TRANSLATION FROM THE ORIGINAL SUMMARY IN SPANISH

Seminar 'Digital Footprint: Servitude or Service?'

Third-party assessment of big data processes

(Summary of the session of February 18, 2021)

The expert committee of the **Seminar 'Digital Footprint: Servitude or Service?'** held its tenth session on February 18 by videoconference. The session began with some comments by the seminar management team about the progress on the topics already discussed, on some issues in which more information is required, and the type of conclusions to be presented at the end of the seminar in June. Conclusions could be formulated mainly on regulation, self-regulation, and education of critical capacity. This session itself dealt with the nature of *artificial intelligence* (AI) *systems* and third-party assessment of big data processes.

The initial presentation was given by José Luis Calvo, Director of AI at SNGULAR, followed by comments from Pablo García Mexía, Consultant-Director of Digital Law at Herbert Smith Freehills, Lawyer of the Courts, and Gloria Sánchez Soriano, Director of the Technology and Transformation Legal Department of the Santander Group.

Demystify AI

Throughout the previous sessions, frequent mention was made of a kind of mythology around autonomous and intelligent systems. For José Luis Calvo, landing the concept means, above all, to remember that AI is just software. Although it is in the last five years when AI has begun to give surprising results, its history dates back to the emergence of the first computers.

The current boom is due to the recent development of so-called deep neural networks, which are part of machine learning techniques. At first, it was based on deterministic algorithms, a set of precise instructions that the machine has to execute. These algorithms were created by hand by developers, who controlled the casuistry that the software would have to deal with. Starting from this deterministic base, which remained stable during the last fifty years, machine learning techniques or probabilistic algorithms have appeared lately. In these techniques, the algorithm has a statistical foundation and is trained with a set of data that already contains a result and conditions the training. Based on a given pattern, the algorithm adjusts until it creates a model or program. In such a way, upon receiving a new data set, the model can generate results on its own.

Unlike the traditional type, software created with AI techniques can create their own model, which goes much farther than a few steps and rules set by a developer. Some of these models are understandable, such as decision trees, but others are complex and difficult to explain: the so-called "black boxes." Advances in models are essentially linked to the availability of large volumes of data and a much larger computing capacity. (Clearly, the concept of "huge amounts of data" is variable: what for a small or medium-sized company is a massive amount of information, is not even representative for an international company).

In summary, the evolution of AI systems in recent decades goes from top-down models (based on given guidelines and their application) to a new type of probabilistic models that are possible thanks to the increase in the volume of data and machine learning capacity: bottom-up models, such as the aforementioned neural networks.

Current and potential problems of AI

The problems usually related to the use of AI are not new. However, these new models' nature implies that these kinds of issues could be potentially harmful on a larger scale and for many more people. The processes are still used to automate activities previously carried out by people, but with a marginal cost that tends to zero. In other words, they could be extended and grown almost without limits. However, experts recognize that these processes demand considerable computing capacity and data storage volumes; It would be interesting to have data on this to estimate the physical limits and the actual costs (externalities) of AI systems. It is also worth wondering about the sustainability of the "business model" built on exploiting the "digital footprint" and the allocation of high advertising costs to advertisers, with substantial profit margins for *big tech*.

On which aspects could third-party evaluation of big data processes and the use of AI have an impact? Four groups of problems are mentioned: misinformation and polarization, privacy, discrimination, and those that arise from the surveillance of users.

Disinformation and polarization problems arise from recommendation algorithms. These types of algorithms are not only used to recommend which movie or book users will like, but also to address them with news and content that promote a possibly biased point of view and favor the spread of fake news. In contrast, there is the risk of a limitation of freedom of expression by States or by *big tech* themselves, justified by the need to avoid disseminating false or abusive content. Platforms are applying censoring practices using algorithms designed to block access to certain content, which raises the question of the legitimacy or arbitrariness of such interventions.

Privacy problems are possibly the most recognized by society since the raw material of algorithms is the data obtained from our "fingerprint" activity. It is the data provided by users that feed and make predictions possible. But in obtaining and manipulating the data, privacy problems arise, as in the cases in which the algorithm, combining different data ceded more or less voluntarily, is capable of reconstructing information that had not been ceded as such. That is the case, for example, when an AI system can obtain information about the sex of a person through the image of their retina. There are many examples in which the model is trained to allow predicting private user information. Thus, users cede some data from which other data are deduced, which they would not necessarily want to make public, and which could be used, not necessarily for their benefit. This is perhaps the heart of the matter of the "behavioral surplus" constructed from our fingerprints and with which it is intended, with or without our assent, to predict our behavior with increasing precision. On this, it would be interesting if the seminar had some additional information elements.

Discrimination problems stem from the fact that the algorithm training is based on a data history from which new patterns are created, reflecting previous selective situations. This can lead to justice problems when someone is valued based on historical data that determine the system's decisions in favor or against the person. Plus, this may imply that previous discriminations are reinforced, and bias is perpetuated.

Indeed, the data sometimes has a hidden bias that the developers may not see at first. In that case, it may result that the new systems are helping to promote discrimination that we as a society want to avoid. At other moments, the bias results from a lack of effort to include diversity in the data, or it may even respond to a manipulative intention on the part of the developer or its constituents. In any case, making automated decisions based only on information from the past implies limits and may produce effects that are not adjusted to the current reality.

The model's accuracy is key to reducing and/or avoiding false positives (such as granting credit to those who cannot pay it) and false negatives (such as not giving credit to those who deserve it). The fine-tuning of the models should attend to both extremes so that, in the example mentioned, reducing false negatives would imply that loans are granted to a more significant number of people who can be good debtors. Reducing false positives would entail ceasing to give credit to people who may not repay, thus avoiding direct suffering to these people and indirectly to the community.

The last-mentioned risk, that of surveillance, is the subject of multiple studies and has been widely debated in the media in different countries, perhaps being the one that implies the most significant threats for the future. It arises, among other situations, in scenarios in which programs such as those that allow traffic control, voice recording and identification, or facial recognition for airport security can be used to carry out other types of monitoring that could violate privacy and the rights of individuals. It is an open debate before which companies such

as IBM, Amazon, and Microsoft, for example, have taken a step back in the commercialization of facial recognition tools.

Controlling AI

Governments, enterprises, and academic institutions have proposed various ethical principles and codes to promote a human-centered AI. The debate is between those who favour government intervention above all - prior or *a posteriori* - and those who see more future in self-regulation, an excellent design of systems, and the possibility of certifying their operation by independent sources.

Among the codes usually suggested, there is a set of principles proposed, among others, by authors such as the Spanish engineer Nuria Oliver ("Fair, transparent, and Accountable Algorithmic Decision-making Processes," an article published in 2017 in the Philosophy & Technolgy magazine): autonomy, justice, beneficence and non-maleficence, that is, the four principles of bioethics, plus one more relating to the transparency necessary for the ethical development of AI models.

The control of the AI models must attend to the three critical pieces of the systems: the data used for the algorithm's training, the type of algorithm concerning the application scenario, and the result of the model.

Concerning the first of these essential pieces - the data - Europe has positioned itself with the General Data Protection Regulation, widely commented on in previous sessions of the seminar. Additional ideas are suggested, such as introducing ethical studies starting from the design of the model or the prohibition of the commercialization of personal data, analogous to organ trafficking. In this perspective, personal data could be used in specific contexts, but they could not be commercialized.

Data is the raw material of AI; therefore, attending directly to aspects such as the quality of the data is essential to avoid problems in the evaluation by third parties. Thus, the data must come from reliable sources; the model must contain the necessary amount of data to prevent or minimize the risk of bias in such a way as to ensure the highest possible accuracy; all this implies that the data has not been altered. If the “training” data meet these requirements, it can be a good starting point.

Regarding the algorithm, the possibility of regulating typologies is mentioned when creating a model. For example, the use of "black box" models could be prohibited when they may affect people, and their effects may pose a greater risk. In these issues, such as video surveillance or labor contracts, only the use of easily understood models could be allowed, thus mitigating the

tension between statistical correctness and individual justice. Other participants see it as more feasible, instead of regulating the algorithm, to control the use cases and the purposes for which it is used, considering the levels of risk they entail. This can be the object of fruitful preventive reflection by companies and institutions. However, trying to obtain these limitations through legal provisions could limit the socially positive uses of the processes and would be equivalent to wanting to "put gates to the field."

Finally, the AI evaluation must attend to the result of the model to ensure they are auditable. An algorithm can use data that respects privacy and biases, be explainable and transparent, but its effects can violate social justice paradigms. Cases like this are, for example, the models used for face identification that work best with white persons. Some voices from the seminar emphasize that it is essential to pay attention to the fact that, if the result reflects reality, even if that reality is not always liked (or is politically incorrect and/or surprising), it is still true or real. In other words, one should avoid falling into the error of not analyzing reality because the information can scare us. Perhaps an analysis could be proposed by comparing the result obtained on the same data but with other models, or maybe we should search for further ways of validation.

About the type of regulation

In the face of all attempts at preventive and/or voluntary regulation or evaluation, some people think that the only effective regulation is *a posteriori*, on the final results. As in other fields, the law would punish situations in which something intolerable has occurred. Given this, the response would be applied in the form of a fine, possibly a heavy sanction. This proposal aims to save the competitive disadvantages that laws bring to countries with strict regulation: regardless of where the algorithm has been developed and under what law, it must respect pre-established norms, or it will be sanctioned with a fine. In addition, this approach is a powerful incentive to preventive behavior on the part of the agents, who will try to avoid the sanction.

However, and without this necessarily being incompatible with the above, some voices defend the need for ethical requirements to start from design. Algorithms should be auditable by design, and interpretability should also be included by design, either by regulations or by codes of conduct. Appealing to design implies defending thinking before doing. Here, it is proposed to transfer the concept "by design" to the problems of data management and thus speak of responsibility "by design," ethics "by design," and control "by design." The said entails understanding the mechanisms, the risks, and the phases to establish the requirements and rules. Also, to monitor compliance, the responsibility of each area must be formally established. A proposal to advance in possible solutions proposes establishing six compliance phases: conceptualization, data, model development, launch, control and feedback, and finally, error

management. These phases cover the control of the various problems that arise in companies' data management. With these six phases in mind, each organization should adapt the control system to its processes. All of this can be independently certified, as is the case with other manufacturing or organizational processes.

In this scenario, there is a need to identify legal responsibility when using algorithms, especially in high-risk cases to integrity and individual rights. In legal matters, the idea (shuffled in Europe after 2015) of building a type of "electronic personality" for intelligent systems, similar to that of companies, associations, or foundations, seems to have been diluted. For the moment, it appears that this idea has been abandoned because in the case of "electronic personality", we would no longer be talking about a "compound" of human beings, but the concept of person would be extended beyond the human sphere. It seems more appropriate to demand that there is always one person, beyond the algorithms, who is responsible for the activity and the results of the system.

The automation bias

Even before the risks mentioned above - such as polarization, privacy, data bias, and surveillance - another deep-rooted risk is in sight, the so-called automation bias, that is, a tendency to blindly believe in the results of automated systems. Given this, it is necessary to claim that human beings should not abandon their judgment capacity nor delegate their reasoning to robotic systems. In today's age, as mentioned in other sessions, the individual has a responsibility to keep free. Society has the responsibility to develop a critical judgment on the issues in its environment, understand reality, try to improve it, and make it fairer. It is human to trust technology, but the problem would be to abandon our reasoning to machines without analysis on the part of humans.

Along these lines, it is interesting to note that there have been occasions throughout history in which ignorance has led to accepting as given some propositions that were finally seen to be untrue. For example: in the middle of the 20th century, with the invention of television, advertisements acquired an unusual force: at first, what was advertised was not questioned, so housewives, office workers, and businesspeople were willing to buy what was advertised without asking about the marketing techniques. Something similar happens now with AI: ignorance plays a relevant role in the importance given to it in society. The opacity with which AI models work generates significant social risks, and the unpredictability of their consequences adds an increased risk factor. But it is possible that "superintelligent" automatic processes are in a still relatively primitive state of evolution and that, with their subsequent development, they will continue to be corrected with elements of the irreplaceable human "common sense".

With or without strong regulation, there remains the need to increase education and the promotion of knowledge in all these matters in order to forge a critical sense. Although the subject is of interest to the whole of society and general social education in the field of AI models should be aimed at, it is essential to start with the technologists so that they can see beyond the development of the system itself and are concerned about responsibility in the design of the AI. Universities should promote the education of engineers so that throughout their training period, they are aware of the problems and social challenges that technology implies. It is also imperative that education on these topics covers companies that buy algorithms or request specific results from engineers and developers. If the top managers of the companies ask the technicians what they want to be done without thinking about the impact, the responsibility will also be on whoever orders a specific service to be delivered, and not only on the model developer. And there needs to be a common language on the matter between those who make business decisions and those who develop the computer programs.

As a tool in the face of automation bias, education is the key to preventing technology from replacing the human being in the reasoning that leads them to make decisions regarding the direction that they socially want and should take. There is a tendency to blindly trust algorithms' results; this should be avoided through regulation and education. What has been said implies a defense of human freedom, a characteristic that is part of humans' intrinsic essence and cannot be overridden by machines. In education issues, it is not so much about training minds in a specific discipline, but rather a training that tends to raise awareness of the reality, so shocking and undoubtedly unknown to many, of the dynamics of the digital world.

As a final thought, some emphasize, once again, the extraordinary benefits that intelligent systems can bring when put at the service of an innovative and selfless research project. An example of this type can be the Human Brain Project led by the Spanish neurobiologist Rafael Yuste, a medical, scientific, and technological project that, through advanced techniques, proposes to know in-depth and map the "black box" of the human brain. The possibility of reading and perhaps altering brain activity opens up immense opportunities and raises the need for ethical criteria in a still unknown dimension. On these topics, the seminar will dedicate its next session, scheduled for April 15.

Attendees:

1. **Alfonso Carcasona**, CEO, AC Camerfirma
2. **Alfredo Marcos Martínez**, Professor of Philosophy of Science, Universidad de Valladolid
3. **Ángel González Ferrer**, Executive Director, Digital Pontificial Council for Culture

4. **Carolina Villegas**, Researcher, Iberdrola Financial and Business Ethics Chair, Universidad Pontificia de Comillas
5. **Cristina San José**, Chief Data Strategist, Banco Santander
6. **David Roch Dupré**, Professor, Universidad Pontificia Comillas
7. **Diego Bodas Sagi**, Lead Data Scientist – Advanced Analytics, Mapfre España
8. **Domingo Sugranyes**, Director, Seminario de Huella Digital
9. **Esther de la Torre**, Responsible Digital Banking Manager, BBVA
10. **Francisco Javier López Martín**, Former Secretary-General, CCOO Madrid
11. **Gloria Sánchez Soriano**, Transformation Director, Legal Department, Banco Santander
12. **Idoia Salazar**, AI ethics expert, Universidad CEU San Pablo
13. **Idoya Zorroza**, Contracted Professor Doctor, Faculty of Philosophy, Universidad Pontificia de Salamanca
14. **Javier Camacho Ibáñez**, Director of Ethical Sustainability and professor at ICADE and ICAI
15. **Javier Prades**, Dean, Universidad Eclesiástica San Dámaso
16. **Jesús Avezuela**, General Director of the Pablo VI Foundation
17. **Jesús Sánchez Camacho**, Professor, Faculty of Theology, Universidad Pontificia Comillas
18. **José Luis Calvo**, AI Director. SNGULAR
19. **José Manuel González-Páramo**, Former Executive Director, BBVA
20. **José Ramón Amor,** Coordinator, Bioethics Observatory of the Pablo VI Foundation
21. **Juan Benavides**, Professor of Communications, Universidad Complutense de Madrid
22. **Julio Martínez s.j.**, Dean, Universidad Pontificia Comillas
23. **Pablo García Mexía**, Digital Jurist, Of Council Ashurst LLP
24. **Paul Dembinski,** Director de Observatoire de la Finance (Ginebra)
25. **Raúl González Fabre**, Professor, Universidad Pontificia de Comillas
26. **Sara Lumbreras**, Deputy Director of Research Results, Associate Professor, Institute for Technological Research, ICAI, Universidad Pontificia Comillas
27. **Richard Benjamins**, Data & IA ambassador, Telefónica