ALGORITHMIC TRANSPARENCY AND DISINFORMATION

LAPIN
LABORATORY OF PUBLIC
POLICY AND INTERNET

Authors:
Gustavo Ribeiro
Julia D'Agostini
Paulo Sarmento
Raquel Rachid

Revision:
Amanda Espiñeira
Cynthia Picolo Gonzaga de Azevedo
José Renato Laranjeira de Pereira

Layout:
Pietra Polo

How to refer this report:
RIBEIRO, Gustavo; D'AGOSTINI, Julia; SARMENTO, Paulo; RACHID, Raquel. **Report on Algorithmic Transparency and Disinformation**: a multisectoral approach. Brasília: Laboratory of Public Policy and Internet, 2021. Available at [LAPIN's weblink]. Access DD mmm. YYYY.

lapin.org.br @lapin.br /lapinbr /lapinbr

ABOUT LAPIN

The **Laboratory of Public Policy and Internet** (LAPIN) is a pioneer nonprofit think tank dedicated to defending digital rights based in Brasília, Brazil. Researchers, lawyers, engineers, and representatives from both the public and the private sectors contribute to LAPIN's goal of analyzing and supporting the development of public policies focused on the regulation of digital technologies.

On the one hand, LAPIN aims to **investigate, analyze and understand** how the internet and new digital technologies influence the law and society. On the other hand, the Lab works to **propose, inform** and **support** Brazilian society and decision-makers on issues such as privacy, data protection, disinformation, artificial intelligence, and respect for human rights online. Moreover, LAPIN counts with a specialized team focusing on **awareness-raising**, which works to produce accessible and uncomplicated content on complex topics to inform the public in general. The think tank also promotes a constructive dialogue between stakeholders from diverse societal sectors.

SUMMARY

INTRODUCTION

Considering LAPIN´s Disinformation Research Team weekly meetings and discussions based on a few papers[1] that proposed algorithmic transparency as an opportunity for an agenda on confronting disinformation, an **interesting intersection came out regarding disinformation and algorithmic transparency**. Based on that, the team decided to seek materials that could theoretically support research focusing on understanding the context of these discussions. Surprisingly, few materials were found on this specific correlation, which was the incentive for setting this project off.

After initial investigations, evidence that the topic could be related to accountability and transparency came to light, so the team decided to **interview specialists in the field** to get their impressions that could potentially support future research about this matter.

On a day-to-day basis, individuals are continuously introduced to content online through the curation of algorithms, especially in the context of digital platforms. These are very sophisticated computer engineering mechanisms. This discussion has a significant impact on how content is selected and how business models are supported by data in a scenario of global maximization of narrative dispute and disinformation.

Social media plays an important role in how people access information. Access to news on messaging apps increasingly grows. According to the Reuters Institute Digital News Report 2019,[2] WhatsApp has become a primary network for discussing and sharing news in countries like Brazil (53%), Malaysia (50%), and South Africa (49%). The 2020 edition of this research concluded that Facebook and other social media groups are now used on average by around a third of the respondents (31%) for local news and information, and those aged 18–24 are more than twice as likely to prefer to access news via social media.[3]

Social media, as a medium by which content is sent from issuers to receivers, has become an influential locus for public debate, access to information, and news that are increasingly governed by algorithms. Platforms operate on a business model that promotes exchange between users, advertisers, and consumers by processing personal data to tailor an experience for each user's particular appeal. This, in turn, increases content interaction.

From the Internet's infancy in the 1990s, policy-makers and societal sectors have discussed these platform's responsibility as information environments. The current regulatory framework

[1] See: Nourani M, Kabir S, Mohseni S, Ragan ED. **The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems**. The seventh AAAI Conf. HCOMP-19 2019; 97-105.
Mohseni S, Block JE, Ragan ED. **Quantitative Evaluation of Machine Learning Explanations**: A Human-Grounded Benchmark. AAAI Conf. HCOMP-20 2020; Mohseni, S, Ragan ED, Ha X. **Open Issues in Combating Fake News**: Interpretability as an Opportunity. Available at https://doi.org/10.48550/arXiv.1904.03016. Accessed 22 Mar 2022.

[2] Reuters Institute for the Study of Journalism. **Reuters Institute Digital News Report 2019**. Available at: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/inline-files/DNR_2019_FINAL.pdf. Accessed 29 Nov 2021.

[3] Idem.

for content moderation in polities such as the European Union,[4] the United States,[5] and Brazil[6] acknowledge their intermediary role.

The premise is that, unlike traditional media, social media platforms do not edit user generated content and may only moderate information in accordance to prescribed terms of service. With growth, moderation became a complex and layered process. Examples on false and misleading information abound. Determining actionable content may involve, in combination, user reporting, feedback from independent fact-checkers, algorithmic decision-making, and human review. Remedies vary from content removal, anti-misinformation banners, reduced distribution, to measures against repeat offenders.[7]

Notably, the discussion's focus has shifted from responsibility for devising the content of online information to social medias' responsibility for their governance over access to and exchange of information. Their ability to detect; to reduce or increase distribution; to steer content and target users; to remove; and to diffuse in mass; both legal and illegal; benign and harmful content; has put the premise of impartial intermediation in check.[8]

On the one hand, platforms are formulating new strategies at both policy and enforcement levels to address the dissemination of harmful content online.[9] On the other, however, there is skepticism as to whether the pursuance of effective solutions is an interest in conflict with their business model.

[4] In the European Union, Article 12 of the Electronic Commerce Directive (Directive 2000/31/EC) provides immunity from liability where the intermediary Internet Service Provider operates as a "mere conduit" for information; that is, it "(a) does not initiate the transmission; (b) does not select the receiver of the transmission; and (c) does not select or modify the information contained in the transmission." See, European Union, Directive on Electronic Commerce, OJ L 178, 17 Jul 2000, pp. 1-16. Available at: http://data.europa.eu/eli/dir/2000/31/oj. Accessed 18 Jan 2022.

[5] In the United States, Section 230 of the Communications Decency Act "gives online intermediaries broad immunity from liability for user-generated content posted on their sites. The purpose of this grant of immunity [is] both to encourage platforms to be 'Good Samaritans' and take an active role in removing offensive content, and also to avoid free speech problems of collateral censorship." Kate Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech' (2017) 131 Harvard Law Review, pp. 1602, 1603-1609. Available at: https://harvardlawreview.org/wp-content/uploads/2018/04/1598-1670_Online.pdf. Accessed 18 Jan 2022.

[6] In Brazil, Article 19 of the Civil Rights Framework for the Internet or, "[t]he Marco Civil, as the Bill is referred to in Portuguese, (…) protects freedom of expression, creating safe harbors for online intermediaries in Brazil, and determining that online platforms will have to takedown specific content when served with a valid court order." Ronaldo Lemos, 'The Internet Bill of Rights as an Example of Multistakeholderism'. In: Carlos Affonso Souza, Mario Viola and Ronaldo Lemos (eds.), Brazil's Internet Bill of Rights: A Closer Look. **Institute for Technology and Society of Rio de Janeiro**, 2017, p. 43. Available at: https://itsrio.org/wp-content/uploads/2018/02/v5_com-capa__pages_miolo_Brazil-Internet-Bill-of-Rights-A-closer-Look.pdf. Accessed 18 Jan 2022.

[7] For a more comprehensive overview of methods, remedies and long-term policies developed by Internet Service Providers and the advertisement industry, see: Code of Practice on Disinformation, Annex II: Best Practices. **European Commission**. Available at: https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation. Accessed 18 Jan 2022.

[8] "Hosting platforms have reached a central role in allowing access to and exchange of information permitting the mass diffusion of any type of content, both legal and illegal. This raised pressing questions on their responsibility in preventing its diffusion, detection and subsequent removal, and platforms' role in the digital realm has morphed, from that of mere hosting providers to that of actors governing how content is displayed and shared online, undertaking certain actions such as moderation, curation and recommendation." Andrea Bertolini, Francesca Episcopo and Nicoleta-Angela Cherciu, '**Liability of online platforms**', European Parliamentary Research Service, Scientific Foresight Unit, Feb 2021, pp. III, 1, 170. Available at: https://op.europa.eu/s/vHsn. Accessed 18 Jan 2022. See also, "[w]e are witnessing a shift in the primary driver of regulation from protecting innovation at all costs to ostensibly protecting aggrieved citizens at all cost." Bruna Martins dos Santos and David Morar, '**The push for content moderation legislation around the world**', Brookings Institution, 21 Sep 2020. Available at: https://www.brookings.edu/blog/techtank/2020/09/21/the-push-for-content-moderation-legislation-around-the-world/. Accessed 18 Jan 2022.

[9] For an overview of automated moderation, see: Spandana Singh, '**Holding Platforms Accountable**: Online Speech in the Age of Algorithms', New America's Open Technology Institute, 22 Jul 2019. Available at: https://www.newamerica.org/oti/reports/report-series-content-shaping-modern-era/. Accessed 19 Jan 2022. For information on platform's policy enforcement, see: Spandana Singh and Leila Doty, '**The Transparency Report Tracking Tool: How Internet Platforms Are Reporting on the Enforcement of Their Content Rules**', New America's Open Technology Institute, 09 Dec 2021. Available at: https://www.newamerica.org/oti/reports/transparency-report-tracking-tool/. Accessed 19 Jan 2022. For a discussion on platform's response to COVID-19 misinformation, see: Nandita Krishnan et. al., '**Research note: Examining how various social media platforms have responded to COVID-19 misinformation**', Shorenstein Center on Media, Politics and Public Policy at Harvard University, John F. Kennedy School of Government, Dec 2021. Available at: https://doi.org/10.37016/mr-2020-85. Accessed 19 Jan 2022.

For instance, Morozov argues that the problem lies not in the misinformative character of content itself, but rather on "**a digital capitalism that makes it profitable to produce false but click-worthy stories.**"[10]

Evidently, these platforms play an important role in how societies interact with information; therefore, **transparency becomes essential in order to understand how the creation, access to and exchange of information is governed online.**

As companies and governments increasingly assign decisions to algorithms that can affect people's lives, this project focuses on different understandings related to the interaction concerning disinformation and algorithmic transparency from a multi-sector perspective (academia, civil society, government, and private sector).

Considering the proposed discussion arises from the relationship between technologies and societies, all possible understandings about it belong to a concept in dispute - both in terms of its definition and concerning the consequences resulting from the application of diverse legal mechanisms in different political contexts.

While the relationship between disinformation and algorithmic transparency has drawn attention, there are many nuances that this initiative intends to address. When online content curation is mostly recommended by algorithms (about which little is known, especially by users), the question arises as to how transparent algorithms should be and whether it is possible to conceive such transparency in the context of trade secrets protection.

Our team listened to the impression of valuable people working in such a complex field in order to get their perceptions. We interviewed a representative from each sector, i.e., the public and private sector, civil society, and academia, to understand the correlation between algorithmic transparency and disinformation from a multistakeholder perspective.

LAPIN wants to express its deepest gratitude to the very special interviewees who contributed to this project and provided us with great insights over the present discussion.

[10] Evgeny Morozov, '**Moral panic over fake news hides the real enemy – the digital giants**', The Guardian, 8 Jan 2017. Available at: https://www.theguardian.com/commentisfree/2017/jan/08/blaming-fake-news-not-the-answer-democracy-crisis. Accessed 19 Jan 2022. See also: Evgeny Morozov, '**Big Tech: a ascensão dos dados e a morte da política**', Portuguese Translation by Claudio Marcondes, Ubu Editora, 2018.

METHODOLOGY

The methodology comprises semi-structured interview research to produce qualitative material and **analysis on the intersection of algorithmic transparency with content moderation and, especially, disinformation**. For this purpose, LAPIN invited four professionals as representatives of societal sectors (academia, civil society, public and private sectors) for interviews based on previously formulated questions. The interviewees were chosen considering their proximity to the discussion and diverse backgrounds.

However, it should be noted that this report aims to conduct qualitative research and by no means claims that the opinions herein disclosed are an exhaustive representation of each sector's view on the matter. It is important to mention that the representation of a voice per sector cares for a certain balance that is also manifested by the fact LAPIN preferred to keep one single interview for each sector.

The standard questions (Table 1) were internally validated by LAPIN as an important methodological step and the interviews were conducted by a representative of our Disinformation Team between July and September 2021. Although they were asked to all the interviewees, small variations on the standard questions were made in order to adapt to the reality of the interviewee or to the answers given to previous questions during the interview.

Standard questions

Table 1.

#1	How would you define disinformation?
#2	How would you define algorithmic transparency?
#3	Would you say disinformation and algorithmic transparency are related topics? If yes, how do you relate them?
#4	Are you favorable to algorithmic transparency? Why?
#5	If you are favorable to algorithmic transparency, how should it ideally work?
#6	Would you say the model you proposed faces social or sectoral resistance? If yes, which kind of resistance?
#7	Would you like to reference any material (academic or not) that you rely on or support you when thinking about the matters you mentioned during this interview?

The interviewees

Gianclaudio Malgieri

ACADEMIA

Gianclaudio Malgieri is an Associate Professor of Law and Technology at the EDHEC Business School in Lille (France), where he conducts research at the Augmented Law Institute and teaches Data Protection Law, AI regulation, Digital Law, Data Sustainability, Intellectual Property Law and Business Law. He got an LLM with honours at the University of Pisa and a Juris Doctor with honours at S. Anna School of Advanced Studies of Pisa.

Natália Neris

PRIVATE SECTOR

Natália Neris is Senior Public Policy Associate at Twitter. Doctoral Student in Human Rights at the Faculty of Law from the University of São Paulo (FD-USP), Master in Law at the São Paulo Law School of the Getúlio Vargas Foundation, holds a Bachelor's Degree in Public Policy Management at the School of Arts, Sciences and Humanities of the University of São Paulo (EACH-USP).

Vidushi Marda

CIVIL SOCIETY

Vidushi Marda is an Indian lawyer and researcher, based in Bangalore (India), who investigates the consequences of integrating artificial intelligence systems in societies. She currently works as Senior Programme Officer at ARTICLE 19, where she leads research and engagement on the human rights implications of machine learning. Moreover, as an affiliate researcher at Carnegie India, she analyzes law enforcement use of emerging technologies in India.

In the past, she collaborated with DATACTIVE at the University of Amsterdam, Privacy International, among others. She is also part of the Steering Committee at RealML, and a member of the Expert Group on Governance of Data and AI at United Nations Global Pulse.

Xabier Lareo

PUBLIC SECTOR

Xabier is a Technology and Security Officer at the Technology & Privacy unit of the European Data Protection Supervisor (EDPS) and member of the internal AI Task Force. He provides advice on technology developments having an impact on privacy and data protection, contributes to policy papers and takes part on EDPS supervision activities.

Artificial intelligence, anonymization and online tracking and profiling are among the topics he focuses on. He graduated as a Computer Engineer and started his professional career in software development. He previously worked as Data Inspector at the Spanish Data Protection Authority.

In the Part I: Consensus and Dissensus, you are going to find a systematization of the interviewees´ main ideas, before having access to their full interviews.

GLOSSARY

In order to facilitate the understanding of the arguments expressed by the interviewees, there are some important concepts that should be explained. These definitions are being provided by LAPIN with the purpose of contextualizing technical terms. It does not purport to reflect the opinion of the interviewee as to the concept's scope.

Black Box

Frank Pasquale describes the '*black box*' as a metaphor with dual-meaning: "It can refer to a recording device, like the data-monitoring systems in planes, trains, and cars. Or it can mean a system whose workings are mysterious; we can observe its inputs and outputs, but we cannot tell how one becomes the other."[11]

Machine Learning

"Machine learning is a branch of artificial intelligence and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy." It can "detect patterns and learn how to make predictions and recommendations by processing data and experiences, rather than by receiving explicit programming instruction. The algorithms also adapt in response to new data and experiences to improve efficacy over time."[12]

Synthetic Media

Synthetic Media refers to the artificial creation or modification of media by artificial intelligence and machine learning. It is usually used to refer to deep fakes. However, deep fakes are only an example of synthetic media. Other examples of synthetic media include AI-written music, text generation, and voice synthesis.

Nudge

"A nudge (...) is any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid."[13]

[11] Frank Pasquale. **The Black Box Society: The Secret Algorithms That Control Money and Information**. Harvard University Press, 2015, p. 3.

[12] See, respectively: IBM Cloud Education. **Machine Learning** (15 July 2020). Available at: https://www.ibm.com/cloud/learn/machine-learning. Accessed 04 Nov 2021; McKinsey Analytics. **An Executive's Guide to AI** (2015). Available at: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/an-executives-guide-to-ai. Accessed: 04 Nov 2021.

[13] Richard H. Thaler, Cass R. Sunstein. **Nudge: Improving Decisions About Health, Wealth, and Happiness**. Apple Books edition, Yale University Press, 2014.

PART I:

CONSENSUS AND DISSENSUS

A. Interviews in Comparison

1.The definition of disinformation

Concerning the concept of disinformation, the representatives proposed definitions that, although subject to particularities, are also partially overlapping and in synergy with one another. The interviewees from civil society, the public sector and academia found consensus that disinformation encompasses two elements:

(i) it aims at misleading and/or manipulating its audience

(ii) with the subjective intent to do so

As identified by Xabier Lareo from the public sector, "disinformation is information which is crafted or presented in a way with the **intention** of **manipulating** people."

Vidushi Marda, speaking from a civil society standpoint, added that disinformation is "often spread by individuals in a **position of power**", thereby identifying it as an instrument of both power game and its maintenance. She also added that it is commonly employed to incite polarization between distinct societal groups, promoting the strategic advantage of certain groups, such as the prevalent ethnic or religious group in a country.

Both representatives from academia and the public sector acknowledged the complexity of creating a definition. Thereafter, the two provided partially-divergent conceptualizations. Although they agreed upon the above-mentioned elements of intentionality and misdirection, they divergently presented other elements to be considered when it comes to conceptualizing disinformation.

Gianclaudio Malgieri, from academia, described two elements that comprise manipulation. The first is referred by him as "**bad resources.**" That is, the source of information itself is biased and explicitly misleading. The second is referred as "**manipulation in the cognitive phase**", whereupon the online platform is used to manipulate and exploit the user's biases instead of the source's. According to his line of thought, only the first moment constitutes disinformation. In other words, disinformation is "giving explicitly misleading information, which is different from implicitly exploiting biases."

Likewise, the interviewee from the public sector recognized false and decontextualized information as disinformation. Xabier also identified disinformation as practices such as the decontextualization of true information; the unilateral promotion of content without contrary information against which the first could be contrasted; and nudging. Therefore, this conceptualization is broader in scope than that presented by the academia interviewee.

Natalia Neris from the private sector, nevertheless, introduced a definition that moves towards a **procedural approach**. This is neither identical nor contrary to the other sector's conceptualizations; in fact, there is a synergy between them. She asserts that disinformation and respective remedies should be analyzed in function of the platform's principles and purposes. For instance, where such a purpose is to enable public discourse, the principle of freedom of expression must be balanced against the safety of users. She argues that this entails the two following concerns.

The first one is that the private institution must entrust third parties - such as opinion leaders, experts, fact-checking agencies and journalists - to deliberate in the sphere of public discourse as to whether a certain piece of content is disinformation, instead of taking upon itself to unilaterally determine it. This seems to be reflective of a procedure that aims at preserving multistakeholderism and public discourse.

The second concern is that interference with public discourse is only acceptable where disinformation is proved to cause real-world harm. The harm caused by disinformation is illustrated by a decision to refuse the COVID-19 vaccine, damages to the integrity of civic procedures, and the dissemination of synthetic or manipulated media.

The synergy here lies in the belief that, as defended by the representatives of other sectors, **disinformation is misleading or false content**. However, truth or falsehood is not left to be solely determined by the private institution in question (Twitter), and the platform's guiding principles must only be interfered with where real-world harm exists as a consequence of disinformation corroborated as such. Accordingly, when disinformation poses real-world risks to safety it should be tackled.

2. The definition of algorithmic transparency

The interviewees from civil society, academia and the public sector argued that **algorithmic transparency entails, at the least, access to the logic of the software**. Gianclaudio Malgieri (academia) defended that this must be achieved in the form of "meaningful information about logic, (...) significance and effect." Xabier Lareo (public sector) phrased it as the ability to "understand and to explain the decision-making" of a system. From a general standpoint, Natália Neris (private sector) adopted a slightly different approach, by which solutions must be developed to enhance transparency. This, it seems, may or may not include the code itself, but the strategic decisions related to developing an algorithm in the first place and giving power to people to choose which algorithm organizes their timeline.

Vidushi Marda (civil society), nonetheless, expressed that, due to the complexity of algorithms and, in particular, machine learning, "having access to just the logic of an algorithm is almost never enough." Indeed, both civil society and the private sector assumed the context of machine learning and the Black Box Problem.

In fact, **all four sectors recognized, to different extents, that other factors should be encompassed**, or at least instrumental to what each interviewee considers to be Algorithmic Transparency. The concept must interact with the whole **socio-technical system** being employed by institutions and actors rather than just the technical aspects of the algorithm itself. Vidushi Marda (civil society) outlined, non-exhaustively, the necessity of also understanding the **context, assumptions, the data-points and features of an algorithmic process**. Moreover, she described Algorithmic Transparency as a first step to achieve accountability and understanding not only of the algorithm, but of the whole socio-technical system.

This goes beyond the algorithm's disclosure. As phrased by the public sector's representative Xabier Lareo, Algorithmic Transparency includes an explanation as to the "reasons and inputs that led to a certain output."

The **level of expertise of the subject** exposed to the usage and the **effects of the algorithm** is also taken into consideration. The interviewees from both academia and the public sector defended that the explanation's level of complexity should be adapted to meet the audience's ability to understand it. This would mean that information disclosed to an auditor and information disclosed to a regular user should be different, in order to make the information more accessible and appropriate to each audience according to their average knowledge on the subject. Gianclaudio Malgieri (academia) highlighted that transparency is, in fact, layered. It has "a general layer, an intermediate level and more specific levels of transparency", corroborating the idea that the target audience is important when considering what kind of information should be disclosed.

Considering the difficulty of conceptualizing Algorithmic Transparency, the academia representative, Gianclaudio Malgieri, argued that Algorithmic Transparency should encompass two questions: "transparency for whom?" and "transparency for what?" **There is no universal concept for transparency**, as it closely relates to the target audience. If an expert, then the explanation may be more technical. However, "if the audience is (...) the average consumer, in this trade-off between complexity and comprehensiveness, comprehensiveness should win." The disclosure of technical reports and the code itself is not accessible to regular users, who are not experts on the subject. Therefore, information should be made available in an understandable and accessible manner.

The interviewee from the private sector, in its turn, argued that fairness, responsibility and transparency are preserved where the institution's principles and purpose are followed. This is, herein, the promotion of public discourse in an open Internet. In her opinion, transparency entails two practices.

Firstly, enabling platform users to exert control over their algorithmic experience in the platform. That is, the user may choose the algorithmic process that determines how online content is presented. Either through moderation by a machine learning algorithm or in the chronological order it is published, without interference. The former's explanation is more complex, while the latter's may be easily understood. Thus, the user has the prerogative to choose the complexity level.

Secondly, the establishment of a diverse and multidisciplinary team specialized in machine learning ethics, transparency and responsibility. This is mandated with analyzing the platform's operation to safeguard data justice, equity, impartiality and fairness of outcomes. It aims at furthering (i) responsibility for algorithmic decisions, (ii) transparency as to its process, as well as (iii) agency and algorithmic choice. The team's findings may lead to the implementation of remedies, such as product changes, the development of new technical standards and the creation of new policies.

Another relevant aspect brought up by the private sector representative is the importance of considering if an algorithm is needed or adequate at all for certain activities. Other than that, equity and impartiality on the effects of the use of algorithms should be considered.

Considering all of the foregoing, **there is partial consensus among the interviewees regarding the concept of Algorithmic Transparency**. The representative of civil society was the one to expand the most on associated concepts and measures to ensure the objectives of Algorithmic Transparency. The interviewee from the private sector was also of the view that the concept must be understood to encompass the socio-technical context of the algorithmic activity. However, it differed as to the means of achieving transparency. Rather than focusing on the algorithm, it focused on enabling user control. Academia and the public sector delved into the need for transparency measures - such as explainability - to meet the user's ability to comprehend a technical subject.

Finally, some interviewees brought up how algorithmic transparency is often understood as opening the black box. However, they disagreed that this is an adequate approach, and defended algorithmic transparency as something way more complex.

It goes further than disclosing the code behind the algorithm. It is more related to actually explaining the reasons and inputs that led to a certain output. That should be understandable and accessible to every individual who is affected by these decisions, not only by experts or enthusiasts on this matter.

3. Algorithmic transparency and discrimination: whether and how both worlds meet

The representatives of civil society, academia and the public sector agreed that **disinformation and algorithmic transparency are related topics**. Vidushi Marda (civil society) explained that social media content is curated by algorithms and raised the question: "What is it that such algorithms optimize for?" The answer, thereafter provided, was that, currently, algorithms optimize for users to "spend as much time on the platform as possible." This, in turn, ends up promoting alarming and shocking content. According to the interviewee, "the line between algorithmic transparency and disinformation really has to come down to what is the business model on which all of these systems are being built and [if we] can scrutinize that business model against the standards of accountability and transparency that we want."

Thus, she defends that the key link between algorithmic transparency and disinformation lies in the objectives and implementation of the companies' business model. This must be scrutinized "against the standards of accountability and transparency." Only then, will we better understand the incentives that are built into the system and if the algorithm optimizes our feed for the best information or for individuals to spend as much time as possible on a particular platform.

The interviewee from academia followed a similar hypothesis. Gianclaudio Malgieri explained that if there is a filter in place that curates which piece of content will be seen by individuals, then we need to understand why each of us is being presented with specific search results or content. This is important not only to better understand the algorithms in place, but also for related decision-making to be more democratic as well as build consensus on how disinformation may be mitigated. **Only by being transparent is it possible to uphold or contest the mechanisms of moderation.**

Xabier Lareo, speaking from the public sector view, explained that both concepts are very closely related. According to him, in modern times, the tools for obtaining information partially shifted from acquaintances and traditional media outlets, such as newspapers and television, to online content. Indeed, "the tools that are supposed to help us not to be overwhelmed by the abundance of content (like search engines, recommender systems or automated content moderation systems) are governed by algorithms."

Lareo points out that the **issue surfaces when users are presented with a piece of content instead of another without understanding why**. Moreover, this becomes even more relevant due to the apparent neutrality of social media platforms and the way they curate content. In reality, people have the impression the content is neutrally moderated when, in fact, companies also have interests that may not always be aligned with that of the user.

The representative of the private sector, on the other hand, argued that disinformation and algorithmic transparency are not related topics and that focusing on disinformation is a narrow view of algorithmic transparency. This is because the responsible use of technology requires looking into many more areas of similar importance. It entails the study of the effects that it may cause throughout time in different areas, not only content moderation (which includes misinformation). When specifically discussing social media that allows individuals to access content generated by who they do not know or follow, Neris argued that the open nature of the platform allows all information, whether true or false to be discussed, counterposed and contested by everyone in an open dialogue.

4. Favorability to algorithmic transparency

The interviewees' answers did not present great disparities when questioned whether they favored or opposed algorithmic transparency. **They all favored more transparency in the process; however, they diverged as to the means by which transparency should be delivered to the public.**

The representative of the private sector highlighted mechanisms in place that should grant more transparency for the platform's users. For instance, the introduction of user empowerment mechanisms so that they could choose how to access the profiles they follow and how their posts would be presented to their own followers.

However, although certainly an important aspect of how users understand the platform, for the private sector the main long-term goal should be, primarily, enabling users with control over the algorithm they interact with. Once this goal is reached, users themselves should be able to decide on which algorithm to adopt among the different ones available on the market. According to Natália Neris, "[A]lgorithmic transparency is an important part of understanding how systems work. **We understand that the long-term goal should be to enable people to have control over the algorithms they interact with and, ultimately, lead to the ability to make our own choices between algorithms."**

The representative of academia brought a different perspective to the subject. Accordingly, transparency cannot be seen as a homogeneous phenomenon, where "one-size-fits-all" transparency is brought to the public eye. Instead, transparency should be adjustable to the context in which explanation is provided.

Therefore, transparency-enhancing measures should consider the complexity of the public affected by algorithms. Otherwise, transparency is unlikely to close the gap it aims to narrow. The interviewee suggested that, since social media platforms have access to and process so much data from their users, they could somehow use this data to target users with more adequate information regarding the functioning of the platform.

Moreover, he indicated **the public should participate in the development of algorithms that may affect them**. For greater transparency, the public should be empowered and provided access to information regarding the algorithm even before it is deployed. This approach is reflected by Article 10(4), of the European Commission's proposal for the Artificial Intelligence Act.[14]

Xabier Lareo (public sector) and Vidushi Marda (civil society) presented a similar approach. According to them, algorithmic transparency should not be viewed as only the disclosure of the platform's source code or the inner workings of a black box.

Lareo argued that the public disclosure of the source code is not always necessary or practical, and that there should be a balance between fundamental rights and intellectual property rights. However, the interviewee suggests that it is unreasonable to refuse from explaining how an algorithm works based on that argument, because the disclosure of information to supervisory authorities would not impact trade secrecy. Not only that, "even without the intervention of authorities, **citizens should have their right to some information that would allow them to know what they are being shown."**

Furthermore, algorithmic transparency should only be the prerequisite for the people affected by the algorithm to have access to the correct information. When people are informed, they should be able to petition the company to interact with them and influence their algorithms. About this, Xabier Lareo argued that "**algorithmic transparency is great, but it is only a prerequisite**; because then you need to be able to take some decision. There are many actors in the online content management that are so powerful dictating their own policies that being transparent is not enough. There should also be the possibility to tweak or to configure the way the content is being shown to you."

[14] Article 10(4): Training, validation and testing data sets shall take into account, to the extent required by the intended purpose, **the characteristics or elements that are particular to the specific geographical, behavioral or functional setting within which the high-risk AI system is intended to be used**. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206. Accessed: 08 Nov 2021.

Similarly, Vidushi Marda (civil society) identified that transparency alone cannot close the gap between algorithms and society. **Accountability is also needed.** According to both, transparency is the first step towards accountability of the whole socio-technical system; thereby including, not only the technology, but also its human operators, the institution behind it and the factors that may influence them. The interviewee argued that "transparency and accountability and, more importantly, (...) how these systems are being used by human actors is also important. So, when we think about transparency, accountability shouldn't just be that of the algorithm, but rather of the system. There is a socio-technical system that is being used by institutions and actors."

Marda also introduced concerns over how companies could employ this supposed transparency to exclude themselves from the debate. By simply disclosing the inner mechanisms of a problematic artificial intelligence system, institutions could wrongly argue to be in full compliance with algorithmic transparency. This, therefore, could serve as an argument to exempt them from taking any subsequent steps to halt deployment or correct a flawed algorithm. The interviewee stated that "it is actively dangerous, (...) if companies said 'well, we've (...) made the logic available to everyone.' Not everyone has the ability to understand that or make sense of it, right? So, if you're an engineer in Silicon Valley and you have a loan denied to you and you have access to an algorithm, then you can say 'well, I'm gonna audit this algorithm to prove that [you know] it's being discriminat[ory] for external reasons.' **But what if you're just not an engineer from Silicon Valley? You still shouldn't be subjected to this kind of arbitrary decision-making and the burden of getting a fair decision shouldn't fall on the individual.**"

This approach would lead to a shift in the burden of access to knowledge on how the system works. Instead of lying with the company, it would befall the user. Evidently, understanding an algorithm is of a much greater effort to a user than it is to a tech company. Moreover, a company is more equipped than users to remedy algorithmic flaws.

"My colleague Frederike Kaltheneur helped me understand this better - when I go to a restaurant, for instance, I will assume that I'm going to eat food that's not going to kill me. I'm assuming that the restaurant is responsible for ensuring that food reaches the table that is healthy for everyone who comes in. I think we need a similar kind of thought process for algorithmics. **The burden shouldn't be shifted onto the individual at all**", stated Vidushi Marda.

5. The ideal model for algorithmic transparency

Relevant aspects of ideal models for algorithmic transparency were brought forth during the interviews, as highlighted below:

a) **Access to algorithmic codes would not solve all problems**. Analysis of the business model behind the code and its prerogatives is also needed. Justifications as to the need for a system and its social impacts beforehand are more effective in preventing societal harm than a reactive audit after the system has already caused harm. This approach could also prevent the financial returns of an algorithm under scrutiny from being such a relevant factor in the decision-making process as to how to remedy its flaws.

Highlighting some of the manifestations, the representative of academia reinforces the necessity to understand a business model behind the code; the representative of civil society mentions the importance of justifying the need of a particular algorithm at the beginning of a project; the representative of the public sector stated each provider should be accountable for their risk analysis; and the representative of the private sector indicated that the so-called literal understanding of "algorithmic transparency" would be limited and tend to fail.

b) Information to be provided to a user in terms of algorithmic transparency should configure more than a mere and simple explanation, and the burden of proof does not belong to the individual who requests justification as to the moderation of content and related data processing.

This was an interesting reflection made by the academia representative when thinking about legal responsibility in this context.

c) All changes proposed in terms of ideal algorithmic transparency models bring some resistance, especially if profit is involved. However, the main point is the discussion concerning the different layers of transparency **regarding diverse types of situations.**

For the academia representative, trade secrecy should not be overestimated as it could not be claimed in light of the allegation of a fundamental right violation. The civil society representative argued that even if companies or governments are not satisfied with eventual changes, all discussions are relevant and could also lead to questioning some technological foundations that may turn an initiative unfeasible - so discussing them beforehand would prevent tools to be banned, for example. In his turn, the representative of the public sector mentioned that society would not resist algorithmic transparency, pointing out that maybe there would be some discussions about different required levels of transparency applicable.

Lastly, the representative of the private sector did not provide a direct consideration about this aspect. However, on a follow-up question regarding transparency mechanisms currently adopted by the platform, she mentioned a program called "Responsible Machine Learning" which, according to Natalia, comprehends a global initiative that aims at bringing higher transparency for the decisions related to algorithms and their respective processes.

This initiative, as stated by her, will bring higher transparency to the process of decision-making carried out by the company's algorithms, guarantee equality and impartiality to the effects of said decisions, and enable greater user autonomy over the company's platform through higher freedom of choice.

One example provided to illustrate this initiative was that, initially, when a user shared an image on their feed, an artificial intelligence system would crop the image to highlight its most important part for users scrolling down the platform's feed. However, the system showed a racial bias because when selecting the most important part of an image with multiple people of different ethnicities, it would most likely highlight people with clearer skin.

Considering the trade-offs between the speed and consistency of automated cropping against the potential risks the research revealed, the study carried out by the private institution concluded that not every piece of content on the platform was a good candidate for an algorithm. In this case, the decision on the cropping of images should actually be made by a human. Thus, the company started testing a new way to display full standard aspect ratio photos to give people more control over how their published images appear while improving how users experience seeing images on homepages.

6. Reading suggestions

Interviewees mentioned some reading suggestions, which are listed below.

- The book "**Desinformación**" (Pascual Serrano) explains the traditional disinformation techniques that are being boosted based on technologies.
- **WeVerify**, an open-source platform that facilitates collaborative and decentralized content verification, tracking, and debunking.
- **Maldita.es** is a Spanish fact-checking non-profit organization that provides tools and recommendations for fighting disinformation.
- **National Institute for Research in Computer Science and Automation** (INRIA).
- Marda, Vidushi. Shazeda, Ahmad. **Emotional Entanglement: China's emotion recognition market and its implications for human rights**. Available at https://www.article19.org/emotion-recognition-technology-report/
- Twitter's Blog, more specifically:
  - Chowdhury, Rumman. **Sharing learnings about our image cropping algorithm**. Available at https://blog.twitter.com/engineering/en_us/topics/insights/2021/sharing-learnings-about-our-image-cropping-algorithm.
  - Chowdhury, Rumman. Williams, Jutta. **Introducing our Responsible Machine Learning Initiative**. Available at https://blog.twitter.com/en_us/topics/company/2021/introducing-responsible-machine-learning-initiative.
  - Agrawal, Parag. Davis, Dantley. **Transparency around image cropping and changes to come**. Available at https://blog.twitter.com/en_us/topics/product/2020/transparency-image-cropping.
  - Roth, Yoel. Achuthan, Ashita. **Building rules in public: Our approach to synthetic & manipulated media**. Available at: https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.

- **Protecting the Open Internet: Regulatory principles for policy makers**. Available at: https://cdn.cms-twdigitalassets.com/content/dam/about-twitter/en/our-priorities/open-internet.pdf

B. Considerations regarding consensus/dissensus

After analyzing the answers given by the interviewees, both separately and in comparison to each other, the following considerations compiling what was presented previously should be made, while also adding some of LAPIN's thoughts and conclusions on the topic:

- The interviewees presented a consensus over the necessity of higher transparency on how algorithms work. However, participants mainly dissented as to how such transparency should be interpreted.

- The representative of academia defended that algorithmic transparency should be contextual and, thus, vary in accordance with the target audience, as opposed to a standardized response. Moreover, contextualization must be proactive, not reactive. It must be accounted for during development, as opposed to only after deployment. In support, he mentioned the European Commission's Artificial Intelligence Act. It proposes that affected groups should be considered during the development of a certain system.

- The representative of academia provided an interesting view regarding algorithmic transparency, by defending that it should not be viewed as a standardized response. It should be contextual and vary in accordance with the target audience. Not only that, contextualization should not be implemented only after the algorithm is deployed for public use; preferably it should be noted during its development. In support, the representative mentioned the European Commission's proposal Artificial Intelligence Act, which points out that affected groups should be considered during the development of a certain system.

- The representatives of the public sector and civil society provided that a strict interpretation of algorithmic transparency will never be enough to fill the public's gap of knowledge on the system.

- According to the public sector representative, a strict interpretation of algorithmic transparency would entail (i) insufficient adherence by the companies due to industrial secrecy concerns, and (ii) insufficient empowerment of the public. Therefore, transparency should be viewed rather as a first step towards a more ethical use of algorithms than a solution itself.

- The civil society representative brought forth a more critical approach to the subject. If not well-delineated, transparency could be invoked by companies to exclude themselves from the debate. It could be wrongly argued that algorithmic transparency is an end in itself and, therefore, the disclosure of the algorithm's inner mechanisms, without any subsequent remedies, would suffice.

- The representative of academia pointed out that algorithmic transparency should not be seen as an excuse by companies to not disclose information to their users. If a target audience is unable to comprehend the workings of a platform or the effects of an algorithm, due to excessive complexity, the burden should still lie on the company to clarify and communicate the information transparently and effectively.

- The academia representative also pondered that algorithmic accountability should not encompass programming and systems isolatedly. It should take into consideration the socio-technical context, the human-operators, the institution, the influences and intentions of each context. This is supported by the discussions on the ideal models for algorithmic transparency and by the view that the burden of proof should not belong to the user requesting justification on content moderation and data processing.

PART II:

TRANSCRIPTS

After the brief analysis proposed by our team, we are glad to share with you the full transcriptions of the four interviews commented on in the first section.

LAPIN is convinced that the more we hear from the interviewees, the more we would like to hear from them; so, this second part comes as a way of providing their full thoughts during the interviews, bearing in mind they have a lot more to say!

GIANCLAUDIO MALGIERI

EDHEC Business School

1. How would you define disinformation?

Well, it's a one-billion-dollar question. In my understanding, I try to conceive manipulation and fake news and all this kind of stuff. I think there are two separate moments. The first one is bad resources, which is disinformation, and the second one is manipulation in the cognitive phase. So, it's different from other forms of online platform manipulation because this bias is the material itself, it is the starting point, which is different from exploiting bias upon the individual, cognitive bias, etc. So, I think disinformation is, more simply, giving explicitly misleading information, which is different from implicitly exploiting biases. It's a very complex question.

2. How would you define algorithmic transparency?

There are different levels of transparency. First of all, I would define algorithmic transparency as a complex multi-layer exercise because transparency is actually based on several layers. We can have a general layer, an intermediate level, and more specific levels of transparency.

Your question needs, from my side, more questions: so, transparency for whom? and transparency for what? So, it's important to understand the target and the audience. For example, if the target is an expert, transparency can be more specific but less sober. Explanation can be more technical. If the audience is more "normal people", like an average consumer, in that case in this trade-off between complexity and comprehensiveness, comprehensiveness should win. Understandability should win to the disadvantage of technicality. So, it really depends on the audience, I guess. Of course, my bias is based on the European Union legislation, the GDPR and the new AI draft have a specific and very wide understanding of what transparency is. So meaningful information about logic, about significance and envisaged effect. This is the word in the GDPR. Or even the comprehensibility in article 14 from the proposed AI act. If you have more specific questions on this point, I am happy to specify.

3. Would you say disinformation and algorithmic transparency are related topics when it comes to digital platforms such as social media and search engines? How do you relate the topics?

I think they are because if there is a filter, like a content moderation algorithm, it is important to understand how it is in order to have a more democratic process in it and to have the consensus on how disinformation could be mitigated, because we can never solve the issue, but mitigate it.

So, in my view it's important that the specific techniques that search engines or social media platforms use to combat and mitigate disinformation are clearly explained. Transparency is something dynamic, it's not just understanding something, but also being able to contest, being able to understand what happens.

I think that content moderation is a lot about that. Facebook has tried this difficult challenge of creating a board for content moderation, a board of experts, and they use it a bit as a shield so that they could justify the decision to block some awkward accounts as the case of Trump. We know that yesterday he decided to sue Facebook, Instagram and Twitter for this.

Maybe it's important to be clear and to be transparent so that it's easier to contest. For example, something which is not exactly disinformation, but censorship and algorithmic transparency. [I will give] an example that happened to me. In Italy, the Ministry of Culture abolished the censorship on movies. I was happy and posted this news from an important journal on Facebook and Facebook deleted my post because the image taken from this important journal (I didn't choose the image; it was just an automatic preview of the image) was from a redline movie. It was something about censorship and Facebook deleted it because they said I was violating [the terms and conditions]. So, you could have a trap like that.

So, it's important to be extremely clear. I understand there might be some false negative and false positive in combating disinformation, but in order to contest false negative and false positive it's important to understand how you trained the algorithm that combats disinformation and on which bases you trained it. If you train child pornography or child-pedo-pornography of course you have some standards that are different from journalists that publish information about movies. So, context and logic.

4. How should transparency be conceived for facing disinformation? How should it ideally work for you?

It's not easy because usually the scholarly approach is to think about problems before, but I always engage with this, sometimes proposing solutions, etc. I can, for example, tell what I think was a good direction in some examples. For example, for Covid-19 both Facebook and Twitter had a lot of different mechanisms to avoid disinformation. For example, I think we all know, it's something quite well established now that every single post or photograph that is related to Covid there is a disclaimer under it saying "This is about Covid, click here to have more information." Something similar happened on Twitter when Trump was tweeting in the famous Capitol, let's say, "accident", or even before that, after the elections, when he was saying "the elections were fake" and etc., and Twitter said "this information is contestable. Click here to have more information."

So, I think it might be nice to have this form of transparency. We should separate two floors. One floor is "I give you more information to combat disinformation" and the other floor is "I am transparent about how I identify the risks of disinformation and how I attack them." So, for the first thing I think they are in the good direction. Every time there is a hot topic, I put a link to some official thing like the World Health Organization website when talking about Covid or an official election website when I talk about the presidential election.

For the second thing I think there's not so much on the table. One thing could be to add to this notice something about the algorithm. For example, not simply saying "We understand that this is Covid related. Stop", but there might be a link "How did we understand this is Covid related" and then, for example, I disclose a long list of keywords that are considered hot keywords. Another example of false positive just now coming to my mind, I'm sure you have even more examples, I have some friends that are TV journalists and, for example, they take a selfie when they are on TV and they talk about things like, now in Italy we have an important law discussed by the Parliament about LGBT, against homophobia. But then broadcasting news have headings underneath. In that case, the heading was about Covid, but it was talking about something else, the post was about something else. Facebook identified that post as Covid related and said "Look, this is Covid related, be sure what you share, click here". [This happens] every time that in the background there is something related to Covid but that is not the main thing. For example, there is a signal about vaccination, but I'm talking about something else, you know, I'm talking about my shirt, there is an automatic thing. I think that if they disclose the techniques, it's easier to contest. So, participation from the individual. I think that participation is a good thing. I have this paper with Michael Kominscjy about algorithmic impact assessment and in that paper we tried to say that if you put codecision in the loop, it's easier. You shouldn't consider codecision as something bad, but as participatory deciding. But this is opening the pandora box.

5. Do you think that systems should be explicable or interpretable equivalently for different kinds of groups such as auditors? Would you think the information concerning a system should be equivalent to users, regulators or people interested in understanding how it's working and should these groups have access to the same information?

This is very much related to the first thing I was saying that transparency is based on the target too. I see your point about groups. My doctoral thesis was about vulnerable people and my proposal was to have a contextualized approach to vulnerable people. You are not vulnerable per se, you are vulnerable in a context. The criticism that I got was "If everything is context and if everyone everywhere is vulnerable, then you never protect them. You have to compromise". I believe that it is important to consider groups, group privacy, all this kind of granularity, but considering subgroups. So not a big group of children, women, but some subgroups.

They know a lot of information about us. They could use these to target better transparency. If they can infer that I'm in a specific vulnerable condition, for example that I am a gay person younger than 20, I can be exposed to many risks (harassment, psychological vulnerability), then you should be explaining information to me. I think another critique could be "How can you do it for every group?". It's impossible. "How can you specify things?". One thing interesting is the AI Act because Article 10, paragraph 4 refers to contextualization of AI. So, they say AI should be tested under specific social and geographical places where it's meant to be used and this is something that we should keep in mind. The second thing is that we should look, in my view, at the main target of something. For example, if the main target of my advertisement is or can include children… Not just the intended target, but also the most impacted target. If I show some advertisements about violence, I should always consider there are vulnerable people victims of violence that could be more exposed. So, to them, the way I give information should be more granular, more sensible. It's not easy at all, I understand. How could we do that? Maybe with participatory decisions. When I design an algorithm, I should call the representatives of all possible impacted groups and ask them "Do you think that treating disinformation is clear enough? Do you think that we should give more? Do you think that information is not enough and we should just block this advertisement?" So, I think it's important to ex ante consider these groups, not just after, in a transparency exercise.

6. Which kind of information should be made available to users, regulators, auditors? How should they be made available?

Thank you for this point, because it's a key point. We don't need the code. Under French law, all algorithms used by public administration should be open-source code, and so the code should be clear and available. This doesn't solve all problems. [It is not enough] if we say "there's a black box and we just make the algorithm available so we have white box". It's not the code that is important. It's important also the business model behind the code, the possibility to interact with the code.

The business possibility. The managerial possibility. How flexible is the algorithm? How contestable is it in your business model? It's not just the code itself. So which information should we disclose? More than mere explanation, I would say that information that should be disclosed is justification. I tried to propose this now. I have an ongoing paper with Frank Pasquale about this justification of algorithms that should be used in my view as a clear report based on an impact assessment in which I explain why and how my algorithm is not discriminatory, is not manipulative, is not unfair, is not inaccurate, etc. It's like putting the burden of proof to the developer and making the algorithm unlawful by default, and they have to flip the proof and say "yes, it's lawful" and explain why and how, because I can relate to statistics related to my algorithm. I can relate to the methods in developing it. I can explain to you why it was not unfair, why we don't underestimate minorities, etc. So, I think maybe I'm a bit biased because of the GDPR. I'm a GDPR guy, and the GDPR is based on data protection impact assessment, but I think that is an important starting point. It has everything. It has the explanation of the algorithm itself, but also something more. How does it interact with fundamental rights? What do I do to prevent this interaction, this impact on fundamental rights? If this is effective. How do I assess the risk? It's not a solution for all problems, but of course it might be much more helpful than having the code. I've seen the code of many tracing algorithms for Covid, but it doesn't change anything to me. What is really important is to understand how the business model is built, the risks that are considered, etc.- to me, of course.

7. Would you say the model you proposed in question #6 faces social or sectoral resistance?

This is a very important point. I've been working on trade secrets in my very first years and I think that trade secrets are a "dangerous monster", but we should not overestimate the coverage of trade secrets. Because I think that an impact assessment report can well respect trade secrets also. You should explain why. The developer should explain in which part of the report there is a violation of trade secrets because we don't have a definition of trade secrets. It's a secret information (so secret to the general audience of people) that is meant to be secret with affirmative steps to keep it secret and of commercial value. This definition, for me, is not comprehensive to every part of the report of an impact assessment. I can understand there are some key points. Also, one thing to keep in mind: trade secret laws. We know there is an international one, the TRIPs agreement (the Trade-Related Aspects of Intellectual Property Rights Agreement of the World Trade Organization). And this is quite a general definition. In Europe we have a Trade Secret Directive in 2016 approved at the same time as the GDPR, just one month before, I guess. But the problem is that there are some exemptions. If the trade secret covers a violation of fundamental rights, interest of workers, etc., this trade secret can be "violated" without any legal detriment to the perpetrator. We don't have that kind of problem because fundamental rights violations cannot be covered by trade secrets. So, you can never use trade secrets to cover a violation of fundamental rights. For example, "my business model discriminates black people". In that case, if you are a whistleblower, you can. I am just being careful about that. Of course, the US can have a different view of trade secrets, I think more economic based than the European one. But at least in Europe I wouldn't be too worried about trade secrets

8. Would you like to reference any material (academic or not) that you rely on when you think about matters mentioned during this interview?

You're asking about materials that relate to disinformation and algorithmic transparency at once, right? I have to think about that. I think a good point of reference in Europe is INRIA, Claude Castelluccia, they are doing great things about both disinformation and about algorithms, so I think they might have something nice.

NATÁLIA NERIS

Twitter

1. How would you define disinformation?

I think that for starters, it's worth talking a little about the purpose of Twitter. Twitter's purpose is to serve the public conversation, which is why our approach to disinformation, as well as the rules that we have the service, is based on a balance between two fundamental values: freedom of expression and the safety of the people who use the platform. In this way, I think it's important to give a context of how we address the issue of misinformation on the platform. For us, the focus is not on determining whether information is true or false. We understand that this type of analysis of this type of discussion is the responsibility of other actors, opinion makers, specialists and journalists checking agencies, people with different perspectives. Our focus is to stop the potential harm and act on content that can deceive people and bring harmful effects to them, influencing the decisions they make in the offline world. So, for example, a person who saw a tweet with misleading information or questionable about the effectiveness of the vaccine on Covid-19 and because of this she made the decision not to take the vaccine and ended up being affected in her life by something she saw in our service. For this, for cases like this, we created rules to give context or to remove from the platform this type of content, which from something demonstrably deceptive and questionable can cause real harm to people. So, our three policies on this approach deal with misleading information about Covid-19, about the integrity of civic proceedings, and synthetic or manipulated media content. So, our policy is much less about a simple true-or-false look and much more about the potential harm that content can cause.

2. How would you define algorithmic transparency?

Ensuring that Twitter meets all public conversation requirements makes the machine learning we employ fairer, more accountable and more transparent. If you will allow me, I would like to point out two areas of work and actions that we have taken in this regard. The first is that, in line with our defense of what we understand as an open internet, we guarantee people the power to choose how they will see their tweets, how they are presented on their homepage. They can choose to view in reverse chronological order as they are published - that is, without any filtering or interference; or they can see the most relevant tweets first, when we use our machine learning to select and prioritize content that might interest each person the most.

That said, I would like to bring up the second point: we have on Twitter what we call "Responsible Machine Learning", a global initiative launched in April that aims to make decisions related to algorithms and the process in which they were made more transparent, as well as ensure fairness and impartiality in the effects of its use, and enable people who use the platform to have more control and choice. As a result of the work of a diverse and multidisciplinary team in the areas of Ethics, Transparency and Accountability of Machine Learning, changes in the product may arise, such as the removal of an algorithm or the already announced possibility of greater control over tweeted images), or even new standards for the development and creation of policies that have a material impact on specific communities.

I could say that the pillars of this Responsible Machine Learning initiative consist of the following: take responsibility for our actions and our algorithmic decisions; equity and fairness of data; transparency about our decisions and how we arrived at them; enable agency and algorithmic choice.

3. Would you say disinformation and algorithmic transparency are related topics when comes to digital platforms such as social media and search engines?

First, it's important to remember that different platforms have different natures and very different workings. Twitter, being open, allows people to access tweets and conversations of people they don't know or follow; moreover, as I mentioned, it is always possible to choose the organization of the homepage in reverse chronological order, without filters or interferences, or in the same way with algorithmic choices. This nature of the platform allows untrue information to be viewed and discussed, opposed or contested in real-time by the community that uses our service, even if this dynamic occurs between accounts that do not follow - that is, independently and beyond the content that may come to be presented to a person from algorithms. Bearing this differential in mind, I would not say that disinformation and algorithmic transparency are related topics, but that focusing only on disinformation is a very narrow view of the importance of algorithmic transparency. The responsible use of technology includes the study of the effects it can have over time and in different areas, and not just in content moderation (which is the activity that addresses the issue of combating misinformation). Our Responsible Machine Learning workgroup is made up of people from different areas of the company, including technical, research, safety and product teams. Leading this work is Machine Learning's Ethics, Transparency and Accountability team, which is a dedicated group of engineers, researchers and data scientists who assess current or future unintended damage to the algorithms we use and help Twitter prioritize which issues will be fought first. Among the working topics of this team are mainly: analysis of racial and gender bias in our image cropping algorithm (bump); equity assessment of our home page recommendations for all racial subgroups; and weighting of content recommendations for different political ideologies in selected countries. We understand that responsible machine learning is a long journey and we are taking the first steps. We want to explore this path with a spirit of openness and the goal of making a positive contribution to the field of technology ethics.

4. Are there algorithmic transparency mechanisms implemented on Twitter?

As I pointed out, we recently announced Responsible Machine Learning, which is a global company initiative that aims to make decisions related to algorithms and the process in which they were made more transparent, ensuring fairness and impartiality in the effects and enabling people to use the platform in a more autonomous way, that they have more control and choice. As a result of the work of a very diverse team, among some decisions we took was to announce the possibility of greater control over tweeted images. In the specific case of image cropping, we conducted quantitative and qualitative research and our results led us to the following conclusion: even though the salience algorithm was adjusted to reflect the perfect equality between the race and gender subgroups, we are concerned about the harm representation of the automated algorithm when people are not allowed to represent themselves the way they want on the platform. The overhang also contains other potential harms beyond the scope of this analysis, including insensitivity to cultural nuances. We considered the trade-offs between the speed and consistency of automated cutting with the potential risks we saw in this research. One of our conclusions is that not everything on Twitter is a good candidate for an algorithm, and in this case, how to crop an image the best decision should actually be made by people. In March, we started testing a new way to display full standard aspect ratio photos on both iOS and Android - that is, without clipping the bump algorithm. The goal was to give people more control over how their images appear while improving their experience of how they see the images on their homepage. We received very positive feedback about this experience and we ended up releasing this feature to everyone. This update also includes an actual image preview in the tweet composer field, so tweet authors know what their tweets will look like before posting them. This release reduces our dependence on machine learning to a function that we agree is best performed by the people who use our products.

5. Are you in favor of algorithmic transparency?

Yes, our recent actions, both the homepage and the image cropping one, reveal this position. However, we understand that algorithmic transparency is an important part of understanding how systems work. We understand that the long-term goal should be to enable people to have control over the algorithms they interact with and, ultimately, lead to the ability to make our own choices between algorithms.

6. How should it work ideally for you?

This is a great question and gives me the possibility to complement the ideas that I presented in your previous question. Much of the policy debate around technology and algorithms has focused on solutions that involve the disclosure of source code for algorithms, a literal interpretation of the phrase "algorithmic transparency". While in a limited context this can provide insights to a small and highly technical audience, it is a limited and flawed approach to dealing with the broader societal challenges that technology services intersect. In short, we think that transparency is not an end in itself. Policy makers must focus on increasing choice, control and competition in this area to shift the balance of autonomy from centralized proprietary systems to open, enabling and decentralized systems.

7. Would you say that the model you proposed (in question #6) faces social or sectoral resistance?

We believe that most debates about content on platforms fall into the trap of focusing on the question of whether content should be removed or not, when in reality content moderation should allow for a series of interventions while establishing clear definitions for the content types. We understand, and it's one of the things we've been working a lot on, that the open internet favors discussions that focus on the latter, and not the former.

8. Would you like to refer to any support material (academic or not)?

This recent Responsible Machine Learning initiative that I mentioned is premised on the internal and external sharing of our learning and best practices to improve the collective understanding of the industry in relation to the topic, and also help in our approach and make us responsible. All the information that we have produced, the research carried out by peers, the trends… data, detailed findings, in general we have been posting on Twitter Blog. So, I think my main recommendation would be to follow this work and these updates on these fronts. In addition, the opinion of a wider audience is very valuable to us, so I would also recommend that, if people who are interested in the topic have any specific questions regarding this work, that they tweet, we have been following all the tweets with questions about it, as long as they follow a #AskTwitterMETA hashtag. We are always monitoring and listening to what our community says so we can respond.

CIVIL SOCIETY

VIDUSHI MARDA

ARTICLE 19

1. How would you define disinformation?

I think disinformation is distinct in a couple of ways. It's information that is intended to mislead individuals. It comes from, like I said, an intentional place. It's not like a mistake. It's not a rumor. That was started off as something and then it blows itself out of proportion. I think there's an intentionality behind it. It's often also issued by individuals in a position of power. That can be something as obvious as political power. It can also be a faction that is a religious majority in a particular country or ethnic majority, if that's even a thing in any form, and it's usually strategically placed in order to disadvantage those that the powers considered the other side. So, it's a very, very intentional piece of information in the phenomenon of information.

2. How would you define algorithmic transparency?

I'm actually very critical of the idea about algorithmic transparency, just because I think the field of machine learning, of course it's existed for like 50-60 years now, but it really gained momentum in the last, I would say, five to six years, and for a long time we talked about transparency as just looking inside the black box and to say "show us how you make this decision, give us the model and give us the data and everything is fine". But that's not actually true because algorithmic transparency is a lot more.

Algorithms and machine learning algorithms, in particular, are much more complex than just that. Having access to just the logic of an algorithm is almost never enough, and it doesn't have any meaningful way of adding to the conversation, right? Even if you're just an expert social science researcher, an algorithm's logic makes no sense to us because we have no context of what were the assumptions and what were the data points and what were the features at the time.

So, when I think of algorithmic transparency, I almost think that it's almost never enough, and what I'm looking forward or looking for instead of transparency is actually accountability, because transparency is the first step. It's not a meaningful end even off itself, it's the first step towards accountability, and I think words like algorithmic scrutability or algorithmic explainability tend to mean more than just transparency.

And the second thing honestly, with respect about algorithmic transparency is that, if we think of machine learning systems, we're not just thinking of the algorithms, we're also thinking of the human in the institutions that use these algorithms and that are using these systems, and so I think transparency and accountability and, more importantly, how these systems are being used by human actors is also important.

So, when we think about transparency, accountability shouldn't just be that of the algorithm, but rather of the system. There is a socio-technical system that is being used by institutions and actors.

3. Would you say disinformation and this accountability of the systems are related topics when it comes to digital platforms? Would you talk about this relation about disinformation and accountability?

If we just take a step back and just look at social media platforms, we're fed information that is determined by an algorithm, right? I mean, that is the first thing that you said. And when we think of the algorithm, you have to ask what it is optimizing for, right? So, is it optimizing for the best information? Is it optimizing for your information diet or is it just optimizing so that you spend as much time as possible on a particular platform? And I think it's the last answer.

Algorithms that curate our social media, curate our news feeds now are optimized for us to spend as much time on the platform as possible. And because humans are slightly problematic just by nature of being human, the algorithm ends optimizing for alarming content, it optimizes for shocking content, it optimizes for often unpleasant content just because that is what it appeals to us and what makes humans more interested in something.

So, I think the line between algorithmic transparency and disinformation really has to come down to what is the business model on which all of these systems are being built and we can scrutinize that business model against the standards of accountability and transparency that we want.

For instance, if I were to be given in a newsfeed algorithmic, it (algorithmic transparency) would make no difference and it would lead me no way closer to accountability at what I am today because I can't make sense of such a complex system. But if we think about accountability of the social technical, then we would also have to look at what is Facebook optimizing for, what are the incentives that engineers are being told to optimize for when they're building these systems, how does this information gets flagged and identified, if at all. I think it's an indirect but extremely important link between the two.

4. You said you were critical about calling this transparency, but do you think that accountability term can encompass all of this?

When I think of the lifecycle of an algorithmic system, there is so much that goes on before this instance exists. So, you have the stage in which an algorithmic system is conceptualized. Then you design it, then you develop it, then you standardize it against some sort of metrics, and then you deploy it.

So, when I think of algorithmic transparency, unfortunately, it seems confined to one stage or the other right? Especially in the classic way that we're talking about algorithmic transparency. What we need to stop thinking about is accountability and scrutability throughout the lifecycle. Some conceptions, like "why do you want this system to exist?; who was asked for this system to exist; what are the incentive structures that led to this conversation?; (when we talk about facial recognition) which government wanted it?; which company is trying to sell it?; (from that point) why are we (the company) going to build it?; How we are going to build it?; how are we are going to design it?; what data are we going to use?; how are we going to test it?; how are we going to audit it?"

With all of that put together, and also testing it against and in context of the humans and institutions that use these systems. For facial recognition, you can't think about facial recognition accountability for law enforcement without thinking about law enforcement accountability. And so, for me, it has to be extremely holistic to mean anything at all, because I think it's easy for companies, for instance, to say "well, we've audited the algorithm. The facial recognition algorithm is 99% accurate. It is great", but what if the law enforcement institution (that uses the system) is extremely problematic? That makes the use of this system far more complex than we otherwise realize.

5. Do you think that systems should be explicable or interpretable equivalently for different kinds of groups such as auditors? Would you think the information concerning a system should be equivalent to users, regulators or people interested in understanding how it's working and should these groups have access to the same information?

Again, I think this is one of those instances where, if I said yes, make the logic of the algorithm and all of the reasoning of the algorithm available to everyone equally, that would be incomplete, right? And it would also be actively dangerous.

It would be incomplete because when we have the logic of the system, we have to clearly identify what we mean by that, so I don't just want the data. I don't just want the algorithm, I want the models, I want the features, I want the incentive documents, I want the design process, I want to be in the room when you're designing these things, that scale complete.

It is actively dangerous, also because if companies said Well, we've done a job, we made this really problematic AI system, but we've made the logic available to everyone. Not everyone has the ability to understand that or make sense of it, right? So, if you're an engineer in Silicon Valley and you have a loan denied to you and you have access to an algorithm, then you can say "I'm gonna audit this algorithm to prove that it's being discriminating for external reasons". But what if you're just not an engineer from Silicon Valley? You still shouldn't be subjected to this kind of like arbitrary decision-making and the burden of getting a fair decision shouldn't fall on the individual.

So, when I go to a restaurant, for instance, I assume that I'm going to eat food that's not going to kill me. I'm assuming that the restaurant is responsible for ensuring that food reaches the table that is healthy for everyone who comes in. I think we need a similar kind of thought process for algorithmics. The burden shouldn't be shifted onto the individual at all.

6. If you were able to build a model, an ideal model, how would it work regarding algorithmic transparency or accountability of the system?

I think for me the important bit would be, first justifying the need for a particular algorithm. I think a lot of times because machine learning is cheaper to build now, it's just built and then we're asking, "well, do we even need this in the first place?" So first it would start with really justifying the premise and the need for these systems.

Thinking about how this particular type of system could benefit, or would actually solve the problem and then it would involve a careful design process that takes into account not just economic incentives of efficiency and scale and things like that, but also be justified against existing legal and regulatory standards.

The thing about auditing is that, if we could audit systems against every problematic outcome: A) we would never have an algorithm that ever made a mistake; and B) It will be easy to gain the algorithm as well, if people knew exactly what to look for. The problem with auditing is that it assumes a certain level of predictability of algorithms which just does not exist, especially when it comes to really complex algorithms.

So instead of saying just auditing, because I don't want to give you a very high-level answer regarding your project, I think testing it in the real world rigorously against all of the standards that it would be subjected to otherwise before deployment is important because often what happens is that you have really problematic systems that deployed widespread and then you start auditing them. Then, no one wants to take them off, right? No one wants to stop using them now because they're used to it.

So that all of this has to happen before you deploy a system. And, once you deploy a system, it has to have a rigorous kind of check on outcomes and impacts on society in general.

7. Do you think this model you just suggested or proposed would face social resistance or sectoral resistance?

I don't think companies would be happy about it. I don't even think governments would be happy about that. But the problem precisely with the field of AI is that governments and companies are the only happy entities as long as they're able to buy and sell these technologies and use them as forms of control.

It seems crazy to think about whether you even need a system like facial recognition because the draw of facial recognition is to be able to sell it and then to be able to wield power through it. So, I don't think it's necessarily an easy way, but I think in the long run, if you think for facial recognition used the US, for instance in Silicon Valley, companies are so quick to make it, sell it and deploy it, and then that there have been these bans and then you see that it doesn't work and then, you know, you're suddenly questioning foundations on which on the basis of which you actually deploy these technologies.

So, the incentive for me would be: wouldn't you want to save money and time and make sure that using these technologies is actually something you can do for the long term?

But again, that's a bit naive. I recognize that, and I think that I'm constantly thinking of ways we can have that conversation and incentivize people in positions of power to take note of those kinds of issues. One day we will see.

8. Would you share any experience you're having in India regarding this topic?

It's not on transparency, but I guess you could say it's all part of it. I actually have an example from China because in, January this year, along with a PhD student at Berkeley, Shazeda Ahmed, I published a paper on emotion recognition market in China. And, what we found was that actors in positions of power, the Chinese Government (at the national and municipal levels), companies and academia were working together to push for demand of emotion recognition technologies. And the question there became: this technology is based on super pseudo-scientific foundations - the foundations and assumptions on which you build this technology have been refuted from the time that they have been around. So how can you have a billion-dollar industry just based on this?

The realization that we came through the research, which was that these technologies exist because actors in position of power enjoy the idea of having and working with these technologies. So, the question of algorithmic transparency for me has also to be around incentives on the political economy of technologies. It can't just be off the system, because often the dangers or the benefits of technologies are determined by individuals and institutions that use them.

I think the second one is in India, where they want to build a centralized facial recognition database that will be an automated face recognition system (AFRS), which would be centralized nationwide. Police stations could exchange information in real time, using pictures from anywhere procured in any way.

We can think about algorithmic transparency to say "facial recognition doesn't work. The only system that you have been using is like 1% accurate, so why would you use that?" Tell them the problems with these systems and things like that. But I think that the bigger and the actual problem there is that there's no transparency as to who is going to get these tenders? Who gets the money? Who is asking for this to be built? On what legal basis? There is complete opacity of the process, complete opacity of the system itself.

By February 2020, the government said that they had picked out like a thousand men from peaceful protests, and they were going to teach them a lesson or whatever. So, there is this complete opacity across the process.

Expanding the view from just a technical system to a political socio-technical-economic-financial reality of the system is something that I've been working on. And I didn't realize that the same thing is happening in Brazil. I think there's a lot of parallels actually between India and Brazil in terms of the fascination for biometric technologies in general.

XABIER LAREO

European Data Protection Supervisor

1. How would you define disinformation?

I will give my definition, because I have seen different studies and different people give different names to similar stuff, so I would say that disinformation is information which is crafted or presented in a way with the intention of manipulating people. So, for nudging people for some reason. Different reasons for different actors.

I would say this is usually tried to be achieved by presenting falsehoods as facts; by decontextualizing information, so you present a fact which is perfectly true but you present it in a different situation. To give an example, it would be like presenting a photo of some country or some place as if it was taken today when it was taken three years ago.

And also promoting certain pieces of information which are biased or falsehoods, promoting disinformation instead of other information that could be contrasted with the first one that would allow people to decide by themselves which person they want to trust.

2. How would you define algorithmic transparency?

I would say that algorithmic transparency is the capacity to be able to understand and to explain the decision-making, the result of the decision-making of an IT system; I will focus on IT systems even if the scope of this could reach perfectly other systems because I understand that is the focus of this conversation. For me, it's about being able to explain not so much each and every step of how you did end up having an automated decision, but on which are the reasons and which are the inputs that led to a certain output so a human can understand - and by a human we can discuss which type of human but I would say that when it comes to the type of automated decision-making systems that deals with misinformation. It should be any human.

So, if I understand that with a certain search result I should be able to understand even if I was not a computer engineer, that should be the case.

3. Would you say disinformation and algorithmic transparency are related topics when it comes to digital platforms such as social media and search engines? How do you relate the topics?

Yes, very much; they are very closely related. The way I see it: internet has an overwhelming availability of information and people are turning more and more into online content for informing themselves. The pandemic has only accelerated this trend. Before, people were informing themselves by speaking to other people, by reading the traditional media like newspapers or by watching TV, so the information was somehow controlled but also somehow safeguarded by certain regulations.

This shift to online content as primary means for informing individuals or citizens has the side effect that the tools that are supposed to help us not to be overwhelmed by the abundance of content (like search engines or recommender systems or automated content moderation systems) are governed by algorithms.

So, the problem here comes when you don't really understand why you are being presented certain search results instead of some others or why something one of your friends has published in a social media is not appearing on the top of your (social media) feed but is appearing somewhere very, very low. When you ask why is that, you have no information.

The bigger problem is that most of people do not know about any of this - they really believe that what they are presented is a neutral narrative when it is not, because all these actors (social media, search engines, etc.) have their own interest (like everybody does) and what we are presented with are information and content tailored to fulfil their interest, which sometimes constitutes ours and sometimes does not.

4. Are you favorable to Algorithmic Transparency?

Yes, I am very favorable to algorithmic transparency but I will make a couple of additions to this first answer because we should not get it wrong. Algorithmic transparency does not necessarily mean public disclosure of the source code - I do not think it is always necessary or practical. Most people will not know what to do with it. I understand that there is a need to protect the intellectual property of the company that invested lots of money and resources into delivering this. But when it comes to fundamental rights, and disinformation is very much related to fundamental rights, you need to balance. You cannot just say "no, I am not explaining you how my search engine algorithm works" or "I am not explaining you how my recommender system works because if I do it then I will lose a lot of money" because for that there are supervisory authorities. Even without the intervention of authorities, citizens should have their right to some information that would allow them to know what are they being shown.

Algorithmic transparency is great, but it is only a prerequisite, because then you need to be able to take some decisions. There are many actors in online content management that are so powerful dictating their own policies that being transparent is not enough. There should also be the possibility to tweak or to configure the way the content is being shown to you. In that respect, I think that the Digital Services Act and the AI Act in the European Union are giving some steps and some provisions that are about transparency of certain AI systems and recommender systems. Of course there is a harder work to be done, I would say, but at least some of these transparency requirements are taken into account in these acts. A little bit of advertisement: we have done also a couple of opinions on this to pieces of legislation which are available online.

5. How should algorithmic transparency ideally work for you?

I have never done this myself because I am not in the industry, I work for a public authority. But I would say that, first of all, you need to make an assessment since the inception of this automated decision-making system, trying to find out which are the risks for the intended users, for the people that are going to be subject to these decisions. And then, of course, there will always be some level of transparency needed, but the level of transparency will be very different depending on the kind of system. I think that misinformation is becoming such a big problem for our agency, for democracy, that I think that a high level of transparency will be needed; but I understand that each and every service provider should do their own analysis and they should be accountable for their analysis. If the supervisory authority asks "why do you think providing these details is enough?" and then they will explain their reasons, that is perfect - but they need to do this analysis. I would say that sampling, very much in the way advertisers do with ads [which are] shown to normal people, in order to question if people liked that or got the message would be a good thing - you give the explanations about what you are providing or the tool you are using to inform people because this should not be about having an extra barrier to an already complex and lengthy privacy policy. This should be about knowing that when you're presented some search results, there's a criterion tailored to you as an individual even if your name is not involved (or even as part of a group - because maybe you are super happy about being part of that group and having tailored results or maybe that is not the case). You should do some sampling because it's not so difficult to get some feedback from regular users and get to know if what you think is understandable would be in fact or not.

6. Would you say the model you proposed above faces social or sectoral resistance?

On the sectoral resistance, I would say it is clearly the case because every time that you are trying to change the way a business model is working and it's making lots of profit, a lot of people will be upset about that. "If this is working nicely, I don't want to make any change", because changes mean: a) money; and b) that I am doing something that is not as ok as it was supposed to be - because if it wasn't for you, I would not be asked to change the way it works.

On the social scale, I don't think that requiring more algorithmic transparency would be something that will be socially resisted - maybe there would be some discussions about different required levels of transparency. My main concern about this is not making the same mistakes we did with the Cookie Regulation. We do not want another cookie banner for automated decision-making systems and algorithmic transparency.

I mean, I - at least - want something that is more practical and that people really understand, so they do not have just to avoid reading whatever. On the other hand, in the long run, we need critical thinking, because confirmation bias is probably the resistance we would need to fight back against. People tend to be super happy when they receive things that confirm their actual beliefs and they attempt to try to avoid other information (other contents). That's not something that could be solved by algorithmic transparency, but, at least, you would be more aware that you are receiving certain content because you are labeled as "whatever" and maybe if you change the order or the filtering you would find some other things. As I said before, it is more a prerequisite.

7. Would you like to reference any material (academic or not) that you rely on or support you when thinking about the matters you mentioned during this interview?

- **Desinformación**, an excellent book by Pascual Serrano. Editorial Peninsula. Explains very clearly and with plenty of examples, the traditional disinformation techniques that are being boosted now using some technologies. I am afraid this book is only available in Spanish. You can find it in online bookshops.

- The EU project **Horizon 2020** funds many projects to help fight disinformation. Among them, WeVerify, an open-source platform that facilitates collaborative and decentralized content verification, tracking, and debunking.

- **Maldita.es** an Spanish fact-checking non-profit organization. Maldita also provides tools, such as browser extensions that warn users about websites with trustworthiness issues or including already debunked news.

Mentioned regulations:

- Art. 24 (online advertising transparency), art. 29 (recommender systems), art. 30 (additional online advertising transparency) in the **Digital Services Act.** The EDPS opinion on this proposal is here: https://edps.europa.eu/system/files/2021-02/21-02-10-opinion_on_digital_services_act_en.pdf.

- Art. 52 (transparency obligations for certain AI systems) in the proposal for the **AI Act.** The EDPB-EDPS joint opinion on this proposal is here: https://edps.europa.eu/system/files/2021-06/2021-06-18-edpb-edps_joint_opinion_ai_regulation_en.pdf.

LAPIN

LABORATORY OF PUBLIC POLICY AND INTERNET