, Katie and Charlotte Werndl. 2020. “Model-Selection Theory: The Need for a More Nuanced Picture of Use-Novelty and Double-Counting.” *British Journal for the Philosophy of Science* 69 (2): 351-375. DOI: 10.1093/bjps/axw024

Stewart, Peter. 2001. “Complexity Theories, Social Theory, and the Question of Social Complexity.” *Philosophy of the Social Sciences* 31 (3): 323-360. DOI: 10.1177/004839310103100303

Sturmberg, Joachim P. and Carmel M. Martin, eds. 2013. *Handbook of Systems and Complexity in Health*. New York: Springer.

Svejvig, Per. 2021. “A Meta-Theoretical Framework for Theory Building in Project Management.” *International Journal of Project Management* 39 (8): 849-872. DOI: 10.1016/j.ijproman.2021.09.006

Swanson, Richard A. and Thomas J. Chermack. 2013. *Theory Building in Applied Disciplines*. San Francisco: Berrett-Koehler.

Tincani, Matt and Jason Travers. 2019. “Replication Research, Publication Bias, and Applied Behavior Analysis.” *Perspectives on Behavior Science* 42 (1): 59-75. DOI: 10.1007/s40614-019-00191-5

Tomasevic, Nikola, Nikola Gvozdenovic, and Sanja Vranes. 2020. “An Overview and Comparison of Supervised Data Mining Techniques for Student Exam Performance Prediction.” *Computers & Education* 143: 103676. DOI: 10.1016/j.compedu.2019.103676

Turchin, Peter. 2003. *Complex Population Dynamics: A Theoretical/Empirical Synthesis*. Princeton, NJ: Princeton University Press.

Turner, John R. and Rose M. Baker. 2019. “Complexity Theory: An Overview with Potential Applications for the Social Sciences.” *Systems* 7 (1): 4. DOI: 10.3390/systems7010004

Uusitalo, Laura. 2007. “Advantages and Challenges of Bayesian Networks in Environmental Modelling.” *Ecological Modelling* 203 (3–4): 312-318. DOI: 10.1016/j.ecolmodel.2006.11.033

Van den Berg, Gerad J. and Bas Van Der Klaauw. 2001. "Combining Micro and Macro Unemployment Duration Data." *Journal of Econometrics* 102 (2): 271-309. DOI: 10.1016/S0304-4076(01)00055-0

Wagner, John A. III. 2022. "The Influence of Unpublished Studies on Results of Recent Meta-Analyses: Publication Bias, the File Drawer Problem, and Implications for the Replication Crisis." *International Journal of Social Research Methodology* 22 (5): 639-644. DOI: 10.1080/13645579.2021.1922805

Wanberg, Connie R. 2012. "The Individual Experience of Unemployment." *Annual Review of Psychology* 63: 369-396. DOI: 10.1146/annurev-psych-120710-100500

Wang, Hao and Steven Segal. 2014. *Paradigm Incommensurability and Multiparadigm Research*. In Paper presented at the 2014 Australian & New Zealand Academy of Management Conference (ANZAM). Sydney. https://www.anzam.org/wp-content/uploads/pdf-manager/1588_ANZAM-2014-039.PDF

Wise, Alyssa Friend and David William Shaffer. 2015. "Why Theory Matters More than Ever in the Age of Big Data." *Journal of Learning Analytics* 2 (2): 5-13. DOI: 10.18608/jla.2015.22.2

Woo, Sang Eun, Louis Tay, and Robert W. Proctor, eds. 2020. *Big Data in Psychological Research*. Washington, DC: American Psychological Association.

Yang, Baiyin. 2002. "Meta-Analysis Research and Theory Building." *Advances in Developing Human Resources* 4 (3): 296-316. DOI: 10.1177/1523422302043005

Zahra, Shaker A. 2007. "Contextualizing Theory Building in Entrepreneurship Research." *Journal of Business Venturing* 22 (3): 443-452. DOI: 10.1016/j.jbusvent.2006.04.007

Zahra, Shaker A. and Lance R. Newey. 2009. "Maximizing the Impact of Organization Science: Theory-Building at the Intersection of Disciplines and/or Fields." *Journal of Management Studies* 46 (6): 1059-1075. DOI: 10.1111/j.1467-6486.2009.00848.x

Zwitter, Andrej. 2014. "Big Data Ethics." *Big Data & Society* 1 (2): 2053951714559253. DOI: 10.1177/2053951714559253

Author Biography

Hermann Astleitner is an associate professor within the Digital Learning Research Group at the Department of Educational Science of the Paris Lodron University of Salzburg, Austria. His research centers on instructional systems design, education and human development as well as intervention methods and theory building in the social sciences.