A Glimpse into the Future

10.2478/abcsj-2023-0026

Big Data and (the New?) Reality

MANUELA MIHĂESCU
Babeş-Bolyai University, Cluj-Napoca, Romania

Abstract
The management of enormous volumes of data, cloud computing technologies, the Internet of Things and all the smart devices we use are strongly related to the *Big Data* phenomenon. Due to the massive digitalization performed in recent years, almost everything, every action and every relationship is digital, transformed into data that can be analyzed and then turned into valuable information. Some of the great benefits of these processes relate to predictive analytics or various studies to understand social and cultural dynamics, aspects which have been successfully exploited by large companies such as Amazon, Google, Microsoft, or Netflix. This article examines Big Data analysis and discusses some challenges that arise from the integration of Artificial Intelligence (AI) in this process. In this digital decade, the influence of Big Data on our understanding of the real world is becoming a key factor. The article also highlights some of the concerns researchers have about the role and the use of AI and the algorithms behind AI in certain contexts. The contribution concludes with a reflection on how data processing, in this digital decade, could affect the way in which we relate to reality. Will it expand *our* knowledge of the real world, or will it increase *our* immersion in the digital world?

Keywords: data, Big Data, AI, information, data exploitation, predictive analytics, digital world, real world

> "*This is the new normal*. . . . The Big Data Era will be led by the individuals and companies that deliver a platform suitable for the new normal – a platform consisting of exploration and development toolsets, visualization techniques, search and discovery, native text analytics, machine learning, and enterprise stability and security, among other aspects. Many will talk about this, few will deliver." (Zikopoulos et al. xvi)

About thirty years ago, speaking about the *explosion of new technologies* (translated as "The End of Continuity") and the transformation of society into a "world economy," Peter Drucker concludes his book *The Age of Discontinuity: Guidelines to Our Changing Society*, with the rhetorical question: "Does Knowledge Have a Future?" Among the extremely interesting topics covered in the book, a particularly important one is that of discontinuity, which the author explains in a 'visionary' manner:

> Discontinuity is what I called these major changes in the underlying social and cultural reality. . . . Revolution is the earthquake or the volcanic eruption that obliterates the familiar landscape and creates a new one. But these revolutions are largely the effects of shifts in the foundations that precede them and make the revolution inevitable. These revolutions result from discontinuities, from the buildup of tension between a new underlying reality and the surface of established institutions and customary behavior that still conform to yesterday's underlying realities. And while revolutions tend to be violent and spectacular, *discontinuities tend to develop gradually and quietly and are rarely perceived until they resulted in the volcanic eruption or the earthquake*. (Drucker xii, emphasis mine)

By association, and through a semantic expansion of the term "discontinuity," the phenomenon that Drucker talks about is, in fact, identical to the one we are experiencing today: the phenomenon of Big Data, which developed subtly and silently, and then "exploded" especially after the Covid-19 pandemic, with the massive digitization of resources and actions, when everything metamorphosed "into data."

The advance towards a "data-driven society" has far-reaching implications, compounded by the unprecedented development of AI. In addition to the important opportunities for innovation and development, there are real challenges related to the management and, above all, to the

ethical use of data that is becoming critical for individuals, organizations and governments.

This article offers a very brief introduction and description of the concept of Big Data, continues with a discussion of the implications of Big Data analysis and especially of some challenges that arise through the integration of AI in Big Data analysis, followed by a reflection on the impact that the processing of this data can have on human behaviour, and on the perception of reality (or what we still perceive in the traditional sense of "reality").

What Is Big Data?

Big Data is not a new term; it has been used with the general meaning of "large data collection" since the '90s (Lohr; Diebold). A brief investigation into the etymology of the term shows that the moment of terminologization is not very clearly established; however, it can be noted that the evolution of the use of the concept coincides with the technological evolution and the massive digitization of society.

It is useful to specify, from the beginning, that not every large collection of data can be called "Big Data." In the specialized literature, we find many descriptions of the term, various authors trying to explain it, to clarify ontologically its conceptual delimitations, functionality and uses depending on its characteristics. Gartner IT Glossary defines the concept referring to quantities of data that have certain specificities (high volume, high speed, high variety) and that cannot be processed or analyzed using traditional processes or tools. The attribute of "big" being relative, the main questions are how big these amounts of data should be, how big should the processing speed be and what "data variety" means exactly.

Over time, the evolution of the concept's meaning shows that not all big data (large collections of data) have all three features (*volume*, *velocity*, *variety*, or the 3'Vs'). The multitude of fields in which we talk about Big Data, the applications and uses of the concept are only some aspects that underlie the diversity of the meanings. But, despite the slight variations in the explanation of the concept, the three characteristics are found in many authors' works. However, in the case of certain large data collections, the

volume, variety or velocity are not sufficient to describe the complexity of the concept, so two more characteristics were added related to the problem of processing and analysis of this data: *the veracity* (authenticity, credibility) of the data and *the value*, or the 5Vs (Hadi et al.; Taylor-Sakyi). Other studies (Kitchin and McArdle) also refer to: exhaustiveness, resolution, indexicality, relationality, extensionality and scalability, important characteristics in processing some types of Big Data, from different areas (linguistics, medicine, communications, transport, economy, security, etc.). Since the number of specialized studies and articles in the field has significantly increased in recent years, other characteristics have been brought into discussion (e.g. *Volume growth rate*, *Volume change rate*, and *Variance in volume change rate*), and there are studies that talk about 7Vs, 10Vs, or, even 12Vs, characteristics that help us understand the complexity of the phenomenon, but also allow a more refined adaptation to the "customer value" and "customer-centric" culture (Loshin and Reifer). Other conceptual clarifications and explanations of Big Data can be found in Zikopoulos et al., Gandomi and Haider, Özköse et al., Schiuma and Carlucci.

Summing up the complexity of the concept's definition and range, we could note that the notion of Big Data is related to the collection, storage and processing at very high speed of a huge volume of data, very heterogeneous and complex (in terms of structure and provenance), as well as the technology involved in their processing. The explosion of data volume is based on the development of the information and communication technology infrastructure and the massive (accentuated, even enforced) digitization induced by the Covid-19 pandemic.

What Exactly Do the 5Vs Mean?

The '5Vs' currently represent the most widely used set for describing the key characteristics or dimensions of Big Data.

Volume. We live in a technological world, a world where three quarters of the population own at least one mobile phone (ITU). Even though not all mobile phones are "smart," the vast majority of users have access to the Internet or own other smart devices such as tablets, watches,

video cameras, cars or various household appliances, all of these containing sensors and generating data all over the planet. According to *CIO Bulletin*, "by 2023, *each person* will have an average of 1.6 networked mobile devices and connections." A huge amount of data comes from all these sensors and is collected and stored for different periods of time. The volume of Big Data refers to this huge amount of data that is obtained from transferring data between these devices, but also from the communication between users through posts on social networks (Facebook, YouTube, WhatsApp, Instagram, TikTok, etc.), or by emails and other ways of interaction mediated by technology. If up until now we were talking about GB of information (10^9 bytes), when we refer to Big Data today, the order goes from zettabyte (which means 10^{21} bytes) to yottabytes (1 yottabyte is 1000 zettabytes), a value that the human mind is not yet ready to accept and certainly cannot represent! And with generative AI, this trend is growing. Statista.com (*Volume of data*) shows that in the year 2022 the volume has grown to 97 zettabytes of data, and, according to their estimations, it will increase to 180 zettabytes by 2025.

Velocity. Data is produced and needs to be processed at very high speeds, mostly in real time. Digital devices, especially IoT (Internet of Things),[i] and the introduction of AI algorithms (and generative AI), produced a lot of data with high speed and frequency, in a unit of time (per second). The streaming data provide a lot of information not only about users, in real time, but also about all the actions or phenomena that take place in the world. This feature is very important for governments, institutions and companies, which need data to be quickly available.

Variety. Data (in Big Data) can be found in various forms. Some are structured (organized in databases/databanks, forms, tables, collections or large archives, etc.), in which case they can be interpreted and processed with relative ease. Others are semi-structured data that do not have very clear or easily identifiable principles of organization. However, the vast majority are unstructured (lacking a preset data model), namely those data obtained from messages/emails, posts on different networks, images or videos, data collected from various sensors, etc.

Veracity refers to the validity of the sources of data. Both volume and speed have a strong impact on veracity. The provenance of very large

volumes of data processed at high speed is extremely difficult to verify. Incorrect data can generate various problems, aspects that are extremely worrying, especially when they are used in the decision-making processes by governments, companies or other influential factors. In addition, because humans are not directly involved in the processing (most processes being automatic), the need to ensure the validity of sources of data is paramount (Rockwell and Sinclair[j].

Value. One of the characteristics of Big Data concerns the possibilities of exploitation (analysis and interpretation), which grow more and more sophisticated. Big Data processing is extremely valuable in statistical analyses, but especially in predictive analyses, in identifying patterns, behaviours, or trends, in simulating scenarios, etc. The potential of collection and the use of data has come to the attention of specialists in the economic field (especially marketing) since the 2000s (Marr, *Big Data: Case Study;* Auditore), when competitions for markets and customer options led companies to collect all kinds of data about users. From purchases to forms containing personal data, financial transactions, posts on social networks, phone conversations, etc., all of these are "digital traces" that define (with fairly high accuracy) the profile, actions and preferences of an individual. Today, this potential is at least doubled by the introduction of cloud technologies and AI into their analysis and is used (nearly monopolized), by certain institutions or by large companies, as not everyone has the skill and the technology to exploit it. Getting (new) information, but also making the business more efficient, and therefore more profitable, are some of the major benefits of processing Big Data.

How Is Big Data Produced and Collected?

Throughout history, the collection of data has been a constant reality of research. The only difference is that in the past data were not accessible to everyone, nor were there adequate tools to process them. In the last years, with a large scale of digital transformation, data became more "visible," its stored and circulating volume has increased, exceeding by far not only the processing capacity of ordinary computers, but also our ability to perceive, to be aware of this massive amount of data.[ii]

Scientific research has long relied on data samples. But when we talk about Big Data, we mean *all* the data that can be collected, integrated and analyzed, and this radically changes the area and the way of investigation and analysis, but especially the statistics or predictability. If, until recently, processing was mainly carried out by humans, helped by various statistical analysis programs, in the case of Big Data this is automated and can provide more extensive analyses (we mentioned the exhaustiveness feature before) and precise and objective results, as a natural consequence of processing very large volumes.

Data production and collection is mainly done in two ways. First of all, data is produced and collected by (official) institutions, companies, and organizations, and it is obtained and recorded either automatically (in most cases) or by humanly managed registration using various methods (resources, forms, surveys, etc.). This is a type of data from which a lot of information, more or less confidential, can be quickly extracted (population records, medical files, files with various information about professional career, etc.). In contrast to previous years, when they were stored in a variety of places and systems, new technologies allow the integration and aggregation of various systems, so that they can be accessed and correlated, in a dynamic way, thus obtaining much more complete and complex information about a person, various actions or large-scale phenomena. One of the largest companies in the world, Amazon, collects data from all users and can provide a lot of information and suggestions about user/costumer preferences. "Just wait. We'll be sending you coupons for things you want before you even know you want them" is the title of an article by Charles Duhigg, where the author explains how retailers subtly track consumer habits to offer them products, anticipating future purchases based on preferences. Similarly, the data collected by Microsoft, through the software they produce, can indicate a user's activity profile by all his/her online activity. Google, one of the most profitable companies today, collects a lot of data about the users' activities, which is then capitalized for example through the advertising system (Marr, *Big Data: Case Study* 4). Banks are big collectors of personal data and can reconstruct the financial/banking profile of each user. Netflix, HBO, and Booking.com are among the companies that collect data from millions of users and can offer

suggestions based on each person's profile/preferences (for travel, accommodation, etc.). The data is then used for statistics, analysis of travel patterns, price setting for hotel rooms, calculation of routes, or calculation of risks. This category may also include the data that exist in large collections, archives, thesauruses, repertoires, international data banks, dictionaries, various resources such as Wikipedia, Google books, translation memories, parallel texts, and which is extremely useful in scientific research, in correlation with other databases, to find more or less common patterns.

The second type of production and collection is automatic, mainly through the IoT system, by automated generation/reading of data, in real time, with very high frequency, at regular time intervals. This is mostly unstructured data and can come from human interactions with devices, from all objects equipped with sensors, or from the interactions between devices. Therefore, a Google search, a WhatsApp message, a tweet on Twitter (currently known as X), a post on Facebook, YouTube or TikTok, a photo on Snapchat, a call on Skype, an online purchase, a GPS location, a Zoom meeting, an interaction with a virtual assistant, a mail, a tag, etc., generate a lot of data every second (for the volume of data generated per day, in 2023, see Wise). *Practically: a click, a movement and, more recently, a look, can generate data*.

Data is collected and stored in massive databases ("data warehouses") or cloud solutions. According to Statista.com (*Number of data centers*), the numbers of storage centres by country in 2022 are: US-2701, Germany-487, UK-456, China-443, etc. Data is collected *from all users*, and then a small number of companies use it to better understand behaviours, preferences and trends.

All this data is processed with a new generation of technologies and architectures designed to analyze and extract information from these huge and complex volumes (Apache Hadoop, Tableau, NoSQL databases, etc.) (Zikopoulos et al.). Although some are free and widely available, they are not easily accessible to everyone.

We are the digital citizens, anonymous in principle, but leaving an impressive trail of data, not only about what we do, but also about what we think and even what/how we feel. That is why it is important to remember

that we ourselves "feed" the growing data collection, every day, through personal data provided (consciously), in different situations, or through data collected/generated automatically, by tracking online activity, or collected from sensors, as mentioned above. As stated in privacy regulations and policies, for the most part, data is anonymized when it is collected, but there are many studies that show that "re-identifying" it is not impossible.

On the other hand, this trend of "anonymization" and transfer towards "unification" and "standardization," practically towards a "global digital memory," which will be able to delete/modify the data or an "individual memory" at any moment, can have far more profound consequences over time.

Big Data analyses with accurate, valid data can help us understand, communicate and function better as individuals and as a society. However, ignoring ethics, by analyses that spread modified, erroneous or false data can produce very serious consequences, of a magnitude that is as yet difficult to assess fully.

What Does Big Data Analysis Actually Mean?

Big Data analytics "refers to collecting, processing, cleaning, and analyzing large datasets to help organizations operationalize their big data" (Tableau). Not without importance, however, is also the interpretation and visualization of the data, their presentation in a certain form, preferably visual, which actually represents a certain perspective. The processes involved are extremely sophisticated, with many challenges related to the heterogeneity of the data, the sources from which they come, as well as the correctness of the data or the amount of noise that often accompany the different types of data.

Some of the types of analyses are given below.

1. *Exploitation of data* to obtain new, valuable information from new connections between data. New relations, new patterns or different anomalies can be identified by sorting large data sets and creating clusters, applying inferential or causal methods. Data is very flexible, and new methods of collecting, processing or querying can lead to different results. Using views from different perspectives helps to gain a deeper

understanding and can lead to finding new patterns, trends and associations not previously discovered (on a macro and micro scale). The various predictive analyses can be used to predict the future outcomes and to identify risks or opportunities.

2. *Correlation analysis* by re-use or use of the data for a purpose other than that for which it was originally collected. Unlike other resources, data can be reused at very low costs, and can be used simultaneously by different people, institutions or companies. By correlations between (apparently) disparate data sets, extremely valuable information can be obtained for specific or contextual situations. For example, data from the population database register can be correlated with geolocation data, in order to obtain information about the number of births from a certain period, in a certain geographical area; or the number of schools can be correlated with the number of children per city/district in order to obtain the distribution and percentage of children attending school, etc.

Regarding data correlation, there are many studies on spurious correlations (Gandomi and Haider; Fan et al.; Flovik). These are real "pitfalls" in data analysis: dangerous speculations that can occur when correlations are forcibly posited between data with no actual causal or inferential relationship, or some that are in fact just incidental links (pure coincidences) devoid of any real meaning. These errors can occur unintentionally, due to the large volume and to the quality of the data, but, unfortunately, they can also be done deliberately, in order to achieve "expected" / "desired" results, which can have a strong negative impact.

3. *Obtaining a more detailed profile* (institutional or individual), with all the associated information (visible or invisible): for example, in a supermarket, the information obtained by analyzing all purchases, from all customers is much more complete and more accurate than that obtained through surveys (based on representative or random samples). Correlated with the data received from the customers' mobile phones, it can be used to analyze the mobility/frequency of store visits, and their preferences for certain products. Or, in medical care, the analysis of a patient's records, located in different departments, can offer crucial information to identify certain diseases in early stages, or trends in the evolution of certain diseases.

As we have seen above, apart from the huge volume or its heterogeneity, Big Data has an extremely important characteristic, namely *value*, which is the *type* of information and *the way it is used*. Raw data has no meaning, it does not communicate anything; it is simply collected and recorded. When data is processed and translated into information, it begins to have value through the meaning it acquires, or, more precisely, the meaning that can be attributed to it. Information can become important, interesting, depending on the context, depending on the interest of certain parties or depending on how it is put to use. The ways of extracting insight from information assets in order to obtain new information depends on the technology used, but also on the ability, the competence of some people to query these databases: "It is not the amount of data that is making the difference but our ability to analyse vast and complex data sets beyond anything we could ever do before" (Marr, *Big Data: Using Smart Big Data* 10). And this is what is called *data science*.

Recently appeared, developed and quickly transformed into a new discipline in major universities around the world, data science[iii] is closely related to the phenomenon of Big Data, the ability to analyze data and turn it into information (*Data Mining* and *Big Data Analysis*). Davenport and Patil called it "The Sexiest Job of the 21st Century," and the researchers specialized in data analysis "the people who can coax treasure out of messy, unstructured data." Regarding this ability to know where to look and what to pay attention to in searching, Hal Varian, chief economist at Google, explained in 2009: "What is it that's really scarce in the Internet economy? And the answer is *attention*. . . . *A wealth of information creates a poverty of attention*. So being able to capture someone's attention at the right time is a very valuable asset" (qtd. in McKinsey & Company). This will be one of the principles of the algorithms used by Google: by collecting data and observing user preferences, the engine offers more and more "personalized" information, going as far as selecting, sorting and summarizing information or news according to preferences and previous searches. On the other hand, the issue that might arise is that using only personalized information, reading only selected news, our perception of reality is gravely altered, limited only to what we want, what we like or what interests us directly. This is just one step away from constructing an extremely beautiful and "personalized," but

subjective reality. The space we build with these "selected data" seems extremely "real," and so do the links, relationships, information, all of these forming a "socialized cocoon" which, however, is merely an ideal we aspire to, and can easily be different from the reality we actually live in.

The impact of Big Data on the market is significant. Started as a "technological phenomenon," Big Data is currently a "technico-social phenomenon," with implications in all fields. Similar to the fields of economics, marketing, management, among the beneficiaries of Big Data analysis we also find the medical field, social sciences or advanced research (Günther et al.). These are fields which have not benefited, until now, from the techniques and tools necessary to process data in such large quantities, in real time. The linguistic field, also, fully exploits these technologies, with the help of new models of natural language processing (NLP) such as GPT2/3, BERT, etc. (Markowitz). By interconnecting and correlating the data, the linguistic context analysis is far extended and can be linked with the cultural, dynamic and complex functional context of the person in society. These techniques are essential in training AI and generative AI algorithms that can thus really improve their performance so as to be similar to human activity ("human-like").

Algorithms and AI

How people interact with, or relate to, the algorithms behind this data is a growing area of research. AI-powered chatbots exchange information with users, mostly, but not exclusively, through conversations. Behind the chatbots are actually algorithms trained on a corpus of data. These corpora and how the algorithms work (a "black box") are not transparent, and this raises the question of the credibility of the responses. There has been a real increase in the percentage of correct answers over the last year (2022), but there are still situations where answers are incorrect, inaccurate or false positives.

Studies of interactions with algorithms show that people relate to them in very different ways, depending on the type of application (social media, entertainment, banking or marketing platforms, etc.) (Bucher; Siles; Swart). In addition to the need to educate users about their existence, as well as when

and how they work, it is also important to consider how they are "enacted and come together to make different versions of reality" (Bucher 152).

Even though they are invisible, people have emotional reactions to them because their expectations are often different (or contradictory) to the way algorithms work, or, more interestingly, they materialise them according to different contexts, not as "coded instructions telling the machine what to do, but rather as emergent accomplishments or socio-material practices" (Bucher 152). Young people's awareness of how they interact with algorithms when using social networks, and how this raises certain cognitive and affective issues, is also discussed by Swart, who concludes: "first, the lack of an established baseline about how algorithms operate; second, the opacity of algorithms within everyday media use; and third, limitations in technological vocabularies that hinder young people in articulating their algorithmic encounters."

In response, the investments made by technology companies show a real interest in integrating AI tools, virtual reality, or augmented virtual reality into their applications/platforms (Davenport and Mittal). The results obtained over the last years have exceeded researchers' expectations, especially in terms of efficiency and augmenting the workload. But, with the ability to process all the data stored and, in particular, to make decisions in modelling these data and in creating (new) representations that can become "almost tangible," these technologies are causing concern among some scientists (Dennett; Hunt; Delcker). In their view, this technology risks to become dangerous, even "toxic," for humanity.

Some Challenges

The huge potential of processing this data does not come without challenges. There are many studies on the benefits, but also on the (potential) costs of Big Data analysis and interpretation, starting with the fact that it is not (always) transparent how the institutions/companies that have access to this data use it and, of course, how they interpret it and then use these interpretations.

A brief review of recent studies related to Big Data and its applications and implications reveals some extremely interesting elements

in terms of the scale of the phenomenon, and highlights concerns that mainly relate to the danger of the illusion of a reality provided by algorithms that humans can no longer control. Thus, algorithms will be able to make decisions in situations that the human mind has no way of predicting, because it does not have the cognitive resources/means to anticipate them.

One of these concerns is related to the use of various large-scale language models that use neural networks and unsupervised learning algorithms. Opinions are divided among researchers, with some calling for caution and discernment in the use of LLMs (Large Language Models or generative pre-trained transformers, GPT) in any context (Bender and Koller; Bender et al.; Marcus and Davis, etc.). Their concern is based on the assumption that a computer will never be able to understand the human mind or reproduce the complexity of language. The illusion of perfect communication with the computer is only a test to the human tolerance for uncertainty and our ability to understand each other.

Other scholars are optimistic about the relevance of these models in understanding language complexity: "Transformers and related architectures seem able to learn both inferential and referential semantic" (Søgaard 443). Describing them as "game changers on all fronts," Steven Piantadosi even appreciates that these models:

> have opened the space of plausible linguistic theories, allowing us to test principles beyond the ones that have traditionally concerned linguists. They allow us to finally develop compelling theories of the interplay of structure and statistics. And they appear to solve many problems that generative syntacticians have worried about, but without using any of their theoretical tools and constructs. Large language models rewrite the philosophy of approaches to language. (29-30)

"You are not a Parrot and a Chatbot is not a human. And a linguist named Emily M. Bender is very worried what will happen when we forget this" is the title of an article by Elizabeth Weil, in which she discusses Bender and Koller's article on the artificial large language models used by chatbots (ChatGPT), to converse in a "natural language," and give more than satisfactory answers to human users. The article explains the methodology of creating these models by extracting patterns/types of linguistic structures from huge corpora of text, which are then used to predict (guess) the right

word (or the right sentence). The good results obtained give the impression of "real or correct," although these models have no referents in the "real world" and they are not produced from anyone's experience. The meaning of the words is obtained strictly by inferential or predictive algorithmic/statistical calculation. "*Please do not conflate word form and meaning. Mind your own credulity"* is Bender's message (quoted by Weil). Bender does not share the assumption that chatbots using artificial language models will help us to better understand linguistic structures and even to communicate better in society. Quite the contrary, she sees a great danger in attempts to dilute the difference/distinction between humans and bots.

A critical attitude towards artificial language models can also be found in Gebru, Wickens and Janus, Ackerson, etc. The discussions focus, in fact, on the referentiality of language, on the way in which words can have meaning and significance. They also address those strange situations called "hallucinations," when language models produce results that are not related to "reality": "instances where a large language model creates information that is not based on factual evidence but may be influenced by its transformer architecture's bias or erroneous decoding" (Ackerson).

This issue of "de-personalization," or "de-individualization" is also current in the field of machine translation, where professional translators talk about how the "perspective of the translator" risks getting lost by using translation memories or machine translation. But questions of originality and copyright exist in all creative fields (music, drawing, painting, etc.). How much can AI tools help us to put people first? Ploin et al. consider that "[t]he creative AI world is about doing cool things with AI tools" (11), reflecting some artists' opinions that AI should and will remain just an aid, and is not a serious threat. In the field of music, generative AI that helps/supports the composer in his or her compositional activity has not yet been credited by authors, who generally argue that music composition should remain strictly human and non-repetitive: *"*AI is just a more complicated extension of digital tools already in use in most music practices today." In translation, on the other hand, professionals want a tool to help them, not a human replacement; translation management systems have proven their effectiveness and are being increasingly used. So the discussion should be about the type and level of help that such tools can

provide, and about copyright when AI is involved in the translation process, because translation risks to be no longer "une activité intellectuelle, nécessaire, bénéfique et profondément humaine" ("a necessary, beneficial and deeply human intellectual activity") (ATLF and ATLAS). This manifest document initiated by these two French associations is set in the context of a massive automation of some occupations, including translation. According to a very recent study (March 2023) on the impact of the LLM on the US labour market, interpreters and translators make up 76.5% of the "occupations with the highest level of exposure" (Eloundou et al. 15).

It is impossible to ignore the enormous advantages of increasing the quality of applications. On the other hand, it cannot be a wise direction for human evolution to empower computers in areas where humans should be making (important) decisions. Slowly, discreetly, humans will be replaced by programs: the two will become "interchangeable" (much like the characters in Eugène Ionesco's play "The Bald Soprano"). It is true that elimination of human error is one of the many benefits of the use of AI. Unfortunately, with errors, humans are also being eliminated from more and more activities where human decisions are (still) important. And the current trend is to move from a human "intuitive" method to an artificial, purely computational one, by transferring both the creative part and the decision-making part (so-called risk assessment tools) to a computer system.

In addition to challenges about the ethics of the use of data, as well as the accuracy and confidentiality of the sources, there are also discussions about the algorithms behind data processing, the "biases" that may be present in their code and the risk of using these models *in any context* (Buolamwini and Gebru; Warner; Denton et al.; Dustin).

The difficulty or even impossibility to study what exactly is included in the training data, and what types of sources are used, are also important issues in Big Data processing (Gebru et al.; Bender et al.). The way in which the data is filtered or selected in the "cleaning" process has direct impact on how the algorithms are trained. Trained AI corpora consist of all data on the internet, mostly open source, common and unverified data. There is also a lot of "negative" information which promotes radical views, related to racism, xenophobia, etc. In this case, the companies, the producers of

technology, make a serious effort to remove it. These time-consuming activities, some automated, use lists of "banned" words. Every day, human employees (from different countries) mark/tag whole sentences for AI algorithms in order for it to learn from the most correct patterns. Cleaning and filtering are not without risk. By removing certain words or phrases, information about certain subjects can simply disappear, turning filtering into censorship. On the other hand, continuous surveillance would be another problem: "Big Data pitfalls may lurk, for example, in the emergence of continuous surveillance facilitated by advances in real-time massive data capture, storage, and analysis" (Diebold 37).

Instead of Conclusions

We now live in a "digital world," a world made up of enormous amounts of data received from us and from the sensors in objects around us. These enormous amounts are processed by AI and provided to us as "a reality." As we increase our interaction with AI (or introduce AI into all our actions), our perception changes discreetly and subtly, becoming more restricted to a "vision," a "form" that AI presents to us and which we can no longer verify or control because, as mentioned above, we do not have the cognitive capacity to process such large amounts of data, nor do we have the speed required to do so over time. Furthermore, the influence of typical social media interaction, the need to "share," to obtain immediate reactions has become "an addiction," and we risk falling into what is called "bounded rationality," i.e. a *thinking oriented mainly towards immediate satisfaction*, to the detriment of a thinking oriented to optimization (Simon), which requires more depth, critical analysis and time (which, indeed, we may no longer have!).

Although statical, data is very flexible, and if we change it, the perspectives change, the queries are different, and we get totally different representations, different shapes. This cycle of actions, of interactions between data-algorithms-results can lead to an extension of human thinking, multiple perspectives being useful for a deeper understanding of phenomena, problems, etc. It helps us perceive the multitude of possibilities of representation, i.e. the complexity of a possible reality, but of a "reality"

created from data, a quantifiable "reality." Our knowledge of reality is largely based on models which approximate reality, but in this case the tendency is to *create large models and force them to fit everywhere*:

> [W]e can share a common enough view [of reality] for most of our working purposes, so that reality does appear to be objective and stable. But the chances of achieving such a shared view become poorer when we try to encompass broader purposes, and to involve more people. (Kent 203)

We assume, in good faith, that the data AI processes is correct and the algorithms are "objective," at least to a point. In "reality," however, data collected manually or automatically has many limitations, if not errors, both objective (technical) and subjective (caused by human intervention). The quality of the data depends on the type of sensor and how it is collected, and much of it is "noise." Calculations are made with this data, so the results are susceptible to errors or "false errors," which we have no way to identify. As mentioned before, data accuracy and security are major challenges in Big Data analysis. On the other hand, actions to maintain data accuracy, to "clean" the data, are not without errors. So even though data are ubiquitous, their value is inconsistent.

If we constantly interact with AI and, more importantly, rely only on it in making certain decisions (i.e. *an indiscriminate embrace of AI*), there is a risk to have a distorted (or erroneous) perception of things, or more precisely, we will have the *AI view*, without even realising it. False positive results can be obtained very easily.

A second major issue is the algorithm behind AI, how it is trained and with what data. Algorithms can be manipulated to produce "convenient," "contextual" results, creating the impression that they are "real" because "all the data has been analysed." And this perception is justified because the data is collected from us, from our habits, from our beliefs, from our actions and ideas. The error therefore comes from the assumption that AI is, or can only be objective. To train algorithms, enormous amounts of various data can be used, but also "filtered, selected" corpora or even "synthetic" data (i.e. artificially manufactured, see generative AI). There are a lot of reasons for using synthetic data, including machine learning applications, data augmentations or data privacy (Islam).

But although it may be mathematically or statistically important, mixing real and fictional data to train algorithms tends to create more confusion and mix the "real" with the "digital."

In addition, we should not forget that AI only works on quantified data. So AI does not really get all the data, because some of it cannot really be quantified. Reality is a continuous, infinite, ever-changing form. It is the unquantifiable aspects of reality that are important, and this is indeed the great challenge: "*we are not modelling reality, but the way information about reality is processed, by people*" (Kent 46, emphasis mine).

Despite the huge and realistic benefits of artificial intelligence, a lot of intervention is needed to mitigate the risks, and governments, businesses and industry have agreed and started to work on a set of regulations for the responsible use of data.

We are in the midst of a paradigm shift, both in knowledge and in our relationship with technologies. Moreover, we could say that we are experiencing a new challenge in human history. In this "digital decade" (see European Commission Strategy), we are building more and more sophisticated and refined intelligent software and objects, hoping that they will help us to augment our knowledge, to better understand the reality around us. All this is true to some extent, because technologies have expanded and improved our knowledge, helped us discover things and facts we had never imagined before. But, with the potential it reaches today, with the access of Big Data, with the algorithms that monitor our action, AI technology is starting to become another "cognitive system," learning from us, to which we transfer, in fact, our habits, our culture, our values. It is a first phase in the *digitization of humanity*. We will soon reach a point from which we evolve together; as is to be expected, after a while, AI will be equally or even better prepared than humans, because it will not be hindered by the physical and physiological limitations of a human body. Or maybe it will help us get out of our thinking routines and preconceptions, only to discover that "real" is "unreal," and the world is actually a set of data – inconsistent, incoherent, irrational... What then?

Notes:

[i] The Internet of Things (IoT) is the network of physical objects that contain embedded technology to communicate and sense or interact with their internal states or the external environment (Gartner Glossary).
[ii] In recent years, especially after the Covid-19 pandemic, the interaction between people has been profoundly modified and has become much more digitized. As a result, the way we perceive and relate to the digital form, as well as the way we "signify/understand" it, is subtly changing towards blurring the difference between human and digital processing.
[iii] In a more detailed definition, "data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and machine learning with specific subject matter expertise to uncover actionable insights hidden in an organization's data. These insights can be used to guide decision making and strategic planning" (https://www.ibm.com/topics/data-science).

Works Cited

Ackerson, Noble. "GPT Is an Unreliable Information Store." *Towards Data Science*, Feb. 2021, towardsdatascience.com/chatgpt-insists-i-am-dead-and-the-problem-with-language-models-db5a36c22f11.

ATLF et ATLAS. *IA et traduction littéraire: les traductrices et traducteurs exigent la transparence*,www.atlas-citl.org/wp-content/uploads/2023/03/Tribune-ATLAS-ATLF-3.pdf. Accessed 14 July 2023.

Auditore, Peter. "Customer-Centricity and The Kings of Big Data – What They Collect About You." *Social Media Today*, 13 June 2011, www.socialmediatoday.com/content/customer-centricity-and-kings-big-data-what-they-collect-about-you.

Bender, Emily M., and Alexander Koller. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 5185–5198, aclanthology.org/2020.acl-main.463/.

Bender, Emily M., et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, 2021, pp. 610–623, https://doi.org/10.1145/3442188.3445922.

Bucher, Taina. *If . . . Then: Algorithmic Power and Politics.* E-Book ed., Oxford UP, 2018.

CIO Bulletin. "How Much Data is Created Every Day and How to Collect It." *CIO Bulletin*, 22 Apr. 2022, www.ciobulletin.com/big-data/how-much-data-is-created-every-day-and-how-to-collect-it.

Buolamwini, Joy, and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR,* vol. 81, 2018, pp. 77-91, proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

Davenport, H. Thomas, and Nitin Mittal. *All-in On AI: How Smart Companies Win Big with Artificial Intelligence.* Harvard Business Review Press, 2023.

Davenport, H. Thomas, and DJ. Patil. "Data Scientist: The Sexiest Job of the 21st Century. Meet the People Who Can Coax Treasure out of Messy, Unstructured Data." *Harvard Business Review*, Oct. 2012, hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century.

Delcker, Janosch. "14 Ways AI Could Become a Detriment to Society." *Forbes*, 14 June 2021, www.forbes.com/sites/forbestechcouncil/2021/06/14/14-ways-ai-could-become-a-detriment-to-society/.

Dennett, C. Daniel. "The Problem with Counterfeit People." *The Atlantic*, 16 May 2023, www.theatlantic.com/technology/archive/2023/05/problem-counterfeit-people/674075/.

Denton, Emily, et al. "Detecting Bias with Generative Counterfactual Face Attribute Augmentation." *ResearchGate*, June 2019, www.researchgate.net/publication/333842250_Detecting_Bias_with_Generative_Counterfactual_Face_Attribute_Augmentation.

Diebold, Francis X. "What's the Big Idea? 'Big Data' and its Origins." *Significance*, vol. 18, no. 1, 2021, pp. 36-37, rss.onlinelibrary.wiley.com/doi/full/10.1111/1740-9713.01490#pane-pcw-references.

Drucker, Peter. *The Age of Discontinuity: Guidelines to Our Changing Society.* Routledge, 1992.

Duhigg, Charles. "How Companies Learn Your Secrets." *The New York Times Magazine*, 16 Feb. 2012, www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

Dustin, Jeffrey. "Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women." *Reuters*, 11 Oct. 2018, www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

Eloundou, Tyna, et al. "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models," 27 Mar. 2023, arxiv.org/pdf/2303.10130.pdf.

Fan, Jianqing, et al. "Challenges of Big Data Analysis." *National Science Review*, vol. 1, no. 2, 2014, pp. 293–314, https://doi.org/10.1093/nsr/nwt032.

Flovik, Vergard. "The Hidden Risk of AI and Big Data." *KD nuggets*, www.kdnuggets.com/2019/09/risk-ai-big-data.html.

Gandomi, Amir, and Murtaza Haider. "Beyond the Hype: Big Data Concepts, Methods, and Analytics." *International Journal of Information Management,* vol. 35, no. 2, Apr. 2015, pp. 137-144, https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

Gartner. "Gartner Glossary." www.gartner.com/en/glossary/all-terms. Accessed 24 May 2023.

Gebru, Timnit. “Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning.” *KDD'20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, https://doi.org/10.1145/3394486.3409559.

Gebru, Timnit, et al. “Datasheets for Datasets.” *Communications of the ACM*, vol. 64, no. 12, 2021, pp. 86-92, arxiv.org/pdf/1803.09010.pdf.

Günther, Wendy, et al. “Debating Big Data: A Literature Review on Realizing Value from Big Data.” *The Journal of Strategic Information Systems*, vol. 26, no. 3, 2017, pp. 191-209, https://doi.org/10.1016/j.jsis.2017.07.003.

Hadi, Hiba Jasim, et al. “Big Data and Five V’s Characteristics.” *International Journal of Advances in Electronics and Computer Science*, vol. 2, no. 1, Jan. 2015, pp. 16-23, iraj.doionline.org/dx/IJAECS-IRAJ-DOIONLINE-1635.

Hunt, Tamlyn. “Here’s Why AI May Be Extremely Dangerous – Whether It’s Conscious or Not.” *Scientific American*, 25 May 2023, www.scientificamerican.com/article/heres-why-ai-may-be-extremely-dangerous-whether-its-conscious-or-not/.

Islam, Ray. “Unveiling the Potential of CTGAN: Harnessing Generative AI for Synthetic Data.” *KD nuggets*, 20 April 2023, www.kdnuggets.com/2023/04/unveiling-potential-ctgan-harnessing-generative-ai-synthetic-data.html.

ITU. “Mobile Phone Ownership,” www.itu.int/itu-d/reports/statistics/2022/11/24/ff22-mobile-phone-ownership/. Accessed 24 July 2023.

Kent, Wiliam. *Data and Reality: A Timeless Perspective on Perceiving and Managing Information in Our Imprecise World.* 3rd ed., updated by Steve Hoberman, Technics, 2012.

Kitchin, Rob, and Garvin McArdle. “What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets.” *Big Data and Society*, vol. 3, no. 1, 2016, https://doi.org/10.1177/2053951716631130.

Lohr, Steve. “The Origins of ‘Big Data': An Etymological Detective Story.” *The New York Times*, 1 Feb. 2013, archive.nytimes.com/bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/.

Loshin, David, and Abie Reifer. *Using Information to Develop a Culture of Customer Centricity.* Elsevier, 2013.

Marcus, Gary, and Ernest Davis. “GPT-3, Bloviator: Open AI’s Language Generator Has No Idea What It’s Talking About.” *MIT Technology Review*, 22 Aug. 2020, www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/.

Markowitz, Dale. “Transformers, Explained: Understand the Model Behind GPT-3, BERT, and T5.” *Dale on AI*, daleonai.com/transformers-explained. Accessed 24 May 2023.

Marr, Bernard. *Big Data: Case Study Collection*. E-book, Wiley, 2015, bernardmarr.com/img/bigdata-case-studybook_final.pdf.

---. *Big Data: Using Smart Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance.* E-book ed., John Wiley and Sons, 2015, bernardmarr.com/wp-content/uploads/2022/05/Big-Data-1.pdf.

McKinsey and Company. "Hal Varian on How the Web Challenges Managers." 1 Jan. 2009, www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/hal-varian-on-how-the-web-challenges-managers.

Özköse, Hakan, Ari, Emin Sertaç and Cevriye, Gencer. "Yesterday, Today and Tomorrow of Big Data," *Procedia – Social and Behavioral Sciences*, vol. 195, 3 July 2015, pp. 1042 – 1050, https://doi.org/10.1016/j.sbspro.2015.06.147.

Piantadosi, Steven. "Modern Language Models Refute Chomsky's Approach to Language." Mar. 2023, lingbuzz.net/lingbuzz/007180.

Ploin, Anne, et al. "AI and the Arts: How Machine Learning is Changing Artistic Work." Report from the Creative Algorithmic Intelligence Research Project, Oxford Internet Institute, University of Oxford, 2022.

Rockwell, Geoffrey, and Stéfan Sinclair. "False Positives: Opportunities and Dangers in Big-Data Text Analysis." *Hermeneutica: Computer-assisted Interpretation in the Humanities*, MIT Press Scholarship Online, pp. 113-136, 2016.

Schiuma, Giovanni, and Daniela Carlucci. *Big Data in the Arts and Humanities: Theory and Practice.* E-book ed., CRC Press, Taylor and Francis Group, 2018.

Siles, Ignacio. *Living with Algorithms*. MIT Press, 2023.

Simon, A. Herbert. "Rational Choice and the Structure of the Environment." *Psychological Review*, vol. 63, no. 2, pp. 129-138, 1956, Google Scholar.

Søgaard, Anders. "Understanding Models Understanding Language." *Synthese*, vol. 200, no. 443, 2022, https://doi.org/10.1007/s11229-022-03931-4.

Statista.com. "Volume of Data/Information Created, Captured, Copied, and Consumed Worldwide from 2010 to 2020, with Forecasts from 2021 to 2025." *Statista*, www.statista.com/statistics/871513/worldwide-data-created/. Accessed 24 July 2023.

---. "Number of Data Centers Worldwide in 2022, by Country." *Statista*, www.statista.com/statistics/1228433/data-centers-worldwide-by-country. Accessed 24 July 2023.

Swart, Joëlle. "Experiencing Algorithms: How Young People Understand, Feel About, and Engage with Algorithmic News Selection on Social Media." *Social Media + Society*, vol. 7, no. 2, 2021, https://doi.org/10.1177/20563051211008828.

Tableau.com. "Big Data Analytics: What It Is, How It Works, Benefits, And Challenges." *Tableau*, www.tableau.com/learn/articles/big-data-analytics. Accessed 24 July 2023.

Taylor-Sakyi, Kevin. "Big Data: Understanding Big Data." 2016, Cornell University, https://doi.org/10.48550/arXiv.1601.04602.

Warner, Andrew. "The False Promise of Generative AI Detectors." *Multilingual*, 13 July 2023, multilingual.com/the-false-promise-of-generative-ai-detectors.

Weil, Elizabeth. "You Are Not a Parrot and a Chatbot Is Not a Human. And a Linguist Named Emily M. Bender Is Very Worried What Will Happen when We Forget This." *Intelligencer*, 1 Mar. 2023, nymag.com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender.html.

Wickens, Eoin, and Marta Janus. "The Dark Side of Large Language Models." *HiddenLayer*, 23 Mar. 2023, hiddenlayer.com/research/the-dark-side-of-large-language-models-2/.

Wise, Jason. "How Much Data Is Generated Every Day in 2023? (NEW Stats)." *EarthWeb.com*, 7 Apr. 2023, earthweb.com/how-much-data-is-created-every-day/.

Zikopoulos, Paul C., et al. *Understanding Big Data. Analytics for Entreprise Class, Hadoop and Streaming Data*. E-book ed., McGraw-Hill, 2012.