Sassan Gholiagha, Jürgen Neyer, Mitja Sienknecht; European New School of Digital Studies, European University Viadrina, Frankfurt (Oder), Germany

Paper prepared for the 64th Annual Convention of the International Studies Association, 15-18 March 2023, Montréal, Canada. Panel: Between Interpretation and Explanation: Using Artificial Intelligence and Machine Learning in International Relations

Word Count: 7979

This is a first draft. Please do not share or cite without permission. Comments are very much welcome.

A New Kid in Town.

Machine Learning and Pattern Recognition in International Relations[1]

Abstract:

The paper discusses the methodological implications of big data and Artificial Intelligence (AI) for the importance of theory in the discipline of International Relations (IR). AI is often connected to the description of objective patterns, and its introduction is accompanied by fear, and sometimes hopes, of an end of theory. Both are grossly overstated, however. IR will remain a theory-centred discipline. AI-driven approaches are welcome complements but are hardly replacing traditional ones. We will make this case by combining epistemological reflections with a report from an ongoing research project. The project uses a combination of Natural Language Processing (NLP) and Machine Learning (ML) to scrutinise a large text corpus. We demonstrate that theory and subjectivity are inseparably connected to big data and AI. The

[1] The paper is part of an ongoing research project that is financed by the Bundesministerium für Wissenschaft, Bildung und Forschung, the Brandenburger Ministerium für Wissenschaft, Forschung und Kunst as well as the Thüringer Ministerium für Wirtschaft, Wissenschaft und Digitale Gesellschaft and conducted at the European University Viadrina and the Bauhaus University Weimar. More information on SKILL can be found here: https://www.europeannewschool.eu/skill. We thank our collaboration partners Bernd Fröhlich, Katrin Girgensohn, Dorothea Horst, Dora Kiesel, Patrick Riehmann, Benno Stein, Magdalena Anna Wolska, and our team of student researchers for discussions and collaboration.

paper closes with a suggestion to begin a debate between traditional and AI-based approaches. AI promises to make big data thick and thick data big, thus providing important new stimulus to IR as a discipline.

1 A New Kid in Town

International Relations (IR) has since long been a discipline organised around theoretical debates. The three debates between 'idealism' and 'realism' in the 1920s and 1930s, between 'history' and 'science' in the 1950s and 1960s, and finally between 'positivism' and 'post-positivism' in the last twenty years have structured the discipline, shaped much of the content of major IR journals and motivated major monographs (Wæver 2010). Artificial Intelligence (AI) might become the next milestone in the development of the debate. AI has been a topic in IR for quite some time. Already in the 1980s, Sylvan and Chan (1984) published a volume on foreign policy decision-making which integrated two chapters on AI. However, both chapters offered only complex computer models but did not discuss self-learning systems. Similarly, the volume edited by Hudson (1991) promised more than it could deliver. However, the technology for self-learning systems was simply not ready yet.

More recent contributions have the benefit of being able to draw from recent technological breakthroughs. Rahal and colleagues offer a brief overview of ML in the Social Sciences (Rahal et al. 2022). Chatsiou and Mikhaylov (2020) discuss the new methodological opportunities for marrying ML and natural language in detail.[2] Kelsey Shoub and Santiago Olivella's work on "Machine Learning in Political Science: Supervised Learning Models" is also of notable relevance here (Shoub and Olivella 2020).

However, most of the discussion remains highly abstract and offers little application to political science. An important issue raised by all of these contributions – implicitly or explicitly – is to question the relevance of theory. Müller and Ritschel (2016, p. 6) declare an end of theory, stating that digital data analysis only requires machines and algorithms, but no theories. Likewise, Anderson (2013a, 2013b) and Geiselberger and Moorstedt (2013) put the relevance of theory for the generation of new knowledge into question. A third – and more cautious – a group of authors points out that theory might still have its legitimate and essential role but that the expected benefits from introducing AI-based methodologies are so significant that the

[2] Machine learning is a subfield of artificial intelligence (AI) that focuses on building systems based on data to make predictions or decisions based on that data. NLP approaches are a specific type of machine learning used to analyze and understand human language by sentiment analysis or chatbot development.

relevance of theoretical progress pales in comparison (Jemielniak 2020). These scholars argue that combining ML and NLP with an enormous amount of data ('big data') provides interpretative meaning by putting data into context. Big data analysed with the help of NLP allows us to engage in discourse analysis, mine for specific arguments and counterarguments, and even discuss findings without explicit reference to theory.

While it remains an open question how AI-centred approaches will shape the role and function of theories, we argue in this paper that theory remains essential for training and developing algorithms and models that might help to generate new knowledge in IR. AI-centred approaches are no alternative to theory but are better understood as an epistemological shift that expands the potential of IR theorising to understand global politics. Big data and AI technologies certainly already do and will change the landscape of IR and its methods even further, but they do not lead to an end of theory. Instead, theory is more needed than ever, especially if one is to make the "methodological twin-move of making *big data thick* and *thick data big"* (Adler-Nissen et al. 2021, 1, emphasis in original).[3] By demonstrating how AI technologies' reflective and critical employment can advance theoretical and empirical knowledge, we introduce a third way of inquiry that does not replace explaining or understanding but adds another approach: *recognising*. Thus, we demonstrate that by focusing on arguments and argumentative structures as a central pattern within social sciences, the power of AI technologies can lead to valuable insights and open new research avenues in IR.

To substantiate our argument and demonstrate that there is no end to theory, we will discuss their epistemological building blocks and check for in-built theoretical content in the following section. Section 3 builds on the preceding section by providing evidence from an ongoing AI-based research project that analyses argumentative structures in IR discourses. The project uses NLP and ML to scrutinise a large text corpus and find patterns in argumentative structures. Finally, section 4 concludes by summarising the argument and suggesting an intensified discourse between traditional and AI-based approaches. AI promises to provide an important new stimulus to IR as a discipline by providing methodological tools that can open new research horizons.

[3] Big data is collected in large quantities which makes it difficult to analyze, thus it must be categorized, defined and aggregated. On the opposite, thick data is qualitatively conducted, providing in-depth information about a certain field, but preventing us from the identification of larger patterns.

2 No AI Without Theory

One of the most puzzling strengths of AI is its power to analyse extensive sets of data quickly and to draw inferences. Combined with big data, AI can describe previously unknown patterns and structures—relations among phenomena which had been overlooked come into the open. A related alleged strength of the combination of AI and big data is its capacity to overcome the biases inherent in any subjective claim about reality.

The analysis of big data through AI technologies aggregates a multiplicity of perspectives, impressions, and interpretations and uses statistical means for inferring patterns. The emerging patterns inferred from sheer endless data promise to unveil structures and relations uninfected by human intelligence's subjectivism and particular theories. It is a new world that has never before been possible to describe empirically. Not surprisingly, the success of AI-based analysis of big data has introduced a new stream of scientific contributions claiming that the practices of social interaction are driven by global patterns and repetitive structures far more than individual decisions (Nassehi 2019; Mayer-Schönberger and Cukier 2013). In the remainder of this section, we aim to demonstrate that while *prima facie* AI technologies and the analysis of big data seem to be able to make no use of theory, the contrary is the case. To this end, we first discuss the epistemology of big data and introduce pattern recognition as a third logic of analysis (2.1). We then show how even within statistical objectivity, subjectivity plays a significant role (2.2).

2.1 The Epistemology of Big Data

Big data has not only found its inroads into descriptions and explanations of social practices but also affected the deeper epistemological foundations of social science. We argue that it has helped to complement the two established epistemologies of "explaining" and "understanding" (Hollis and Smith 1990) with a third one, which we suggest to call "recognising". Recognising is a holistic epistemology which deals with the seemingly naive question of "what is?".[4] It rejects the long-held conviction that we can only overcome subjectivity by adopting the lenses of good theories but holds that data can be objectively interpreted if we only have enough of them (Mayer-Schönberger and Cukier 2013). Data becomes powerful by becoming big. The object of empirical analysis is not the individual datum but a pattern emerging from the observation of a large set of data. Observing individual actions and decisions is an unnecessary

[4] Wendt notably has identified "what?" questions as relevant for explanations and not only as descriptive (Wendt 1998, p. 110) But often what questions are understood and used in the descriptive manner.

and probably even analysis-distorting form of scientific activity in a recognition-based approach. In this sense, one could argue that actors often follow a logic of regularity rather than a logic of appropriateness or rationality. Therefore, it is not the individual action but the aggregated pattern that is the appropriate benchmark of empirical research.

AI-driven approaches also differ from standard social science methodology in their concept of truth. Standard social science methodology is built on a deterministic understanding of truth, which holds that we can make either right or wrong statements. Inferences are assumed to be the product of a more or less careful data selection and meaningful choices about conceptual tools and theories. Most scholars believe we can infer correct statements if science is conducted correctly and all methodological steps have been taken according to sound standards. If our procedures have been deficient, however, inferences will most likely be wrong, i.e. be rejected by future findings. Recognition-based approaches are different. They employ a concept of truth which is probabilistic rather than deterministic. Inferences are the product of statistical operations and have a quantifiable likeliness to be right or wrong. They hardly ever claim to be perfectly true but only to have a certain probability. A typical statement would be that, for example, a state with the characteristics a and b decides in favour of option A and against B with a probability of ß per cent (cf. table 1). Truth is, therefore, a probabilistic concept.

Table 1: Three Logics of Analysis

	Logics of Analysis		
	Explaining	**Understanding**	**Recognising**
Question	Why?	How?	What?
Answer	Problem-solving	Subjective meaning	Pattern
Logic of action	Logic of consequences	Logic of appropriateness	Logic of regularity
Truth	Deterministic	(Inter-)subjective	Probabilistic

Source: Authors

2.2 Theory and Bias in Statistical Objectivity

Big data analyses often claim to be less subjective than standard social science. AIs trained with an extensive data set infer findings rather than interpret data. Not surprisingly, a transformative language model such as GPT3 that is trained on an extensive data set denies that it produces subjective interpretations but claims to be programmed to be "neutral and objective".[5] According to its reply, its trainers had been chosen according to criteria of diversity and non-discrimination and the text corpus it draws its replies from is rich with all kinds of data and perspectives. Its replies are thus not subjective interpretations but objective findings. It promises a scientific world in which unnecessary debates about facts and their meaning are reduced to writing prompts and reading replies. We no longer have to engage in endless debates about the proper interpretation of empirical facts and have, say, constructivists and realists debate their underlying ontological and epistemological premises in order to understand why Russia invaded Ukraine. We simply would ask the machine. Not surprisingly, the truth is far more complex.

Language-centred AI is usually trained with a text corpus that is necessarily limited and designed according to some selection criteria. Even if an AI had been trained with all data available on the Internet plus all analogue data in the world's libraries, it would still have a selection bias as it overlooks non-written data and all information that had not been published. This is far from banal as it encompasses, for example, classified information, marginalised voices and, last but not least, all interpretations repressed for political, religious or other reasons. Even the published data is biased regarding the perspectives it represents, as those interpretations of reality shared by most sources will be found most often in large text corpora. An algorithm working with a text corpus collecting all mainstream IR journals will most likely follow some combination of Realism and Liberalism and mix it with a good dose of Constructivism (and maybe a scent of Marxism and Feminism). On the other hand, an algorithm trained to follow a probabilistic concept of truth will identify those views held by most people with the "true" interpretation and thus will always reflect the points of view and the majority's interpretations. As we know from history, however, the majority can sometimes be highly subjective in its interpretation of facts.

Theory is also evident in the fact that NLP necessitates human input. Annotators must be trained by instructors and be taught how to interpret data. When data is simple to read and if

[5] Such a claim of neutrality is of course challenged by research that identifies ethical and other biases in such models (Weidinger et al. 2021).

unambiguous decisions are easy to adopt, then human input involves little interpretation. A decision about, for example, whether a picture shows a cat or a mouse will rarely demand much human interpretation. When interpreting sentences and deciding whether a specific sentence is a claim, a rebuttal or a warrant and if human language mixes different illocutionary components in one sentence, things become far more complex. Language-centred AI often works with some model of argumentation. Language models are essential because they help understand texts' semantic structure and establish connections between words and sentences. Models are always abstractions, however. They respond to a specific interest by highlighting some components, intentionally abstracting from others.

Two standard models of argumentation and arguments – which is where our key focus lies – are the narrative model (Bex and Bench-Capon 2014) and the Toulmin model (Toulmin 2003). The narrative model of arguing relies on storytelling to make an argument. This model emphasises personal anecdotes, vivid descriptions, and emotional appeals to persuade the audience. It is often used in persuasive speeches, advertisements, and political campaigns to connect with the audience on a personal level and create a sense of empathy and understanding. The Toulmin model is more formalistic. It consists of a claim, evidence, and a warrant (an assumption that connects the evidence to the claim). The Toulmin model also includes qualifiers (indicating the claim's degree of certainty or probability) and rebuttals (anticipating and responding to counterarguments). We are entering here the realm of objectifying subjective meanings, i.e., a practice that entails transforming a sentence with a meaning contingent on the specific interpretation of a reader into a sentence with a clear and unequivocal meaning independent of any subjective interpretation. In NLP, this practice is guided by a so-called gold standard. A gold standard is a theoretically informed definition of what is to be understood as the proper choice when annotating a sentence. It can be understood as a subjective categorisation that nevertheless claims objective status.

All these factors speak a clear language: Large language models and the AI-based on them require theory-guided decisions that can guide deliberate choices among options. The selection of the underlying corpus, the analytical reasoning model, and the formulation of a gold standard all require theoretically grounded guidance to avoid becoming arbitrary. Therefore, there is no end to theory but rather the need for a thorough theoretical reflection – and hence a moment of reflexivity (Neufeld 1993) – on the procedures of big data and AI-driven analyses. The following section demonstrates how an AI-based approach is set up to reflect upon the theoretical and methodological aspects discussed before.

3 Evidence: Insights From SKILL

The theoretical reflections reported above can be well observed in building an AI. The Social Sciences AI Laboratory for Research-Based Learning (SKILL) is a joint venture between the European New School of Digital Studies (ENS) and the Center for Teaching and Learning (ZLL) at the European University Viadrina, as well as the research units Web Technology and Information Systems (Webis) and Virtual Reality and Visualization Research (VR) at the Bauhaus University in Weimar. Together with students, it is building the most extensive annotated text corpus in the field of IR and is developing an AI-based argument search engine, a visualisation of argumentative structures, and an innovative didactic concept for using AI in higher education. The theoretical reflections discussed above present some discussions that have evolved in developing SKILL. The following section provides insight into how those debates resonate with the methods and data we employ in the project.

In the project, we start very traditionally by carefully choosing a research question to focus our empirical attention on in order to safeguard that any observed patterns would produce more than cats or dogs (sec. 3.1). In a second step, we develop a thick big data set by collecting a large volume of scientific articles, disaggregating it into individual sentences and adding layers of information on each of them (sec. 3.2), We finally train a group of annotators to train an algorithm with the power to conduct a systematic analysis of argumentative structures in IR-theories (sec. 3.3). In all three steps, theoretically informed decisions have to be adopted.

3.1 Research Question: The Migration of Arguments in IR

Arguments are a central component of both scientific and political debates. A good argument is key in theories, scientific writing and policy-making debates. It is crucial for political decision-making and necessary when engaging with the polity. In the SKILL project, we thus ask: what makes a theoretical argument convincing? Under what conditions does it migrate from one theory to another? Or even beyond academia and into politics? And how much do an argument's structure and quality matter for its reception by what kind of audience?

Those questions are hardly new. They have been the focus of a significant number of contributions, starting with Aristotle's Rhetoric more than 2500 years ago and running all the way down to more recent contributions in IR (Hanrieder 2011; Holzscheiter 2017; Müller 2004; Risse 2000; Zangl and Zürn 1996). Good arguments are considered a strong instrument in political negotiations (Johnstone 2011) and crucial in European politics (Neyer 2012). They are analysed as tools for facilitating political integration and have been described as the standard

modus operandi of democratic politics (Habermas 1997) and international relations (Müller 2004). In computer science, the rise of ML and NLP has inspired many works focusing on argument mining (i.e. the identification of arguments within a text) and the analysis of arguments (Kiesel et al. 2021; Al-Khatib et al. 2016).

Almost all of these contributions assume that arguments matter and that their reception in academic or political settings differs according to the merits of their quality. Relevant standards of quality include different features depending on theoretical provenance. Positivist epistemologies emphasise the empirical verifiability of claims and the repeatability of lines of evidence (King et al. 1994). Constructivist epistemologies emphasise the subjectivity of observation and underline the detailed and plausible reconstruction of meaning to make them comprehensible and thus understandable (Berger and Luckmann 1967 [1966]; Kratochwil and Ruggie 1986; see Jackson 2011 for an overview of different scientific logics in IR). Regardless of the respective scientific theoretical orientation, almost all authors agree that arguments have additional plausibility when supported by empirical evidence. Most authors also share the idea that empirical data only become relevant through their explicit integration into a theoretical context. They furthermore both assume that theoretical perspectives gain traction to the degree that they are explained through an explicit exposition of their premises. The idea that quality matters for arguments to be considered seriously also applies to scientific policy advice. When scientists advise policymakers, they usually assume that their arguments will be more likely taken into account when they comply with scientific standards.

However, the assumption of a high relevance of argumentation-specific features for their reception is not undisputed. Receptions within the scientific community might also be influenced by the integration of authors into established research networks (Risse et al. 2020) and sometimes even citation cartels (Teodorescu and Andrei 2014). Intellectually challenging positions that deviate from the majority opinion are easily ignored if particularly strong arguments and evidence do not back them. Complying with lower standards is often good enough for arguments replicating the mainstream. Thomas Kuhn has prominently pointed out that research programs have their internal logic, selectively receiving content based on whether it fits into dominant paradigms (Kuhn 1962). Despite high formal quality, arguments would be easily ignored if they ignored dominant understandings of problems and solution strategies (paradigms) and followed unorthodox trajectories.

For policy advice, the assumption applies analogously that policymakers only receive scientifically sound arguments if they can be reconciled with prevailing political calculations,

i.e., are politically opportune (Böcher 2022, Lepgold 1998). Lepgold describes "a communications gap between many IR theorists and practitioners" (Lepgold 1998, p. 43): whilst practitioners often feel pressure to make complex decisions under tight time constraints, academics are increasingly pushed by their community to engage in thorough reflection and to work on the leading edge of theoretical generalisation. The gap is further broadened by policymakers' desire for precise predictions about the results of possible policy choices. That is something, however, that no theory or academic reflection can provide. Social science works with abstractions and scope conditions often not met by reality. Practitioners often also have difficulty applying academic findings as most scholars rarely care whether their explanatory variables can be controlled by political intervention. For example, scholars interested in the forces leading to civil unrest might point to cultural heterogeneity, which is of only limited interest for someone trying to limit unpeace.

Luhmann's thesis of different societal functional systems, each with its language codes and rationality criteria (Luhmann 1984), also suggests that the idea of a search for truth that integrates functional systems and is based on argumentation is at least very optimistic: In science, knowledge is generated within the framework of disciplinary concepts and prevailing epistemological interests. It often sits squarely with the logic of politics in which solutions must be negotiated, and compromises will often be based on power asymmetries rather than claims to truth. Science also involves a continuous critique and problematisation of findings, thus inevitably rejecting any conclusive certainty. This irrevocable uncertainty in science is, in turn, difficult to reconcile with the expectation that policymakers can make effective decisions that inspire consent and confidence (cf. Böcher 2022).

The tension between the thesis of an argumentation-based dynamic of scientific discourse, on the one hand, and the indications of non-scientific factors influencing the reception of arguments, on the other hand, gives rise to two interrelated questions: What is the significance of the quality of a scientific argument for its reception and the change of another's opinion, and to what extent can a systematic connection between reception intensity and specific quality features of scientific arguments be empirically proven? Is there a connection between the two spaces of scientific and political communication, and if so, in what direction and under what argumentation-structural conditions do arguments migrate between discursive arenas?

3.2 Thickening Data

The power of AI Algorithms is primarily defined by the volume of data it has access to (bigness) and the amount of information (thickness) stored in the data. To establish a sufficient quantity of data, we set up a large corpus of high-quality scientific articles from leading English-speaking political science journals dealing with IR and international relations. The corpus comprises more than 2,000 articles, with approximately 600.000 sentences (assuming 300 sentences per article). Whilst the amount of data is comparatively easy to produce by simply giving the algorithm access to many articles, its thickness is trickier to realise. We can distinguish here between two strategies. A first sequential strategy produces thickness by adding a layer of qualitative research on top of a layer of quantitative research (Jemielniak 2020, p. 27). It starts with conducting large-scale and big data analysis of thin data. It analyses, for example, the geographical distribution of connections among Facebook users, their gender or ethnicity, or the frequency of contacts. In a second step, a sequential understanding of thickening data focuses on an especially interesting subset of data (for example, non-binary users in Alabama) and conducts thorough empirical ethnographic analysis. This strategy's benefit is bringing the strengths of both quantitative and qualitative analyses to bear on the data. However, it is a strategy that is difficult to employ when the research interest is tied to understanding patterns in a large data set and when no meaningful hierarchy among subsets of the data can be established.

A second strategy is more promising, which we suggest calling the coating strategy. We follow Wang (2013) and Latzko-Toth et al. (2017), who use the metaphor of the onion to explain the process of thickening data as adding layers of information. It is a framework based on the idea that data collected from human experiences is multi-layered and complex, much like the layers of an onion. To thicken data is to 'coat' them with additional layers of metadata – in the literal sense of data on data. Data are thus "trimmed, prepared and dressed" (Latzko-Toth et al. 2017, p. 203) before use. The process of annotating markables with domain knowledge and discourse knowledge can thus be understood as a process of "coating", i.e. thickening data by adding layers of information.

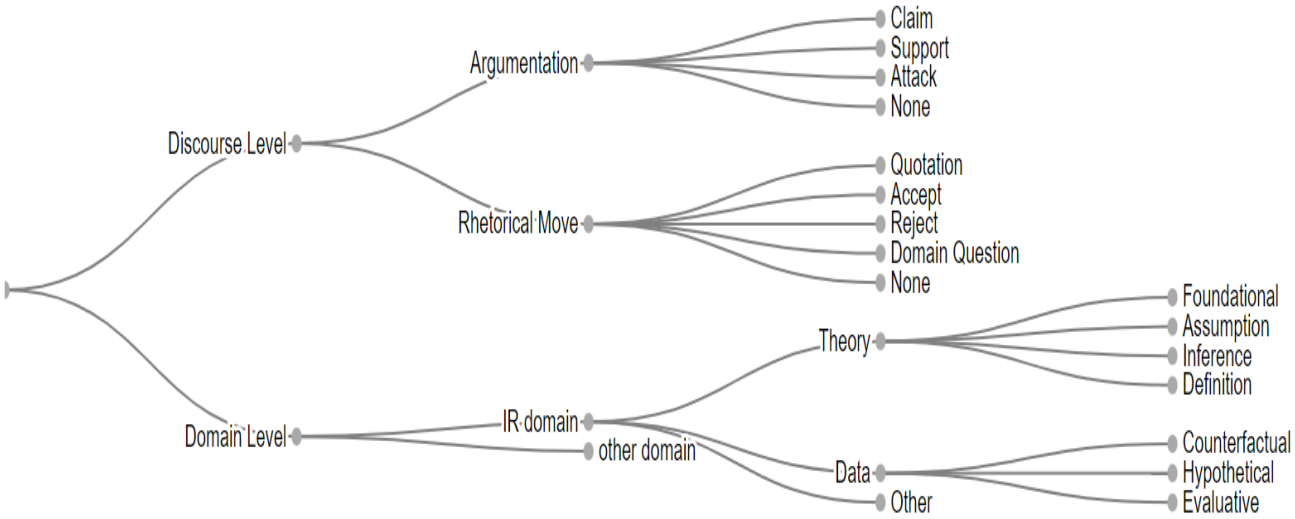
For guiding the thickening process, we take inspiration from an analytical framework for analysing and constructing arguments developed by Stephen Toulmin (Toulmin 2003). The model distinguishes between claims, grounds, warrants, rebuttals, backing and qualifiers and

provides a clear and structured framework for organising and evaluating arguments.[6] Each component of the model serves a specific purpose, which helps to analyse whether an argument is well-constructed and effectively presented. It also recognises that arguments are often complex and multifaceted and that multiple lines of reasoning may need to be considered to reach a conclusion. The model is not without its shortcomings, however. It offers a stylised image of an argumentative process that is hardly ever met in reality and is thus difficult to apply in practice. It speaks the language of an ideal type rather than a set of empirical categories and therefore needs to be adapted to the specific context in which it is applied.

To adapt it to the empirical analysis of a debate among theoretical arguments and to provide them with additional layers of information, we started with distinguishing between the two levels of discourse and domain. Annotation on the discourse level is separated into two branches: the argument level and the rhetorical level, which are both concerned with the illocutionary aspects of a sentence. Sentences can be a claim, support or an attack. Argumentations also have a rhetorical component and accept or reject some claim or involve a quote. The second level of thickening our data refers to its domain. In our case, this is IR. As sentences within IR scholarly work may also draw on other domains, annotators do also indicate if a sentence stems from another domain. On the domain level, the central distinction is between theory versus data. Sentences that fall in the theory category are distinguished according to whether they contain information that is foundational, present an assumption or an inference. Data markables are distinguished according to being counterfactual, hypothetical or evaluative.

[6] This model is often used in academic writing to help writers structure their arguments and make them more persuasive. While both, the Toulmin and the narrative model have their strengths, the Toulmin model is generally considered to be better suited for annotating. It provides a clear structure for identifying the different elements of an argument and evaluating their effectiveness. By breaking down an argument into its constituent parts, you can analyze each element separately and determine whether it is convincing and relevant to the overall argument. In contrast, the narrative model is less apt for annotation because it relies on subjective elements such as personal experiences and emotions. While narratives can be powerful and persuasive, they are often less structured and more difficult to evaluate.

Figure 1: Revised Category System



Source: Categories: The Authors. Tree design: Dora Kiesel, Bauhaus Uni Weimar

The simplified and adapted model is applied sentence by sentence to allow for fine-grained text analysis. For example, different sentences may contain different information; annotating each separately can help capture these nuances. Another strength of annotating sentence by sentence is to increase clarity which parts of the text correspond to each annotation. This can make it easier to communicate the annotation results to others and maximise inter-annotator reliability. Every sentence is thus attributed a certain meaning and provided with additional layers of information that allow the algorithm to identify it accordingly.

The decomposition of texts is not unconditional, however. Provisional or trailing sentences are used as an additional resource of information for annotation if they provide important information without which sentences cannot be properly understood. The process of decomposing texts into sentences is also contextualised by adding relationships between sentences. Sentences that refer to each other and provide an explicit argumentative context are connected graphically, indicating a relation between them. For example, if sentence one contains a claim and sentence two lists the supporting evidence, then both sentences are annotated as relating to each other.

It is important to realise that there are limits to the thickness of a NLP-based method. According to Geertz (1973) thick data refer to qualitative data rich in context, emotion, and meaning. It is often gathered through ethnographic or other qualitative research methods in which interviews have been conducted or participatory observation has been employed. A “thick description” involves going beyond the surface-level meanings of an event or behaviour and seeking to

understand its deeper cultural and symbolic meanings. Rather than simply describing what happened, thick description seeks to understand why it happened and what it means within its cultural context. Thick description thus involves a careful and nuanced approach to understanding cultural phenomena, recognising that the meanings and significance of events are deeply embedded in their cultural context. It emphasises the importance of understanding the broader social and cultural systems that give rise to specific behaviours and practices rather than just focusing on the behaviours themselves.

Algorithms trained on large text corpora can only, to some degree, process thick data. As they have no direct experience or understanding of human culture and social dynamics, their understanding of reality is limited to applying the parameters they have been trained to apply. The thickness of data that can be processed is defined by the set of categories being applied in annotation. Important contextual information that is not reflected in the text itself, such as nonverbal cues or situational factors, which may be crucial for fully understanding thick data, and is not added by means of annotation, will thus get lost. An adequate term for the method used in SKILL might thus be "medium-grained". It refers to an analysis that is more detailed and nuanced than "thin" data or analysis but not as detailed or nuanced as "thick" data or analysis. Medium-grained data promise to capture important nuances and complexities of a phenomenon while allowing to process a large amount of data.

3.3 Objectifying Subjectivity in Practice

The thickening of data involves acts of interpretation. When adding layers of information, those who add the information must identify the proper analytical categories. Interpretative processes, however, are necessarily infected with much subjective assessment. That becomes an important issue in every annotation process as a rather large team usually conducts it with significant interpretative variance. In addition, annotators have different educational, cultural and other backgrounds and will most likely show different practices for interpreting sentences. In order to mitigate those differences, annotation is preceded by intensive training of the annotators in which specific rules for interpretation are established, and a commonly shared understanding of the analytical concepts (layers of information) is developed.

This often-cumbersome process of realising inter-annotator reliability involves the establishment of a so-called gold standard. A gold standard is a reference dataset used as a benchmark for annotator performance. In this gold standard, the trainers define a specific mode of annotating sentences to convey the underlying principles to the annotators so that they can

understand and apply them autonomously.[7] As soon as the annotators have learned to comply with the gold standard, they start to annotate individually and thus train the algorithm. Constant checks of gold standard comparison and inter-annotator reliability remain essential for ensuring sufficient annotation quality. In this step, the algorithm learns to identify arguments relating to theory-specific propositions, to tell, for example, an assumption from an empirical reference and to distinguish between different types of empirical references. In the final step of this process, the algorithm is given access to the full-text corpus and starts to (semi-)autonomously annotate and learn from its successes and failures. It will be closely guided by the annotators and monitored to see if the annotations comply with the gold standard.

This third step leads to a sizeable argumentative repertoire of the algorithm and, thus, significant usability. The repertoire should allow both the systematic search for arguments by users and infer statements about correlations of domain level features and illocutionary arguments. This opens a promising way for answering the research question about the relevance of successful, i.e. persuasive arguments and their domain- and illocutionary features. At a later stage, the algorithm may then be applied to a larger corpus of IR journals or even different corpora, such as debates in the United Nations or the European Union.

4 A New Research Agenda

SKILL bridges between qualitative and quantitative approaches. It combines large pattern recognition with thickened data. It establishes a text corpus collecting thousands of journal articles and disaggregates them into individual sentences (so-called “markables”). These approximately 600,000+ sentences carry a subjective meaning and are being interpreted as performing a specific function in terms of domain knowledge and argumentative status. The big data component of SKILL is being mastered by an ML-driven algorithm recognising the semantic patterns of, for example, inferred theoretical statements about processes or empirical illustrations. Having been trained to identify those patterns, it can process vast amounts of data and uncover many other cases of assumptions, inferences or empirical statements. It can also

[7] The practice of annotation is trained in the beginning with four central texts that are characteristic for the four theoretical perspectives of neorealism, liberalism, constructivism and feminism. They are: Kenneth N. Waltz: The Emerging Structure of International Politics, International Security Vol. 18, No. 2 (Fall, 1993), pp. 44-79; Robert D. Putnam, Diplomacy and Domestic Politics: The Logic of Two-Level Games, International Organization, Vol. 42, No. 3 (Summer, 1988), pp. 427-460; Finnemore, Martha; Sikkink, Kathryn (1998): International Norm Dynamics and Political Change. In International Organization 52 (4), pp. 887–917 and Zalewski, Marysia (1995): 'Well, What is the Feminist Perspective on Bosnia?'. In *International Affairs* 71 (2), pp. 339–356.

set them in context and will be able to combine assumptions, inferences and data if they belong to a specific theoretical perspective or contradict it.

4.1 The Theory in AI

In discussing the theory and practice of building an AI, we hope to have demonstrated that it cannot be thought of independently of theoretical reflections. Theoretically sound decisions must be made in all phases of the development of an AI. Willingly or not, theory is everywhere in AI-centred social science research.

(1) The selection of the text corpus reflects a conscious decision about the relevant data material. The project opted for integrating the most established data sources (i.e. journals from ´prestigious publishers like Oxford University Press, Cambridge University Press, Sage, and Taylor and Francis). In doing so, it deliberately applied a theory of relevance which prioritised reputation in the mainstream academia of the Western world over other criteria. By implication, it abstained from integrating journals from non-mainstream backgrounds. As a result, the text corpus de facto excludes Asian or African Journals unless published in English by one of the publishers listed above. This decision can and ought to be criticised. It is biased in its repertoire of arguments and has little to say on, for example, postcolonial approaches or on reporting arguments that do not fit with the established debates. It also is relatively weak in collecting arguments that do not comply with the established standards of proper science, such as beginning with a clear-cut research question, providing a state of the art, etc. pp. Its theoretical bias is thus strongly mainstream.

(2) SKILL operates with an abstract concept of arguments which is not only a rough simplification of the complexity of reality as it overlooks rhetorical and emphatic components of speech acts (Aristotle famously distinguished among logos, ethos and pathos as essential components of speech acts). It thus interprets reality by highlighting some aspects and abstracting from others. It is also necessarily selective in its understanding of the world as it is sensible only to those interpretations of reality which can be analysed in the given categories of a model of argumentation. Art, music, and all other non-deliberative types of expressing interpretations of the meaning of facts are necessarily ignored or underrated. The selection of the Toulmin model expressed a conscious decision for a deliberative and scholarly mode of articulating insights and opinions. The language model is closely tied to deliberative concepts of argumentative interaction and in contrast to all approaches which criticise deliberation as a different term for academic elitism. This choice is justified because we are interested only in

scholarly debates. There is, therefore, nothing wrong with it. However, it is essential to remember that a different model might yield different results.

(3) The gold standard used in SKILL depicts a process of agreement between three domain experts who have made their individual annotation practices the standard for the development of the AI. All of them have been educated in well-established universities and socialised in mainstream research institutions such as the European University Institute in Florence or the Wissenschaftszentrum Berlin. Whilst they represent both genders and have significant international experience in US and European contexts, they lack a comparative non-Western background and related understandings of how to interpret text. While this setting may be well justified for pragmatic reasons, it will most likely produce interpretative bias. The gold standard represents a subjective rationality that claims hegemonic status by means of academic hierarchy.

4.2 New Opportunities for Research

This methodology introduced in this paper promises to inspire innovative IR research on four levels. First, on a metatheoretical level, it challenges the hegemony of the two paradigms of explaining and understanding. As opposed to these two epistemologies, it holds that the seemingly naive question of "what is" might be more promising to pursue than most scholars would have it. A recognition-based approach uses large data sets to infer patterns and structures of social reality that are hard to challenge empirically. Third, it offers an innovative methodology which can answer both critiques raised by critical rationalism and interpretivism. In reaction to the claim of critical rationalism that only falsification can overcome subjectivism and leads to proper insights, it submits that data speak an objective language if they are only big enough. In reaction to interpretative approaches, it holds that big data can be thick if analysed through a theoretically informed lens. The preoccupation of much of IR with a theoretical debate that has progressed only marginally since the times of Thukydides and that still seeks answers to the same old questions can hardly suffice. A recognition-based approach opens the debate for a new epistemology which brings innovative ML/NLP tools to bear on the pressing questions of the discipline.

Second, combining NLP and ML promises to provide sociological insights into the discipline. The major theoretical debates in IR are about the origins of war and peace, cooperation and conflict, and the power of ideas versus those of fixed interests. All those questions have been treated ever since by theories emphasising either state power and national interests or normative

ideas and social forces. Although both sides have become more sophisticated over time and elaborated their argument ever more, no consensus has emerged. Whilst some might interpret this as giving expression to the fact that both sides have a point, a recognition-based approach allows for a more critical approach to explaining the limited progress in the debate. It can use NLP methodology for scrutinising the large amounts of arguments and counterarguments raised over time in pertinent journals and identify whether the theoretical opponents have listened to each other. Did Mearsheimer take Wendt seriously, and vice versa? Have they learned from each other? If so, what have they learned, and what have they ignored? Suppose it were the case that scholarly arguing was a performative practice that hardly ever saw prominent participants take their opponents arguments seriously enough to consider changing their opinion. In that case, one might be tempted to conclude that progress in scholarly debate is indeed hard to foresee. Such a finding would throw a challenging light on the practices of IR and invite demands for methodologies that allow data to speak for themselves.

NLP and ML can thirdly produce significant new insights for some of the recent theoretical debates. For example, it can contribute to understanding arguments' role in international politics (Müller 2004; Risse 2000). Large text corpora like the United Nations General Assembly meeting records can be compared to social scientific journals. By extracting arguments from speeches (and scientific articles), whether and what arguments are employed by whom can be analysed. An annotated text corpus combined with an argument search engine could identify how, where and what kind of arguments travel between academia and politics (cf. Lepgold 1998; Adler-Nissen et al. 2021). It could answer questions of direction, i.e. whether politics learns from academia, whether it is the other way around or whether both social spaces are basically self-referential. Where do the arguments come from that are employed? Or do arguments follow interests and relationships of power only and show no connected patterns between academia and politics?

A NLP/ML approach would finally be helpful for understanding and explaining pertinent conflicts. It could, for example, analyse major newspapers in democracies and ask about the impact of public opinion on foreign policy. Such analyses have indeed been undertaken before. However, the power of ML and NLP-based analysis of vast text corpora has never been employed for inferring patterns. Such an approach would not be limited to a set of newspapers and a specified period but could read all newspapers in all democracies and ask for the connectedness of changing sentiments in op-eds and shifts in foreign policy. If the algorithm has only been trained to identify the relevant structures in texts, it could do so with little extra effort across countries, policies, and time.

In combination with NLP, ML represents a methodological innovation whose potential can hardly be foreseen today. SKILL promises to make quantitatively corroborated statements about the argumentative content of the discipline of IR and will be capable of describing basic patterns of how and under which conditions arguments migrate across theories, journals and time. In perspective, many other applications of NLP, ML and Big Data can be imagined in IR. The annotated text corpus generated within the framework of SKILL is used to develop an argument search engine that can answer a large number of inquiries in a meaningful way and thus massively facilitate the research work of students and scientists. It can also be extended in perspective to journals from other regions of the world and thus allow systematic access to knowledge that promises to open the cognitive container of a discipline still dominated by US and European contributions.

Publication bibliography

Adler-Nissen, Rebecca; Eggeling, Kristin Anabel; Wangen, Patrice (2021): Machine Anthropology: A View from International Relations. In *Big Data & Society* 8 (2), 1-6. DOI: 10.1177/20539517211063690.

Al-Khatib, Khalid; Wachsmuth, Henning; Kiesel, Johannes; Hagen, Matthias; Stein, Benno (2016): A news editorial corpus for mining argumentation strategies. In ACL Anthology (Ed.): Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. COLING 2016, the 26th International Conference on Computational Linguistics. Osaka, Japan, pp. 3433–3443. Available online at https://aclanthology.org/c16-1324/.

Anderson, Chris (2013a): Das Ende der Theorie. Die Datenschwemme macht wissenschaftliche Methoden obsolet. In Heinrich Geiselberger, Tobias Moorstedt (Eds.): Big Data: Das neue Versprechen der Allwissenheit. 2nd ed. Berlin: Suhrkamp Verlag, pp. 124–130.

Anderson, Chris (2013b): The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Available online at https://www.wired.com/2008/06/pb-theory/, updated on 6/23/2022, checked on 9/11/2022.

Berger, Peter L.; Luckmann, Thomas (1967 [1966]): The Social Construction of Reality. First Anchor Books Edition. New York: Anchor Books.

Bex, Floris; Bench-Capon, Trevor (2014): Understanding narratives with argumentation. In Simon Parsons, Nir Oren, Chris Reed, Federico Cerutti (Eds.): Computational models of argument. Amsterdam: IOS Press (Frontiers in artificial intelligence and applications, 266), pp. 11–18.

Böcher, Michael (2022): Wie funktioniert wissenschaftliche Politikberatung? Available online at https://www.forschung-und-lehre.de/politik/wie-funktioniert-wissenschaftliche-politikberatung-4759, updated on 6/2/2022, checked on 2/6/2023.

Chatsiou, Kakia; Mikhaylov, Slava Jankin (2020): Deep Learning for Political Science. In : The SAGE Handbook of Research Methods in Political Science and International Relation: SAGE, pp. 1053–1078.

Geertz, Clifford (1973): The interpretation of cultures; selected essays: Basic Books.

Geiselberger, Heinrich; Moorstedt, Tobias (Eds.) (2013): Big Data: Das neue Versprechen der Allwissenheit. 2nd ed. Berlin: Suhrkamp Verlag.

Habermas, Jürgen (1997): Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy: Blackwell Publisher.

Hanrieder, Tine (2011): The false promise of the better argument. In *International Theory* 3 (3), pp. 390–415. DOI: 10.1017/S1752971911000182.

Hollis, Martin; Smith, Steve (1990): Explaining and Understanding International Relations. Oxford: Clarendon Press.

Holzscheiter, Anna (2017): Was vom arguing übrigblieb… Der Nachhall der kommunikativen Wende in den Internationalen Beziehungen. In *Zeitschrift für Internationale Beziehungen* 24 (1), pp. 143–159. DOI: 10.5771/0946-7165-2017-1-143.

Hudson, Valerie (1991): Artificial Intelligence and International Politics. Boulder: Westview.

Jackson, Patrick Thaddeus (2011): The conduct of inquiry in international relations. Philosophy of science and its implications for the study of world politics. London, New York: Routledge (The new international relations).

Jemielniak, Dariusz (2020): Thick big data. Doing digital social sciences / Dariusz Jemielniak. Oxford: Oxford University Press.

Johnstone, Ian (2011): The power of deliberation. International law, politics and organizations / Ian Johnstone. New York, Oxford.: Oxford University Press.

Kiesel, Dora; Riehmann, Patrick; Wachsmuth, Henning; Stein, Benno; Froehlich, Bernd (2021): Visual Analysis of Argumentation in Essays. In *IEEE transactions on visualization and computer graphics* 27 (2), pp. 1139–1148. DOI: 10.1109/TVCG.2020.3030425.

King, Gary; Keohane, Robert O.; Verba, Sidney (1994): Designing social inquiry. Princeton, N.J., Chichester: Princeton University Press.

Kratochwil, Friedrich V.; Ruggie, John G. (1986): International Organization: A State of the Art on an Art of the State. In *International Organization* 40 (4), pp. 753–775.

Kuhn, Thomas S. (1962): The structure of scientific revolutions: Chicago Univ. Press.

Latzko-Toth, Guillaume; Bonneau, Claudine; Millette, Mlanie (2017): Small Data, Thick Data: Thickening Strategies for Trace-based Social Media Research. In Luke Sloan, Anabel Quan-Haase (Eds.): The SAGE Handbook of Social Media Research Methods. London, Thousand Oaks, New Delhi, Singapore: SAGE, pp. 199–214.

Lepgold, Joseph (1998): Is Anyone Listening? International Relations Theory and the Problem of Policy Relevance. In *Political Science Quarterly* 113 (1), pp. 43–62.

Luhmann, Niklas (1984): Soziale Systeme: Grundriss einer allgemeinen Theorie. Frankfurt am Main: Suhrkamp.

Mayer-Schönberger, Viktor; Cukier (2013): Big Data: Die Revolution, die unser Leben verändern wird. 3. Auflage. München: Redline Verlag.

Müller, Harald (2004): Arguing, Bargaining and All That: Communicative Action, Rationalist Theory and the Logic of Appropriateness in International Relations. In *European Journal of International Relations* 10 (3), pp. 395–435.

Müller, Thomas; Ritschel, Gregor (2016): Big Data als Theorieersatz? In Thomas Müller, Gregor Ritschel, Alexander Amberger, Stefan Böschen, Roland Broemel, Ulrich Busch et al. (Eds.): Big Data als Theorieersatz. Berliner Debatte Initial 4/2016. 1. Auflage. Potsdam: WeltTrends (Berliner Debatte Initial, (2016) 4), pp. 1–8.

Nassehi, Armin (2019): Muster: Theorie der digitalen Gesellschaft. München: C.H. Beck.

Neufeld, Mark (1993): Reflexivity and International Relations Theory. In *Millennium: Journal of International Studies* 22 (1), pp. 53–76.

Neyer, Jürgen (2012): The justification of Europe. A political theory of supranational integration / Jürgen Neyer. Oxford: Oxford University Press.

Rahal, Charles; Verhagen, Mark; Kirk, David (2022): The rise of machine learning in the academic social sciences. In *AI & SOCIETY*. DOI: 10.1007/s00146-022-01540-w.

Risse, Thomas (2000): 'Let's Argue!': Communicative Action in World Politics. In *International Organization* 54 (1), pp. 1–39.

Risse, Thomas; Wemheuer-Vogelaar, Wiebke; Havemann, Frank (2020): Theory Makes Global IR Hang Together. Lessons from Citation Analysis. Available online at http://dx.doi.org/10.17169/refubium-28510, updated on 11/6/2020, checked on 9/11/2022.

Shoub, Kelsey; Olivella, Santiago (2020): Machine Learning in Political Science: Supervised Learning Models. In : The SAGE Handbook of Research Methods in Political Science and International Relation: SAGE, pp. 1079–1094.

Sylvan, Donald A.; Chan, Steve (1984): Foreign Policy Decision Making. Perception, Cognition and Artificial Intelligence. New York: Praeger Publishers.

Teodorescu, Daniel; Andrei, Tudorel (2014): An examination of "citation circles" for social sciences journals in Eastern European countries. In *Scientometrics* 99 (2), pp. 209–231. DOI: 10.1007/s11192-013-1210-6.

Toulmin, Stephen (2003): The uses of argument. Updated ed. Cambridge, New York: Cambridge University Press.

Wæver, Ole (2010): Still a Discipline After All These Debates? In Timothy Dunne, Milja Kurki, Steve Smith (Eds.): International Relations Theories: Discipline and Diversity. 2nd edn. Oxford: Oxford University Press, pp. 297–318.

Wang, Tricia (2013): Big Data Needs Thick Data. Available online at http://ethnographymatters.net/blog/2013/05/13/big-data-needs-thick-data/, updated on 5/13/2013, checked on 3/6/2013.

Weidinger, Laura; Mellor, John; Rauh, Maribeth; Griffin, Conor; Uesato, Jonathan; Huang, Po-Sen et al. (2021): Ethical and social risks of harm from Language Models. Available online at https://arxiv.org/pdf/2112.04359.

Wendt, Alexander (1998): On Constitution and Causation in International Relations. In *Review of International Studies* 24 (The Eighty Years' Crisis 1919-1999, December 1998), pp. 102–117, checked on 1/14/2020.

Zangl, Bernhard; Zürn, Michael (1996): Argumentatives Handeln bei internationalen Verhandlungen: Moderate Anmerkungen zur post-realistischen Debatte. In *Zeitschrift für Internationale Beziehungen* 3 (341-366).