The Lab's Quarterly
2025/ a. XXVII / n. 3 – ISSN 2035-5548

FROM DIGITAL TRACES TO ARTIFICIAL INTELLIGENCE
New boundaries from representativeness

by *Beba Molinari*[*]

Abstract

The abstract aims to highlight what contribution Artificial Intelligence can make in the field of social research by setting as a fixed point the tools to date established in research methodology, whether traditional or closely related to e-methods. It is necessary to stop and think about how and what data AI provides us with by asking: are we in the same scope of analysis as e-methods? Instead, can we continue to handle such data through traditional analysis techniques, or should we think of AI as totally new data/information? These are just a few questions that will be attempted to be answered without any claim to exhaustiveness of course, but aimed at discussing representativeness and margin of error not only statistically, but understood in a much broader sense.

Keywords

Artificial Intelligence, Big Data, E-methods, Machine learning

* BEBA MOLINARI is a researcher of Sociology at the University of Rome Tor Vergata. Email: beba.molinari@uniroma2.it

DOI: 10.13131/unipi/x6c5-2t23

1. INTRODUCTION

The ever-increasing use of remote work that occurred in 2020 with the spread of the SARS-CoV-2 virus, which was followed by an increasingly high-performance response through ICT solutions with a consequent acceleration of digital that ensued in the following years, has highlighted how much the digital dimension is an integral and pervasive part of our lives through the solving of real problems with an increasing number of practical applications, most of them based on Artificial Intelligence (AI) (Thottoli, 2024).

AI offers a range of thoughts and opportunities unexplored to date. While we may think of it as a very useful tool that will touch much of our lives, consciously and unconsciously, we ignore the procedures and ways in which AI moves with great adaptive capabilities on our devices. We will have the opportunity in the next 7 years to have a considerable amount of information coming from AI, even more than previously assumed, as the European Programming 2021-2027, particularly the Euromed and Alcotrà programs, focus much more than in previous plans on Technological Innovation with an established interest in the inclusion of AI and the direct consequence of a series of increasingly large experiments both nationally and internationally.

It has already been about 15 years since a strong debate was sparked in Italy about the use of new 2.0 research tools in the social sciences (Corposanto and Molinari, 2022).It is not possible to talk about e-methods without discussing Big Data, inevitably the discourse marries on the "construction" of the data, then on the plausibility of the survey instrument adopted thought in relation to the object of study (Marradi, 1989). It is inevitable to understand how quickly we move from an extremely innovative vision to a discourse rooted in "traditional" research methodology.

This is not the first occasion in which the author questions around the plausibility of new online search tools that allow us to analyze large amounts of information collected within databases. The focus of the discussion revolves around the concept of representativeness, the reliability of the data (Molinari and Corposanto, 2023) with a particular focus on e-methods and how they can be subject to bias, not forgetting that most of the classical tools in the researcher's toolbox are not exempt from such risks. It is important, therefore, to ask what is meant by bias, a word that varies greatly in its meaning depending on the survey instrument under consideration. It is certainly not the first time that attention has shifted in the methodological field to the fidelity of the data

in its dimensions related to: the sincerity of the response, the congruence of meaning, the a priori classification of the responses and their respective exhaustiveness (Bethlehem and Biffignandi, 2012). This is not the place to discuss such topics, about which so much has already been written, but it is necessary to make a clarification that sets the stage for this article. While we have a number of assumptions on which the foundations of social research methodology are based, it is necessary to ask how many of these assumptions can still be valid with an immense amount of information and data provided and analyzed by the network, shifting the focus to data used by artificial intelligence (AI).

It is necessary to stop and think about how and what data AI provides us with, so some questions are necessary: are we in the same scope of analysis as e-methods? Can we continue to handle such data through traditional analysis techniques, or should we think of AI as totally new data/information? Does the data that is "released" to us by AI totally fall under the well-established definition of Big Data? Or are they "other" data because they pose us a number of more purely technical questions (format, size, etc.)?

Before going into the specifics and new methodological frontiers that AI poses to us, let us make a small clarification. We often discuss data quality, but in Web 2.0 contexts, let alone Web 4.0, does the locution "data" still make sense? Or does the connotation change depending on the context, the type of information and the provenance, i.e., the data warehouse[1] from which it was extracted?

In this case, according to the author, given that the level of information changes depending on the type of analysis to be carried out (context analysis, sentiment analysis) and even more so depending on the Application Programming Interfaces (API) used to carry out the analyses, the quality of the data is constantly changing. In light of these considerations, it would be more appropriate to think about the potential of data warehouses and the possible mash-ups of data, that is, the many combinations that can be achieved with data of different nature and origin[2] (Corposanto and Molinari, 2022).

[1] The data warehouse is a kind of second level of a 'database' within which is contained a series of information oriented towards individual subjects, be they users and/or consumers, integrated with several databases from which it 'imports' specific information previously identified and thus of interest to the programmer/researcher.

[2] By way of example alone, it is now possible to use an API to extract data from YouTube, to make topographical maps of real or presumed risks, to build an ego network from facebook pages, to carry out a contest analysis via Twitter, etc.

2. OLD TOOLS, NEW RESOURCES

Before going into what are the survey tools of today's researcher, it is necessary to understand what was the path that led us to an increasingly diverse range of data born with the web.[3]

We have always imagined the moment of data collection as at that stage when the researcher with his or her toolbox goes into the field to begin the administration of the research instruments (Palumbo and Garbarino, 2006). Therefore, the first question we might ask is whether with the tools made available by Web 2.0 the survey phase changes: it is clear that the first difference is clearly visible, the researcher no longer moves in the field, but rather navigates the web and through it moves the everyday real-world to Web 2.0 contexts from blogs, to social networks, to the use of search engines, etc.. It is important to point out, however, that the intromission of the Web does not become apparent exclusively in the phase of navigation on the part of the user, but emerges even earlier, for the researcher in particular it is already configured in the preliminary phase of study, in which the boundaries of the object of research, the declination of hypotheses and the understanding of the population under study are delineated. The use that can be made of it then is manifold, from bibliographic collection, to statistical data, to the identification of the most suitable indicators to understand the phenomenon (Grimaldi, 2005).

It is not our goal here to discuss the more purely epistemological aspect of detection tools in Web 2.0 contexts, but rather to understand the possible developments of such tools by turning our gaze toward AI, taking a first step toward a bumpy and complex path that will seek to understand, whether e-methods can be considered a continuation, or a new Web 2.0 toolbox, and especially whether AI falls within the Web 2.0 toolbox or is something totally different.

Before going into the specifics of the use of AI in social research, it is necessary to set some milestones.

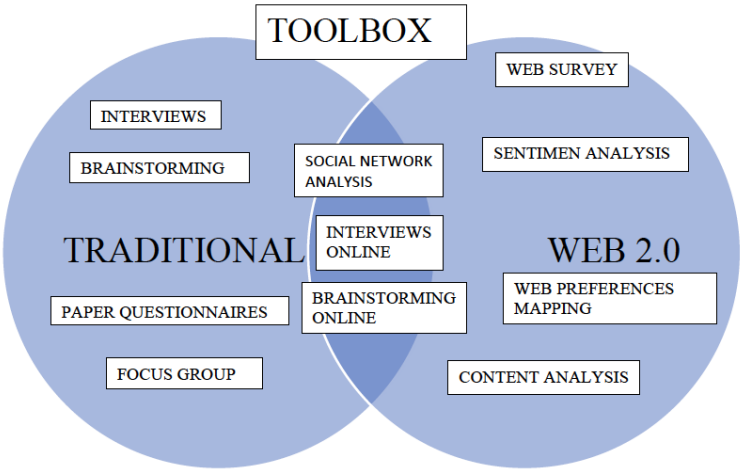
The first major difference, which determines the choice of in-depth analysis adopted in the articulation of the next paragraphs, is the inescapable link between tools and data analysis, an aspect that also arises for the traditional toolbox, but not as strongly constraining as it is for web techniques.

It is the author's (Molinari and Corposanto, 2022) belief that there are

[3] First and foremost, a substantive premise is necessary, namely that the tools determine the researcher's choice to conduct a particular analysis rather than another (Bucchi, 2019). The very nature of the data, whether etic or emic, affects the subsequent choice regarding the type of analysis to be adopted (Corposanto and Molinari, 2017).

as many different toolboxes as there are interpretive lenses that the researcher intends to use. To decide that a tool is within one cassette rather than another would be to claim to contain within a categorization an ever-changing reality. Having made that clarification let us try to define to date a first reorganization of research tools by distinguishing between the traditional and the web 2.0 cassette.

Figure 1. The researcher's toolbox in modern times



Thanks to the use of a Venn we have separated the two toolboxes into two sets, between the two is clearly visible a common area within which are represented the web tools that we consider to be the direct continuation of the traditional tools, on the right instead are enclosed the tools that in our opinion can be considered innovative compared to their predecessors, thus bringing that "something different" compared to the traditional tools.

The choice to include one tool rather than another within an area was dictated by considerations made in different application contexts, with different objects of study.[4] In this regard, the following aspects were considered:

[4] The following has been extensively discussed in an article, here is the bibliographical reference: ---, Analizzare dati di microblogging con la Sentiment Analysis. Quale rappresentatività?.Sociologia Italiana. 11: 123-132.

- survey procedure,
- ype of data to be analyzed,
- quality of the data as a whole,
- type of analysis that allows us to carry out the data obtained.

The first difficulty was that of being able to distinguish between the type of instrument and its analysis.It thus seems clear from the definition of the areas common to the two sets that interviews, brainstorming and focus groups can also be declined in web contexts with the proper adjustments, they are something different from the classical tools, but at the same time they can be analyzed with traditional analysis techniques. However, there are a number of entirely different tools such as data mining, search engine preference mapping, web surveys themselves, sentiment analysis, and content analysis that use totally different data where the application of so-called traditional analysis would certainly be a stretch for some, while for others virtually impossible.
The striking example of the latter case is the mapping of websites, where there is no other way of knowing the preferences of Internet users than through network analysis, the data that emerged can in fact be used only in this sense, an analysis that falls squarely within the scope of web crawlers' programs.[5]

In this context then, how does AI position itself? Let's get into it by trying to shed light on the uses of AI in social research contexts and the respective analyses and, of course, the algorithms on which AI itself is based.

3. WHAT ARTIFICIAL INTELLIGENCE IS

In addition to all these aspects related to the more classical conception from search tools to the arrival of methods related to online contexts, we must say that we live in a context of algorithms that we find in our everyday life without even realizing it. The algorithm conceived in the Big Data context was born to solve complex problems that result in our everyday life, in finding the shortest way to get to the nearest restaurant that at the same time satisfies our food preferences, what are the people you might know and add them on social networks, but they were thought mostly out of a pressing need to adequately manage amount of information seen before.

[5] In this regard, it is also worth mentioning the semantic web, which allows us to monitor the content and semantic context of a web page by means of tags (Trobia, 2014:67-104).

In a note dated April 8, 2019, the European Parliament drafted a document to provide clarity with respect to what is the prerogative of the AI. The Commission Communication on AI clarifies certain aspects of AI as a scientific discipline and as a technology, with the aim to avoid misunderstanding, to achieve a shared common knowledge of AI that can be fruitfully used also by non - AI experts, and to provide useful details that can be used in the discussion on both the AI ethics guidelines and the AI policies recommendations.

In other words, AI is the ability of a machine to exhibit human capabilities such as reasoning, learning, planning and creativity. AI enables systems to understand its environment, relate to what it perceives, and solve problems and act toward a specific goal. The computer receives data (either already prepared or collected through sensors, such as a camera), processes it and responds. AI systems are able to adapt their behavior by analyzing the effects of previous actions and working autonomously.

In more detail (European Parliament, 2021) AI is divided into two major macro areas, software and embedded intelligence. The former include virtual assistants, image analysis software, search engines, facial and voice recognition systems, while the latter include robots, autonomous vehicles, drones, and the Internet of Things. If the former are what we most commonly use in everyday life, the latter are what we at first glance refer to as AI.

From a recent study conducted by the author there is a lot of confusion regarding AI from which there is clear evidence of not only the now widely debated aspect referring to the generation gap of Digital Devide, but across the board a rather marked confusion referring to AI itself. In more detail, out of a sample of 7,900 people only 14 percent chose the correct answer that involves both macro areas of AI, software and robotics, while for 37 percent of the participants AI is related only to robotics, 23 percent think of AI as software, and the remaining 26 percent have no idea what it is.

This is not the place to go into the details of the above study, but it certainly allows us to confirm, what is already partly said in everyday life, namely that the topic is rather thorny and complex. If most people do not have a correct idea of what AI is certainly it is surrounded by it in everyday life and provides useful data for processing a range of information that changes our lifestyles: dietary, relational, educational, etc. We are discussing a volume of information that is difficult to even imagine, from a total of 33 zettabytes in 2018 it is assumed to increase to 175 zettabytes at the end of 2025 where a zettabyte is equivalent to a

trillion gigabytes, with this amount of data it is difficult to think of a repository that can hold it all.

Let's get into how AI actually works, albeit in broad strokes. What we usually see and relate to when dealing as users with AI is the chatbot (Xu, 2019). The latter is a software that simulates and processes human conversations, whether written or spoken, in order to make users interact with digital devices having the impression of communicating with a real person. Chatbots are of different types depending on how you program them, there are some extremely simple ones that answer a single question, while others are much more complex such as digital assistants that learn from their own mistakes and the data the web provides them with in order to provide increasingly personalized levels that reflect the user's interests.

So our first approach with AI usually is through a chatbot with devices integrated to smartphones, PCs, or home automation devices. As anticipated usually one formulates a question, either vocal or written, on a device and waits for the answer, we could say more or less correct. What are the broad steps of the mechanism that triggers the response and how this may, or may not, be related to big data and social research methodology?

The chatbot before it is able to respond must be trained, and it is at this stage that AI, or machine learning, comes into play (Liu, Wang, Whang, 2012) through which the model underlying the chatbot learns through a huge amount of data of different kinds (text, images, videos) that are provided to it for free from the web. The learning phases are divided into three successive steps: a first one in which language skills and general notions are learned, and by learning from the mistakes made, the chatbot becomes autonomous in its ability to give correct answers, but only with respect to language proficiency. In this step the role of the programmer comes into play he "adjusts" the errors made through the randomization of correct words and a predictive process formulated as a "learning model." In the second step, the chatbot learns specialized skills through a series of multiple-choice questions against which the model formulates a task. In the third and final step, on the other hand, a team of computer scientists verifies the effectiveness of the responses and defines the "style" and genre with which the chatbot responds. Educated the software behind the chatbot actually could be used even without a network connection, it would evaluate the most reliable answer in relation to the information (big data) it had previously incamerated, but it would lack the continuous data updates that would allow it to properly "train" its skills on specific questions.

Can what we have just described fit into the now widely accepted

definition in the methodological field of Big Data? In this regard, a decade has already passed since Burrow-Savage and Kitchin (20214) defined Big Data as all those data that are distinguished from the more common data with respect to: volume breadth, high speed, variety, exhaustiveness, high resolution, relationality, flexibility (Burrow and Savage 2014, p.1; Kitchin 2014 p. 262).
It is such a broad definition that it encompasses several categories of data because it adopts an exclusionary interpretative philosophy, translated one could say that Big Data is everything that is not commonly known as 'classical' data in research contexts, in fact it is assumed that they are distinguished by 'a norm'.

Thus, the 175 zettabytes of supposed information our days - supposed because it is practically impossible to definitively number the data that travels the net - is in fact Big Data, obviously of different nature and origin. Within this mass of information, we have data derived from the traces we leave behind from shopping on the net, from the most banal online searches, from cyber security, from the ever more present personal digital assistants, from automatic translations, whether carried out within the more classic search engines such as google translator, or made and inserted within the social network, from sensors for home automation, from intelligent infrastructures, from robotically guided vehicles, etc. etc. It should be emphasised that AI and Machine Learning are not synonymous, they are two closely related, yet completely different technologies. The former has designed and will design an intelligent architecture, the latter, on the other hand, allows us to develop a learning system with ease.

4. WHAT REPRESENTATIVENESS FOR ARTIFICIAL INTELLIGENCE

The possibility of mining Big Data and being able to analyze such a vast amount of data that we can identify, a lot of information from different online sources on multiple objects of study is a wealth of information that we often forget. Especially as social researchers, we underestimate the fact that all the software that allows us this kind of analysis, from data mining to its subsequent analysis, is done through algorithms. Different algorithms that mutate at an impressive rate (Quarteroni, 2020, 2013).

This is a reason that the social researcher today must simultaneously have strong computer and mathematical skills. However, it is crucial to consider that while Machine Learning, i.e., the ability of machines to solve problems by giving them the tools to autonomously learn the

correct methodology to operate, allows us analysis of online contexts opening new frontiers of interest without any kind of error, it is equally true that we are talking about mathematical calculations only, thus lacking the reasoning that lies upstream on detection techniques and the identification of the different forms of analysis most appropriate for that particular object of research. With the use of online Big Data analytics and related algorithms, error is around the corner and is increasingly evident with the use of AI. In these AI contexts, such as Recommendation Engines that are responsible for choosing and directing the user to targeted ads and information, or Gbl virtual assistants that interact with the customer via chat, the relationships between social actors are completely lost and give way to the algorithm as a substitute form of knowledge and capable of reasoning.

Let us take a step back, in the most classic social research contexts, there is a tendency, when carrying out quantitative analyses, to regard the very concept of statistical representativeness as indispensable, which on the other hand the social sciences borrow from the experimental type of research. In itself, carrying out an experiment in the field of social phenomena is extremely complex, the 'noise' is too deafening, i.e. the environment within which the experiment is to be carried out is impossible to control, the 'intervening' variables are too many and difficult to identify, and there is also the aspect linked to the effect of the experimenter, i.e. the interference of the researcher himself in the 'choice' of experimental subjects (Rosenthal and Jacobson, 1968).[6] There is also the age-old problem of non-representativeness, i.e. in the sense that the results of an experiment are often not generalisable to the entire population, or to segments of the population other than those being studied. The basis of the experimental method is defined by the possibility of varying the independent variable and keeping all the other variables under control, on large numbers this is practically impossible, or at least, in the very few cases in which it is possible to succeed in controlling the dependent variables, we are in any case in the realm of understanding the cause-effect relationship of a phenomenon.

It should be remembered that the inferential representativeness of the sample is also lost in all those cases in which the researcher interferes in the sampling procedures and somehow forces the predetermined procedure, thus passing from a probabilistic sample to a non-probabilistic

[6] They have been purposely named experimental subjects and not social actors, because it makes clear the idea with which the experimental research context considers subjects to be part of the research, where the distance between being part of the research in the most 'aseptic' way possible lays the foundations of the method itself.

one (Kruskal and Mosteller, 1980). It should be noted that the concept of inferentiality relies on two basic requirements, the representativeness and randomness, both decisively challenged by Marradi (1997, p. 23-87) who claims that randomness makes it impossible to assume the representativeness of a sample and vice versa, and that the tendency is to continue to invalidate the concept of inferentiality. However, of equal importance is the Central Limit Theorem[7] and

> the extent of the sample is directly proportional to the desired level of confidence in the estimation and the variability of the studied phenomenon, and inversely proportional to the error that the researcher is willing to accept. This means that the size of the population is of no major importance in determining the sample size, in fact, for example, a sample of 1,000 cases may be sufficient to arrive at the same levels of precision for estimates for populations of 10,000 and 100,000 elements. At the most, if precise estimates to two percentage points are desired, 2,500 cases are sufficient for any population size, including global. (Corbetta, 1999: 320).

There have been authors in the past, among them Kish and Frankel (1974), who have challenged this view of the plausibility of the experimental method in the field of social research, but today with Big Data and machine learning the situation becomes even more complicated.

Starting from these assumptions, it is worth emphasising that there is not just one type of error, there is the systematic error more commonly known as bias, the accidental error, the selection error, errors made at the indication, operationalisation, selection and observation stages. Let us try to clarify, in the more purely theoretical phase there are errors of indication, i.e. a type of error that encompasses a series of aspects linked to the coverage of the sample (mentioned in the previous paragraph), of the universe being surveyed and the relative non-responses. In the more purely empirical phase, on the other hand, there are errors of operationalisation, i.e. those linked to the intrusion of the interviewer, the respondent and the way in which the survey itself was carried out.

We have thus highlighted how error in its different meanings is an overrated aspect, especially when discussing sampling. If we think at error from a mathematical, statistical and/or probabilistic point of view and if we thought we could get around it by using e-methods, we will again make a mistake. They are completely different realities; they are two completely, different epistemological points of view. At the same

[7] By increasing the sample size n, the average distribution of a sample from any population with a finite variance approximates the normal distribution (mean and standard deviation) regardless of how the variable used to calculate the average is distributed in the population.

time, it would be a mistake to think of e-methods as a panacea for all the weaknesses of traditional probabilistic approaches. Certainly, they can help us to analyze data that we once did not have at our disposal, they can provide a different, new point of view and help us understand a complex and different reality. Having made these considerations, social relations in themselves are already extremely fluid, social media and the algorithms with which we analyze most of the data available to us are equally fluid, as is the complexity on the cognitive plane in which we move is extremely changeable.

Those considerations done, when we talk about bias we use to attribute a purely negative connotation to it, but this is not always true; in fact, unexpected influences should not always be perceived as "mistakes", i.e. "errors" that contaminate the quality of the recorded information and thus make the information itself to be considered as unusable. We could, instead, think of biases as a space-time continuum, in which possible distortions can actually modify the Data Quality to varying degrees and, in some cases, allow us to discover new aspects to which not much weight had been given, becoming non-negligible areas of research for the study of the phenomenon.

In this regard, it is easy for data from the web to be affected to varying degrees by distortions that we could distinguish into two macro categories: on the one hand, we have the possibility of incurring errors due to the tool preparation, as it happens in traditional research contexts, in other cases, instead, the biases derive from big data and from the data warehouses themselves, which can alter any information through different digital formats; therefore, it will be up to the researcher to understand these differences and "transform" the format of the data in its useful form, aimed at the analysis that is intended to be carried out.

So far we have considered all the different forms of error possible from the point of view of the 'classical' methodology of social research, superimposing them, when possible, on e-methods, but with AI the situation becomes even more complicated. If, on the one hand, when we discuss e-methods, albeit from a point of view more strictly linked to the operational aspects of data analysis, we have been able to distinguish and find a sort of compromise between the classical techniques and those in use today thanks to web contexts, it is much more complex to examine the concept of representativeness and the margin of error in contexts linked to AI, to Big Data understood in its broadest sense, i.e. linked to machine learning. We start from the assumption that AI and machine learning are strongly connected, and this has been reiterated several times within this article, but how are they connected?

AI can be described according to Burstein, W. Holsapple, & Power (2008) as a decision-support system consisting of: a set of data (observations), the input of an analysis process from which indications can be obtained to define decisions and corresponding actions that, evaluated with measurable results, can have an effect on the achievement of an overall goal (value growth). The basis of this theory is data distinguished in terms of volume, velocity, veracity and variety, fully echoing the definition of Big Data as already mentioned several times. AI can thus be used to provide us with a range of information aimed at telling us what has happened (TRON, 2020), or what is happening (Randazzo et al., 2018) and why it is happening (Benanti and Maffettone, 2024). We are thus linked to a field of research that carries out results related to reporting, or to the description or diagnostics of the event that is being investigated, but AI also answers predictive questions such as what is best to do in order to achieve a set goal (Warren, Lipkowitz, Sokolov, 2019). Precisely in the latter case, the link between AI and machine learning is extremely obvious; it is predictive analysis that makes the automation process the most appropriate one to provide us with such results. As discussed in the first paragraph, chatbots are the interface with which we relate within this human- machine relationship. The machine learning algorithm is used to simulate intuitive capabilities, designed in a data-driven manner thanks to previously learned information, i.e. in the code-writing phase there is a phase during which - according to the scientist interpretation - the algorithm would learn 'reality' in an 'objective' manner from the data set defined a priori by the programmer. The choice regarding which data set to use for the algorithm to learn information is dictated by an exclusively 'human' choice, certainly defined by cultural norms and rules of belonging, which can influence through systemic distortions the model itself (Airoldi, 2022). This type of problem certainly falls within the biases we have previously discussed. It is a type of error which in its declination is difficult to ascertain and define aprioristically and which in some ways comes close to the error as we know it in social research concerning the intrusion of the researcher.

5. CONCLUSIVE CONSIDERATIONS

Even before AI entered our everyday lives as pervasively as it is perceived to be today, in the pre-pandemic period (Kyriakidis, 2019), there was already international discussion of the black box that characterises the use of systems based on algorithms and computational

methods. In particular, the opacity with which the mechanisms of operation are characterised was already being discussed, aspects that among other things still arouse a certain reflexive scepticism, not only for those who use them as end-users, but also for researchers, who often do not possess such specific computer skills as to understand the mathematical and informatics mechanisms. Some of these therefore opt for a scientistic view, while others distance themselves from it by questioning the reliability of many AI solutions, which turn out to be poorly understood and scarcely usable, and for this reason, even less safe and efficient, especially when it comes to automated systems, think for instance of the use of robotics in the workplace.

The article highlighted some basic questions of the concept of representativeness, understood in a broad sense, looking not only at statistical representativeness, but especially at all those aspects with a broader scope. Obviously, an attempt has been made to provide answers starting from a classical view of social research methodology, passing through e-methods up to AI, which gives us a new point of view with respect to the concept of data reliability (Neresini, 2015).

According to the author, it is therefore a question of drawing new boundaries of investigation, because not only are the tools not conceived as human-centred (Amershi et al., 2014) and thus designed not so much to improve human capabilities, but rather to replace them, but they are also obscure for the majority of researchers, who have to deal with new research horizons, understand their potential, versatility, but above all their limits (Xu, 2019). AI is thus not inclusive, certainly it is increasingly easy for end-users to make use of chatbots even for those who have a significant digital divide, but understanding the mechanisms and using AI research fields moves us onto a completely different epistemological plane. According to Shapin's (1982) argument following the Science Studies Unit (SST) strand in Edinburgh, people produce knowledge on the basis of the knowledge they inherit within their own culture, their own collectively situated purposes and the information they receive from natural reality, and not least the role of the 'social', i.e. the importance of pre-structuring, not precluding (the scientist's) choice (Shapin 1982: 196-198). This view of the SST school of thought can also relate well to AI, and it is of the utmost importance to make conscious and culturally defined choices with respect to the choice of the data base on which to base the roots of machine learning, the process underlying AI, thus greatly limiting the 'noise' that can result. Not a few mistakes have been made in this regard in the past, with extremely evident consequences on the use of AI with blatant racial and gender discrimination, etc. etc. On

the other hand, AI certainly responds to the need for collectively situated purposes through predictive information, reporting, etc.

When discussing new research frontiers, it is necessary to understand where we are, how we got there and where we are going. This is why the first paragraph presented the author's view of the borderline areas between the classical researcher's toolbox and that within which e-methods fall. From this initial point of departure, a series of arguments concerning representativeness, error and the purpose of understanding the plausibility of AI in social research were discussed.

To recapitulate, therefore, speaking of error, of representativeness in the field of AI in social research is, in my opinion, obsolete, because only the phases of construction of a model at the basis of machine learning can in some way be the object of attention on the part of the researcher and limit any errors linked to the choice of the data base, data set, ect. On the other hand, it is becoming increasingly difficult to enter into research activities that outline an algorithm from scratch, rather than the choice of a specific dataset. We are now in a field where we are mainly users of a wider blockchain, branching out across different datasets, with machine learning algorithms now defined and possibly modified in real time according to specific problems and/or interests already defined ex-ante. In the rare cases in which, the researcher can intervene within the process of building an algorithm and/or a chatbot, the social researcher must have a broad mathematical and statistical knowledge in order to be able to really contribute effectively within a research team dealing with the creation of new AI models and to be able to possibly benefit, at a later stage, from some of the Big Data processed, created, collected, etc.

In my opinion, too many steps are currently missing for the social researcher to be able to answer the questions concerning the margin of error of the research, the concept of replicability, and the plausibility of the tool. Certainly, AI is not a fire that burns out and goes out quickly, it will be the future and in the field of research there are certainly several ways in which it can be used.

The examples are countless, I will list a few without any claim to exhaustiveness. You can summarise a text, a manual, an article and be able to interact in real time with a chatbot to discuss it together. Being able to summarise a paper, a book, a longer or shorter text and turn it directly into a podcast to listen to it in audio format. Asking the AI to carry out a bibliographic reconnaissance from a lemma. There are also some chatbots that transform an analysis of quantitative data into a research report, not only with the most appropriate graphs to describe the phenomenon under study, but with a commentary below it (Mayr and

Schreder, 2014).[8]

I have listed just a few examples to reflect on the fact that there are merits and demerits to each one. Certainly it is useful to be able to summarise some of the texts, provided that we go into them in greater detail later on when the need arises. As the dissemination of scientific information increases, one extremely important aspect is missing, however, and that is to consider that it is fundamental to design AI without forgetting that social, environmental, economic, legal, etc. aspects must coexist in order to respect all the ethical values for each part that makes up the blockchain. The focus then shifts from the concept of representativeness, which as already mentioned is not applicable as we know it in the field of social research, to the ethical aspects, where the ethical design and development of AI systems requires a continuous synergy between a research team that constantly re-addresses the values.

AI's abilities to display human capabilities such as reasoning, learning, planning and creativity will become increasingly powerful (Kissinger, Huttenlocher, Schmidt, 2023). We are already at the stage where AI is able to disentangle itself within complex systems and understand its environment, relate to what it perceives and solve problems accordingly. That is why it is no longer so much a question of representativeness, the boundaries are already extremely blurred, because they are extremely fast, changeable with a level of continuous implementation that one can hardly even imagine. It is therefore crucial that AI is trained on a social-ethical level, with strong values that lay the foundations so that the transcription of the algorithms underlying machine learning is not self-reproduced by the machine itself in continuous evolution, but is driven by people, for a process that is no longer machine-centric, but human-centric.

REFERENCES

AIROLDI, M. (2022). *Machine Habitus. Toward a sociology of algorithms*. Cambridge: Polity Press.

AMERSHI, S., CAKMAK, M., KNOX, W.B., KULESZA, T. (2014). Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine*. 35(4): 105- 120.

[8] Among the most famous is the International System Of Typographic Picture Education, a data analysis and visualisation system that uses a visual language created and reshaped from previous data visualisation experiences to create new hybrid forms of communication between statistics, information, graphic design and data visualisation. It deals with demographic trends, the environment and the economy.

BENANTI, P. MAFFETTONE S. (2024). *Noi e la macchina. Un'etica per l'era digitale*. Roma: Luiss University press.

BETHLEHEM , J. BIFFIGNANDI, S. (2012). *Handbook of Web Surveys. Wiley Handbooks in Survey Methodology*. New Jersey: John Wiley & Sons.

BUCCHI, M. (2019). *Scienza e società. Introduzione alla sociologia della scienza*. Milano: Raffaello Cortina Editore.

BURROWS, R., SAVAGE, M. (2014). After che Crisis? Big Data and the Methodological Challenges of Empirical Sociology. *Big Data and Society*. 1: 1-6. doi: 10.1177/2053951714540280.

BURSTEIN, F., W. HOLSAPPLE, C., POWER, D. J. (2008). Decision Support Systems: A Historical Overview. *Handbook on Decision Support Systems*. 1: 121–140. https://doi.org/10.1007/978-3-540-48713-5_7

COCHRAN, W. (1953). *Sampling Techniques.* New York: Wiley.

CORBETTA, P. (1999). *Metodologia e tecniche della ricerca sociale*. Bologna: Il Mulino.

CLOITRE, M., SHINN, T. (1985). Expository practice: Social, cognitive and epistemological linkages. In T. Shinn, R. Whitley R., *Expository Science. Forms and Functions of Popularization* (pp. 31-60). Dordrecht-Boston: Reidel Publishing Company.

CORPOSANTO, C., MOLINARI, B. (2014). La piattaforma online come strumento di rilevazione e fonte di possibili scenari interpretativi. *Salute e Società*. XIII(3): 103-117.

CORPOSANTO, C., MOLINARI, B. (2017). Big Data and the evaluation of policy. *Riv Rassegna Italiana di valutazione*. XXI(68):84-102. DOI: 10.3280/RIV2017-068006.

CORPOSANTO, C., MOLINARI, B. (2018). Analizzare dati di microblogging con la Sentiment Analysis. Quale rappresentatività?. *Sociologia Italiana*. 11: 123-132.

CORPOSANTO, C., MOLINARI, B. (2020). Dai Big Data alla valutazione passando per la metodologia della ricerca sociale. In S. Gozzo, C. Pennisi, V. Asero, (et all) (a cura di), *Big Data e processi decisionali. Strumenti per l'analisi delle decisioni giuridiche, politiche, economiche e sociali* (pp. 73-83). Milano: Egea.

CORPOSANTO, C., MOLINARI, B. (2022). How Does the Error from Sampling to Big Data Change?. *Italian Sociological Review*. 12(7S): 665-684. https://doi.org/10.13136/isr.v12i7S.576.

CORPOSANTO, C. MOLINARI, B. PAGANO, U. (2024). Online research to capture the essence of social relationships in digital society?. *Sociologia Italiana. AIS Journal of Sociology*. 25: 53-69.

EUROPEAN PARLAMENT (2021). Artificial Intelligence Act. (21 April

2021). "Proposal for a regulation of the European Parliament and the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts." EUR-Lex - 52021PC0206 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELLAR:e0649735-a372-11eb-9585-01aa75ed71a1.

GRIMALDI, R. (2005). *Metodi formali e risorse della Rete. Manuale di ricerca empirica*. Milano: FrancoAngeli.

HENRY, G.T. (1990). *Practical Sampling, vol. 21*. Newbury Park: Sage Pubblications.

KISSINGER, H. HUTTENLOCHER, D. SCHMIDT, E. (2023). *L'era dell'intelligenza artificiale. Il futuro dell'identità umana*. Milano: Mondadori Editore.

KITCHIN R. (2014). Big Data, New Epistemologies and Paradigm Shifts. *Big Data & Society*. I. 1: 1-12. doi: 10.1177/2053951714528481.

KYRIAKIDIS, M., DE WINTER, J. C. F., STANTON, N., (et all) (2017). A human factors perspective on automated driving. *Theoretical Issues in Ergonomics Science*. 20(3): 223–249. https://doi.org/10.1080/1463922X.2017.1293187.

KISH, L., FRANKEL M.R., (1974). Inference from Complex Samples. *Journal of the Royal Statistical Society: Series B (Methodological)*. 36(1): 1–22. https://doi.org/10.1111/j.2517-6161.1974.tb00981.x.

KNORR CETINA, K. (1981). *The manufacture of Knowledge: An Essay on the constructivist and Contextual Nature of Science*. Oxford: Pergamon.

KRUSKAL, W., MOSTELLER, F. (1980). Representative Sampling, IV: The History of the Concept in Statistics 1895-1939. *International Statistical Review.* XLVIII: 169-195.

LATOUR, B. (1983). Give me a laboratory and I will raise the world.In K. Knorr Cetina, M. Mulkay, (eds.), *Science Observed*, (pp. 141-170). London: Sage.

LIU, Y., WANG, Y., & ZHANG, J. (2012). New machine learning algorithm: Random forest. *Computer Science*. 7473 LNCS: 246–252. https://doi.org/10.1007/978-3-642-34062-8_32.

MARRADI, A. (1989). Casualità e rappresentatività di un campione nelle scienze sociali: contributo a una sociologia del linguaggio scientifico. In R. Mannheimer, *I sondaggi elettorali e le scienze politiche. Problemi Metodologici* (pp. 51-53). Milano: FrancoAngeli.

MARRADI, A. (1997). Casuale e rappresentativo: ma cosa vuol dire?". In P. Ceri (a cura di), (1997), *Politica e sondaggi* (pp. 23-87). Torino:

Rosenberg & Sellier.
MAYR, E., SCHREDER, G. (2014). Isotype Visualizations as a Chance for Participation & Civic Education. *Conference for E-Democracy and Open Governement*. 6: 136-150.
MOLINARI, B. (2014). Survey e questionari online?. In C. Corposanto, A. Valastro (2014) *Blog, FB & TW*, (pp. 17-42). Milano: Giuffrè.
MOLINARI, B., CORPOSANTO, (2018). Big Data and the evaluation of policy. *Riv Rivista Italiana di valutazione*. 68(2): 84-102.
MOLINARI, B., CORPOSANTO, C (2023). The error role in risk perception. *Salute e Società*. XXII (1): 45-57.
NERESINI F. (2015). Quando i numeri diventano grandi: che cosa possiamo imparare dalla scienza. *Rassegna Italiana di Sociologia*. 3(4): 405-431.
PALUMBO, M. GARBARINO, E. (2006). *Ricerca sociale: metodo e tecniche*. Milano: FrancoAngeli.
QUARTERONI, A. (2020). *Le equazioni del cuore, della pioggia e delle vele. Modelli matematici per simulare la realtà*. Bari: Edizioni Dedalo.
QUARTERONI, A. (2013). *Matematica numerica. Esercizi, laboratori e progetti.* Berlino: Springer Verlag.
SHAPIN, S. (1982). History of Science and its Sociological Reconstructions. *History of Science*. 20(3): 157-211. https://doi.org/10.1177/007327538202000301.
RANDAZZO, L., ITURRATE, I., PERDIKIS, S. (et all) (2018). Mano: A Wearable Hand Exoskeleton for Activities of Daily Living and Neurorehabilitation. *IEEE Robotics and Automation Letters.* 3(1): 500–507. https://doi.org/10.1109/LRA.2017.2771329.
ROSENTHALM, R., JACOBSON, L. (1968). Teacher Expectations for the disadvantaged. *Scientific American.* 218(4): 19-23.
THOTTOLI, M.M. (2024). Leveraging information communication technology (ICT) and artificial intelligence (AI) to enhance auditing practices. *Accounting Research Journal.* 37(2): 134-150.
THRON. (2020). THRON: Digital Content Management Software. Retrieved December 8, 2020, from https://www.thron.com/en/.
TROBIA A. (2014). Web Mining e Application Programming Interfaces per la ricerca sociale: caratteristiche, strumenti, prosepttive e limiti. In C. Corposanto, A. Valastro (2014) *Blog, FB & TW*, (pp. 67-104). Milano: Giuffrè Editore.
WARREN, J., LIPKOWITZ, J., SOKOLOV, V. (2019). Clusters of Driving Behavior from Observational Smartphone Data. IEEE *Intelligent Transportation Systems Magazine*. 11(3) : 171–180.

https://doi.org/10.1109/MITS.2019.2919516.

WONNACOTT, T.H. (1969). *Introductory Statistics*. Hoboken: John Wiley and Sons.

XU, W. (2019). Toward Human-Centered AI: A Perspective from Human-Computer Interaction. *Interactions.* 26(4): 42-46.