IN-DEPTH ANALYSIS
Requested by the IMCO committee

European Parliament

Artificial Intelligence (AI): new developments and innovations applied to e-commerce

Challenges to the functioning of the Internal Market

Policy Department for Economic, Scientific and Quality of Life Policies
Directorate-General for Internal Policies
Authors: Dino PEDRESCHI, Ioanna MILIOU
PE 648.791 – May 2020

EN

Artificial Intelligence (AI): new developments and innovations applied to e-commerce

Challenges to the functioning of the Internal Market

Abstract

This in-depth analysis discusses the opportunities and challenges brought by the recent and the foreseeable developments of Artificial Intelligence into online platforms and marketplaces. The paper advocates the importance to support **trustworthy, explainable AI** (in order to fight discrimination and manipulation, and empower citizens), and **societal-aware AI** (in order to fight polarisation, monopolistic concentration and excessive inequality, and pursue diversity and openness).

This document was provided by the Policy Department for Economic, Scientific and Quality of Life Policies at the request of the committee on the Internal Market and Consumer Protection (IMCO).

This document was requested by the European Parliament's committee on the Internal Market and Consumer Protection.

AUTHORS
Dino PEDRESCHI, University of Pisa, Italy
Ioanna MILIOU, University of Pisa, Italy

ADMINISTRATORS RESPONSIBLE
Mariusz MACIEJEWSKI
Christina RATCLIFF

EDITORIAL ASSISTANT
Roberto BIANCHINI

LINGUISTIC VERSIONS
Original: EN

ABOUT THE EDITOR
Policy departments provide in-house and external expertise to support EP committees and other parliamentary bodies in shaping legislation and exercising democratic scrutiny over EU internal policies.

To contact the Policy Department or to subscribe for updates, please write to:
Policy Department for Economic, Scientific and Quality of Life Policies
European Parliament
L-2929 - Luxembourg
Email: Poldep-Economy-Science@ep.europa.eu

Manuscript completed: May 2020
Date of publication: May 2020

This document is available on the internet at:
http://www.europarl.europa.eu/supporting-analyses

For citation purposes, the study should be referenced as: Pedreschi, D., *Artificial Intelligence (AI): new developments and innovations applied to e-commerce*, Study for the committee on the Internal Market and Consumer Protection, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, Luxembourg, 2020.

CONTENTS

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
BBX	Black-Box Explanation
DL	Deep Learning
EU	European Union
GDPR	General Data Protection Regulation
GPS	Global Positioning System
H2020	Horizon 2020
ML	Machine Learning
US	United States
XBD	Explanation By-Design

LIST OF FIGURES

EXECUTIVE SUMMARY

Background

Artificial Intelligence (AI) has progressed to the point where it is an essential component in nearly all sectors of today's modern economy, with a significant impact on our private, social, and political lives. **AI** refers to an ecosystem of models and technologies for **perception**, **reasoning**, **interaction** and **learning**. The return of AI to the limelight in the recent years is mainly related to learning from data, **Machine Learning (ML)**, that made a jump ahead thanks to the emergence of **Big Data**. The mix is threefold: data reach the critical mass of examples to learn from, algorithms discover predictive patterns and patterns hidden in the data, the high-performance architectures succeed to compensate for the computing and storage resources needed. Based on this mix, AI managed to successfully tackle **long-standing open challenges**, such as understanding and translation of texts and speech, recognition of the content of images and video, and other tasks believed to require intelligence. Approximately ten years ago, it was noticed that some long-known learning models, hitherto ineffective to the mentioned tasks, if properly generalized and trained on a large set of example data, can make a qualitative leap. Indeed, such methods can learn, from the pixels of the images or from the words of the example texts, the general "concepts" that allow to recognize and classify accurately new images and new texts.

The current stage of development of AI exhibits strengths and weaknesses. On the one hand, the learning capacity of AI models is growing, bringing extraordinary progress in robotic vision and autonomous driving, in automated text and speech translation, in medical diagnosis, in risk assessment, in predictive maintenance. On the other hand, the gap with other aspects of AI is growing, in particular, with **reasoning** and **person-machine interaction**, central aspects for the development of a human, ethical, and anthropocentric AI, that is the focus of the European approach. The opacity and nature of black boxes of AI models are growing, together with the risk of creating systems exposed to biases in training data, systems that even experts fail to understand. **Tools are missing** to allow AI developers to certify the **reliability** of their models. It is crucial to inject into AI technologies, ethical values of **fairness** (how to avoid unfair and discriminatory decisions), **accuracy** (how to provide reliable information), **confidentiality** (how to protect the privacy of the involved people) and **transparency** (how to make models and decisions comprehensible to all stakeholders). This **value-sensitive design** approach, yet to be fully operationalised, is strongly needed for boosting widespread social acceptance of AI, without inhibiting its power.

Furthermore, as increasingly complex sociotechnical systems emerge, consisting of many interacting people and intelligent and autonomous systems, AI acquires an important **societal dimension**. A key observation is that **a crowd of intelligent individuals (assisted by AI tools) is not necessarily an intelligent crowd**. On the contrary, it can be stupid in many cases, because of undesired, unintended **network effects** and **emergent aggregated behavior**. Examples abound in contemporary society. For example, using a car navigation to avoid a traffic jam on the main route can cause additional jams in the local alternative routes. In the field of opinion formation and diffusion, a crowd of citizens using **social media** as a source of information is subject to the **algorithmic bias of the platform's recommendation mechanisms** suggesting personalised content. This bias will create echo chambers and filter bubbles, sometimes induced in an artificial way, in the sense that without the personalisation bias the crowd would reach a common shared opinion. Recommendations provided

by AI systems may make sense at an individual level but they may result in undesired collective effects of information disorder and radicalisation.

The flow of information reaching us via the online media platforms and ecommerce marketplaces is optimised not by the information content or product quality but by **popularity** and **proximity to the target**. This is typically performed in order to maximize platform usage. The **recommendation algorithms** suggest the interlocutors, the products and the contents at which we get exposed, based on matching profiles, promoting the most popular choices for people similar to us. As a result, we observe the appearance of the "**rich get richer**" phenomenon: popular users, contents and products get more and more popular. In the online marketplaces this would mean that a few businesses get the biggest share of the market while many have to share the rest. These businesses become the hubs of the network, gathering most of users' purchases to the detriment of the vast majority. In the social media, that would mean that certain popular peers or contents gather all the user's attention, becoming the hubs of the social network. As a consequence of **network effects** of AI recommendation mechanisms for online marketplaces, search engines and social networks, the **emergence of extreme inequality** and **monopolistic hubs** is artificially amplified, while **diversity** of offers and easiness of **access to markets** are artificially impoverished.

Aim

There is a wide consensus that AI will bring forth changes that will be much more profound than any other technological revolution in human history. Depending on the course that this revolution takes, AI will either empower our ability to make more informed choices or reduce human autonomy; expand the human experience or replace it; create new forms of human activity or make existing jobs redundant; help distribute well-being for many or increase the concentration of power and wealth in the hands of a few; expand democracy in our societies or put it in danger. Our generation carries the responsibility of shaping the AI revolution. The choices we face today are related to fundamental ethical issues about the impact of AI on society, in particular, how it affects labour, social interactions, healthcare, privacy, fairness, security, and markets.

The current technological advancements and developments of AI that can occur in the near future, if driven along the path of a **human-centric AI**, could represent an important transformation factor for ecommerce/digital services and for the Internal Market. Novel AI platforms for ecommerce and digital services based on human-centric AI interaction mechanisms have the potential to **mitigate monopolistic concentration**, deliver more **open and resilient markets**, and better connect the diverse demands of European consumers to the diverse offer of European products and services, by fostering **diversity "by-design".** AI-based recommendation and interaction mechanisms may help departing from the current purely "advertisement-centric" model, focusing on the interests of platforms in maximising intermediation revenues, to a systemic approach where focus is on the interest of citizens in accessing and sharing of high quality contents, the interest of consumers to broaden their choices and opportunities, and the interest of advertisers in broadening their audience and customer base.

Within this landscape, the European strategy for the next-generation digital services and online platforms is of utmost importance, with impacts that will go far beyond consumer protection, towards shaping the digital society that will emerge. Coherently with the recent strategic white paper of the European Commission "*On Artificial Intelligence – A European approach to excellence and trust*",

we recommend to develop the European provision on artificial intelligence in the area of ecommerce and digital services - in the context of reforming the E-commerce directive and introducing the Digital Services Act - with the aim to properly address and exploit the transformative impact of upcoming human-centric AI developments, to the purpose of social progress. Accordingly, our proposed **recommendations to EU policy makers** follow a double line of reasoning and interventions: **topic-wise**, and **instrument-wise**.

Topic-wise, it is crucial to address and operationalise the following challenges:

a) trustworthy, explainable AI in order to **fight novel forms of discrimination and manipulation** and **empower citizens**, and;

b) societal-aware AI in order to **fight polarisation and inequality** and **pursue diversity and openness.**

Instrument-wise, it is important to realise that the scientific and technological landscape is not yet fully mature to address all the open challenges discussed here. Therefore a mix of policies is needed, that tackle the problem at three levels:

a) a bold **EU investment in fundamental and applied research in human-centric AI;**

b) a bold **EU investment in creating novel online platforms and digital services** embodying human-centric AI mechanisms (and/or in supporting the scaling of existing coherent initiatives);

c) a coherent set of **EU regulations concerning AI, big data and digital services**, designed not only to seize the opportunities and mitigate risks, but also to inspire research & development in AI, big data and digital platforms towards an inclusive, equal and diverse society.

1. ARTIFICIAL INTELLIGENCE, BIG DATA, MACHINE LEARNING

KEY FINDINGS

Modern AI is mostly learning from data.

Pro:

- ability to deal with unsolved problems requiring "intelligence", such as understanding natural language and recognising contents of images.

Cons:

- other aspects of intelligence are still lagging behind, such as reasoning, adaptation to human environments and human-machine interaction;
- quality of AI systems are totally dependent on quality of data.

Artificial Intelligence (AI) has progressed to the point where it is an essential component in nearly all sectors of today's modern economy, with a significant impact on nearly all areas of our private, social, and political lives.

The term **Artificial Intelligence** refers to an ecosystem of models and technologies for **perception**, **reasoning**, **interaction** and **learning**. The return of the term AI to the limelight in the recent years is mainly related to this last aspect of learning from data: **Machine Learning (ML)** that made a jump ahead thanks to the emergence of **Big Data**. The mix is explosive: data reach the critical mass of examples to learn from, algorithms discover predictive patterns and patterns hidden in the data, the high-performance architectures succeed to compensate for the computing resources and storage needed.

In this new spring, arriving after the long winter originating from the disillusionment of the eighties, AI managed to successfully tackle open challenges, such as understanding and translation of texts and speech, recognition of the content of images and video, and other tasks believed to require intelligence. Approximately ten years ago, it was noticed that some long-known learning models, such as artificial neural networks, hitherto ineffective to the mentioned tasks, if equipped with a huge number of internal variables (neurons) and associated parameters, and properly trained on large set of example data, suddenly can make a qualitative leap and are able to generalize, from the pixels of the images or from the terms of the example texts, the general "concepts" that allow to recognize, classify, and accurately predict new images and new texts. A learning mode that requires new high-performance computing paradigms. This frame, **Artificial Intelligence = Big Data + Machine Learning**, is widely confirmed today.

As **data-driven** decision making becomes visible in everyday life, the interest around AI and Big Data is mounting. Society has shifted from being predominantly "analog" to "digital" in just a few years: society, organizations, and people are increasingly interconnected. As a consequence of digitization, data is collected about anything, at any time, and at any place. The spectacular growth of the digital universe, summarized by the term **Big Data**, makes it possible to record, observe, and analyse the behavior of people, machines, and organizations.

The term **Big Data** has been used diversely in different communities, according to their topics of interest and priorities. A set of attributes that characterize Big Data: velocity, volume, variety and veracity, are often cited, and named the four **V**'s**. Velocity** relates to the streaming nature of data and its speed; **volume** is about the sheer size of such data, calling for filtering/compression strategies or special measure for storing and processing it; **variety** stands for heterogeneity of the data sources that can be involved, both in terms of formats and representation, and in terms of semantics; **veracity** points to data quality issues and trustworthiness of the information. In the recent discourse, a fifth V has been added to the list, standing for **value**, aimed to emphasize the fact that turning Big Data sources and associated analytics tools into value (economic, social or scientific) is both important and far from trivial.

Machine learning methods exploit large "training" datasets of examples to learn general rules and models to classify data and predict outcomes (e.g., classify a taxpayer as fraudulent, a consumer as loyal, a patient as affected by a specific disease, an image as representing a specific object, a post on a social media as expressing a positive emotion). These analytical models allow scientists to "produce reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical relationships and trends in the data. Machine learning tasks are typically classified into two broad categories, depending on whether there is a learning "signal" or "feedback" available to a learning system: supervised and unsupervised learning. In supervised learning tasks the computer is presented with example inputs and their desired outputs, typically given by a "human expert", and the goal is to learn a general rule (or function, model) that maps inputs to outputs. In unsupervised learning tasks, there are no labels given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

In this context, all the strengths and weaknesses of the current situation are highlighted. On the one hand, the capacity of machine learning models to generalize from training data that are larger and of higher quality, is growing; this capacity explains the extraordinary progress in image recognition and in robotic vision, in understanding text and speech, in automatic translation, in medical diagnosis, in risk assessment, in predictive maintenance. On the other hand, the gap with other aspects of AI is growing, in particular, **reasoning** and **person-machine interaction**, central aspects for the development of a human, ethical, and anthropocentric AI, that is the focus of the European approach. Accordingly, AI created unprecedented opportunities but also new risks, such as the **opacity** of AI models that makes it difficult, even for experts, to explain the rationale of their conclusions. This can represent a critical point in technological and social terms, since the risk is real, as demonstrated by recent episodes, of training systems that are compromised by prejudice and discrimination **bias**, learned from the training data.

2. AI-RISKS AND CHALLENGES

KEY FINDINGS

Machine learning models are often black-boxes:

- The logic of a decision-making AI model may be hidden even to developers and experts.

Risks:

- Learning from biased data (either by accident or by necessity);
- Trained models for, e.g., on-board vision and obstacle recognition, may inherit bugs harmful for safety;
- Profiling models for, e.g., predictive justice, may inherit discrimination of known (or new) vulnerable groups.

Direction to pursue: explainable, trustworthy AI:

- Key for effective human-machine collaboration in high-stakes decision making

AI created unprecedented opportunities but also new risks. The exponential growth of the capabilities of AI models allows to reach levels of value and generalization never attained before. However, the **opacity** of these models has also grown and their nature of **black box** makes it difficult, even for experts, to explain the rationale of their conclusions. This can represent a critical point in technological and social terms, since the risk is real, as demonstrated by recent episodes, of training systems that are compromised by prejudice and discrimination **bias**, which learnt from the training data. It is therefore possible that "learning from the digital traces of past decisions can lead to incorporating invisibly existing prejudices into the resulting models, perpetuating them", as reported in the work that launched the study of discrimination-aware data mining in 2008[1].

Paraphrasing Frank Pasquale, author of *The black box society*[2], we have seen increasingly opaque algorithms spread, employed to infer **intimate traits** of individuals. These algorithms yield classification and prediction models of behavioral traits of individuals, such as credit score, insurance risk, health status, propensity to crime, personal preferences and orientations, using personal data disseminated in the digital environment by citizens, with or sometimes without their awareness. Such automated decision-making systems are often **black boxes** that, observing the characteristics of the users, predict a class, a judgment, a vote and suggest decisions; but without explaining the reason of the proposed prediction, or recommendation. It is not just a matter of transparency. The models are trained on examples reconstructed on the basis of the digital traces of user activities, such as movements, purchases, online searches, opinions expressed on social networks. As a result, the models inherit the prejudices and defects - the **biases** - that are hidden in the training data, hiding them, in turn, in the decision algorithms that risk suggesting unfair, discriminatory or simply wrong choices, possibly without the awareness of the decision maker and the subject of the final decision. If a chat box, an AI that conversates with users on social networks, learns from the wrong examples,

[1] Pedreschi et al. 2008.

[2] Pasquale 2015.

e.g., racist utterance, it will be racist in turn - and its creators will have to quickly and silently remove it. Many cases that have already happened, like the one of the Twitter bot Tay released by Microsoft in 2016 and quickly shut down after it learned offensive and racist behavior from online conversations[3], warn us that delegating choices to black box algorithms may reveal a bad idea. The cases in the next sections bring further evidence to this claim.

2.1. Black boxes

- *Compas*, owned by Northpointe Inc., is a predictive model of the risk of criminal recidivism, used until recently by various US Courts of Justice in support of judges' decisions on release requests. Journalists at *Propublica.org* have collected thousands of use cases of the model and shown that it has a strong racist bias: blacks who will not actually commit a crime again are assigned a double risk compared to whites in the same conditions[4]. The model, developed with ML techniques, presumably inherited the bias present in the historical sentences and is affected by the fact that the American prison population over-represents blacks over whites;
- The top three agencies for credit risk in the United States, Experian, TransUnion and Equifax are often discordant. In a study of 500,000 cases, 29% of credit applicants had a risk assessment with differences of more than 50 points from the three companies, which can mean tens of thousands of dollars of difference in overall interests. Such a wide variability suggests very different evaluation hypotheses, as well as opaque ones, or a strong arbitrariness[5];
- In the 1970s and 1980s, St. George's Hospital Medical School in London used software to filter job applications which was later found to be highly discriminatory towards women and ethnic minorities, inferred by surname and place of birth. Algorithmic discrimination is not a new phenomenon and is not necessarily due to ML[6];
- A classifier based on DL can be very accurate with respect to training data, and at the same time completely unreliable, for example because it has learned from bad quality data. In a case of image recognition aimed at distinguishing husky wolves in a large dataset of images, the resulting black box was dissected by researchers only to find out that the decision to classify an image as 'wolf' was based solely on the snow in the background[7]! The fault, of course, is not on DL, but on the accidental choice of training examples in which, evidently, each wolf had been photographed on the snow. So, a husky in the snow is automatically classified as a wolf. Now we carry this example on the vision system of our self-driving car: how can we be sure that it will be able to recognize correctly every object around us?
- Various studies, such as the one mentioned in the note[8], show that the texts on the web (but also on the media in general) contain bias and prejudices, such as the fact that the names of whites are more often associated with words with a positive emotional charge, while names of black people are more often associated with words with a negative emotional charge. Therefore, models trained on texts for the analysis of sentiment and opinions are highly likely to inherit the same prejudices;

[3] Tay (bot) https://en.wikipedia.org/wiki/Tay_(bot) [19-4-2020].

[4] Machine Bias https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [19-4-2020].

[5] Carter et al. 2006.

[6] Lowry – Macpherson 1988.

[7] Ribeiro et al. 2016.

[8] Caliskan-Islam et al. 2016.

- Bloomberg data journalists[9] have shown how the automatic model, used by Amazon to select the neighbourhoods of American cities to offer "same-day delivery" for free, has an ethnic bias. The software, without the knowledge of the company, systematically excludes from the offer areas inhabited by ethnic minorities in many cities, while including neighbouring ones. Amazon replied to the journalistic investigation that it was not aware of the practice, because the ML model was totally autonomous and based its choices on previous customer activity. In short, it is the fault of the algorithm.

2.2. "Right to Explanation"

Through Machine Learning and Deep Learning (DL) we are creating systems that we don't know fully understand yet. The European legislator has become aware of the trap, and perhaps the most innovative and forward-looking part of the General Data Protection Regulation (**GDPR**), the new privacy regulation that came into force in Europe on May 25th of 2018, is precisely the right to **explanation**, or the right to obtain meaningful information on the logic adopted by any automatic decision system that has legal effects, or "similarly relevant", for the people involved. Without a technology capable of explaining the logic of black boxes, however, the right to explanation is destined to remain a dead letter, or to outlaw many applications of opaque ML. It's not just about **digital ethics**, about avoiding **discrimination** and **injustice**, but also about **security** and **corporate responsibility**. In areas such as self-driving cars, robotic assistants, home automation and manufacturing IoT systems, personalised precision medicine, companies launch services and products with AI components that could inadvertently incorporate wrong decisions, crucial for safety, learned from errors or by spurious correlations in the learning data. For example, how to recognize an object in a photo by the properties not of the object itself but of the properties of the background, due to a systematic bias in the collection of learning examples. How can companies trust their products without understanding and validating their operation? **Explainable AI technology** is essential to create products with reliable AI components, to protect consumer safety and to limit industrial liability. Accordingly, the scientific use of ML, as in medicine, biology, economics or social sciences requires comprehensibility not only for trusting the results, but also for the open nature of scientific research, so that it can be shared and progressed. The challenge is tough and stimulating: an explanation must not only be correct and exhaustive, but also **understandable** to a plurality of subjects with different needs and competences, from the user being the subject of a decision, to AI solution developers, to researchers, to data scientists, to policy makers, supervisory authorities, civil rights associations, journalists.

What is an "**explanation**" has already been investigated already by Aristotle in his Physics, a treatise dating back in the 4th century BC. Today it is urgent to give a functional meaning, as an interface between people and the algorithms that suggest decisions, or that decide directly, so that AI serves to enhance human capabilities, not to replace them. In general, the explanatory approaches differ for the different types of data from which you want to learn a model. For tabular data for example, the explanation methods try to identify which variables contribute to a specific decision or prediction in the form of if-then-else rules or decision trees.

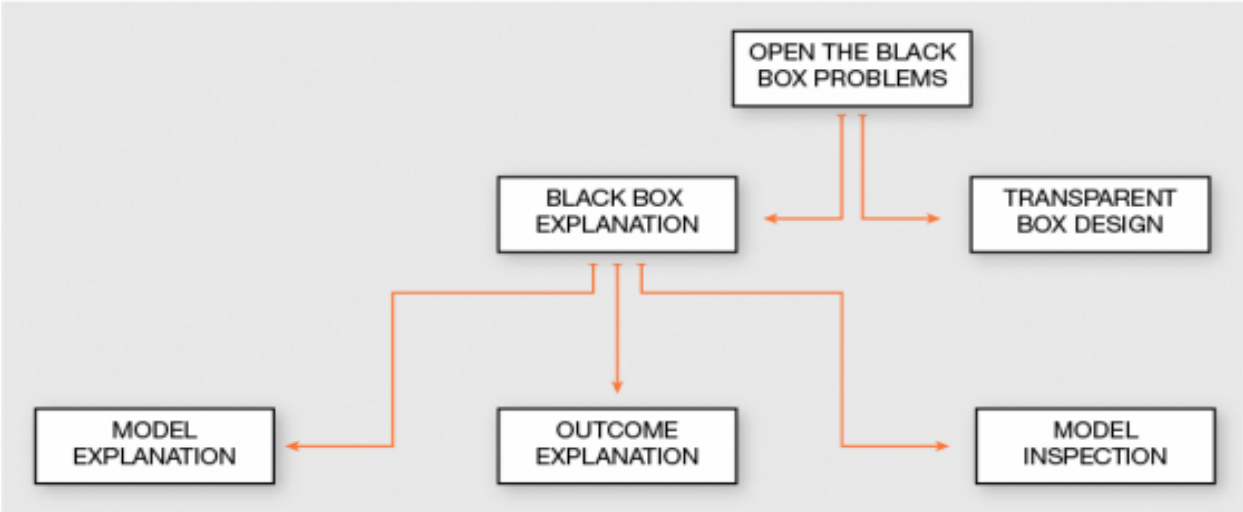
Over the past two years there has been an impetuous research effort on **understandable AI**, but a technology for the explanation that is practical and systematically applicable has not yet emerged. There are two broad ways of dealing with the problem:

[9] Amazon Doesn't Consider the Race of Its Customers. Should It? https://www.bloomberg.com/graphics/2016-amazon-same-day/ [19-4-2020].

- **explanation by-design**: **XbD**. Given a set of decision data, how to build a "transparent automatic decision maker" that provides understandable suggestions;
- **explanation of the black-boxes**: **Bbx**. Given a set of decisions produced by an "opaque automatic decision maker", how to reconstruct an explanation for each decision.

The most recent works in literature are discussed in the review[10], organizing them according to the ontology illustrated in the figure below. Today we have encouraging results that allow us to reconstruct individual explanations, answers to questions such as "Why wasn't I chosen for the place I applied for? What should I change to overturn the decision?"

Figure 1: Open the Black Box Problems



Source: Guidotti et al., 2018. Open online access at: https://dl.acm.org/doi/10.1145/3236009

The first distinction concerns XbD and Bbx. The latter can be further divided between Model Explanation, when the goal of explanation is the whole logic of the dark model, Outcome Explanation, when the goal is to explain decisions about a particular case, and Model Inspection, when the goal is to understand general properties of the dark model.

We are rapidly evolving from a time when people coded algorithms, taking responsibility for the correctness and quality of the software produced and the choices represented in it, to a time when machines independently infer algorithms based on a sufficient number of examples of the expected input / output behaviour. In this disruptive scenario, making sure that the AI black boxes are open and understandable is functional not only to verify their correctness and quality, but above all to align the algorithms with human values and expectations and to preserve, or possibly to expand, the autonomy and awareness of our decisions[11].

2.3. Direction to pursue

For all these reasons, the strategy for AI must develop on two synergistic levels: on one hand, it is crucial to innovate, experiment and systematically evaluate the application of the state of the art in AI in **all scientific, social and industrial sectors**, but it is not enough. On the other hand, it is necessary to intensify **research** efforts to deal with **open problems**. A **trans-national** and **multidisciplinary**

[10] Guidotti et al. 2018.

[11] Pedreschi et al. 2018.

approach is needed, that pivots on excellence in AI in research institutions to build a strong alliance with all other disciplines, including the social sciences, cognition and ethics.

The opacity and nature of black boxes of AI models are growing, together with the risk of creating systems exposed to biases in training data, systems that even experts fail to understand. **Tools** are missing to allow AI developers to certify the **reliability** of their models. It is crucial to inject into AI technologies, ethical values of **fairness** (how to avoid unfair and discriminatory decisions), **accuracy** (how to provide reliable information), **confidentiality** (how to protect the privacy of the involved people) and **transparency** (how to make models and decisions comprehensible to all stakeholders). This **value-sensitive design** approach is aimed at boosting widespread social acceptance of AI, without inhibiting its power.

Europe has set its own approach to AI, not only with the forward-looking **GDPR**, but with many more initiatives in order to put AI at the service of its citizens and economy. In its Communication[12], the European Commission puts forward the three main pillars for AI:

a) Being ahead of technological developments and encouraging uptake by the public and private sectors;

b) Prepare for socio-economic changes brought about by AI;

c) Ensure an appropriate ethical and legal framework.

On February 2020, a White Paper was published aiming to foster a European ecosystem of excellence and trust in AI, along with a Report on the safety and liability aspects of AI. The GDPR was a major step for building trust but actions are also necessary to ensure legal clarity in AI-based applications. There are several similar initiatives taken in the U.S.A.[13] and China[14] to promote the use and development of AI, but also to regulate it.

We are already seeing the rate of **AI's penetration** in real-world applications being limited by most systems' ability to adequately interact with humans, by related issues of user acceptance, and by the capability of dealing with complex, dynamic, unpredictable environments. To be truly adopted on a social and industrial scale, AI must be able to amplify human potential, especially on a cognitive level, not aiming at replacing human expertise. Ethical constraints are stimuli to create better innovations, not obstacles to innovation. The **GDPR**, to put it with Tim Cook, Apple's CEO, is not a problem but a great opportunity. Values are principles of good design for a technology aiming at individual and social wellbeing. It is an exciting scientific and technological **challenge**, which requires fundamental multidisciplinary research for the development of new models. From statistical learning and correlations to causal reasoning. From black box decision models to interpretable models, to design AI systems that **converse** with humans to help them improve the quality of their **decisions**. To measure, predict and preserve the **sustainability** and **resilience** of the complex (techno-) social systems we inhabit. To design new ways of **interaction** between people and machines in a way that the former reach higher levels of **awareness** and the latter higher levels of learning and understanding of the context and of human **reasoning**. Finally, to design new decentralized and **distributed** ways for collection and data management, the energy that feeds the AI, going beyond the current phase of extreme data centralization by a few monopolistic corporations. Indeed, these are hindering open science, innovation, **common good** and, with their monstrous date centres, **environmental sustainability**.

[12] Artificial Intelligence for Europe 2018.

[13] https://www.whitehouse.gov/ai/.

[14] A New Generation of Artificial Intelligence Development Plan, 2017.

Whoever leads the way in **future generations** of AI technology will set the standards for the **values** and **ethics** embedded in it. AI's development to date has brought us to an important crossroads: depending on the direction of future development, AI could destroy or create new jobs; empower citizens or impair their autonomy and privacy; increase the complexity of a globalized, interconnected world, thereby increasing the risk of systemic failures, or provide us with transparent, usable, understandable tools for taming that complexity, e.g., in epidemic emergencies such as the Covid-19 pandemic we are experiencing in these days. The goal to aim for is a **Human-centric AI** at the benefit of humans at an individual and at a social level, systems which incorporate European ethical values by-design, which are able to understand and adapt to real environments, interact in complex social situations, and expand human capabilities, especially on a cognitive level.

While there are aspects of AI where big non-EU corporations and countries may be better positioned, a specific **European brand of AI** that focuses on European values could be beneficial. The empowerment of humans and a benign impact on society can provide Europe with a unique competitive advantage and make a significant contribution to the well-being and prosperity of European citizens. On the big data landscape, while big corporations have an advantage on consumer behavioral data, Europe is better positioned for data coming from the public sector, such as **health, education**, etc., thanks to the well developed public administration and welfare systems, as well as for **industrial data**, coming from the ecosystem of large, medium and small industries that turned digital.

3. THE SOCIAL DIMENSION OF AI

KEY FINDINGS

Machine learning from Big Data:

- ML is a form of collective intelligence, which discovers general patterns from human experience encoded in data.

Discovered patterns of collective behaviour are applied to individual users/consumers:

- online platforms apply general patterns/profiles to individuals;
- based on an advertising logic;
- targeted at maximising capture of users' attention, clicks, likes, purchases.

As increasingly complex sociotechnical systems emerge, consisting of many (explicitly or implicitly) interacting people and intelligent and autonomous systems, AI acquires an important societal dimension. A key observation is that **a crowd of (interacting) intelligent individuals is not necessarily an intelligent crowd**. On the contrary, it can be stupid in many cases, because of undesired, unintended **network effects** and **emergent aggregated behavior**. Examples abound in contemporary society. For example, using a car navigation to avoid a traffic jam on the main route can cause additional jams in the local alternative routes. In the field of opinion formation and diffusion, a crowd of citizens using **social media** as a source of information is subject to the **algorithmic bias of the platform's recommendation mechanisms** suggesting personalised content. This bias will create echo chambers and filter bubbles, sometimes induced in an artificial way, in the sense that without the personalisation bias the crowd would reach a common shared opinion. Recommendations provided by AI systems may make sense at an individual level but they may result in undesired collective effects of information disorder and radicalisation.

The interaction among individual choices may unfold dramatically into **global challenges linked to economic inequality, environmental sustainability, and democracy**. Aggregated network and societal effects of AI and their impacts on society are not sufficiently discussed in the public and not sufficiently addressed by AI research, despite the striking importance to understand and predict the aggregated outcomes of sociotechnical AI-based systems and related complex social processes, as well as how to avoid their harmful effects. Such effects are a source of a whole new set of **explainability**, **accountability**, and **trustworthiness** issues, even assuming that we can solve those problems for an individual machine-learning-based AI system. Therefore, we cannot concentrate solely on making individual citizens or institutions more aware and capable of making informed decisions. We also need to study the emerging network effects of crowds of intelligent interacting agents, as well as the design of mechanisms for distributed collaboration that push toward the realization of the agreed set of values and objectives at collective level: sustainable mobility in cities, diversity and pluralism in the public debate, and fair distribution of economic resources. Societal AI is emerging as a new field of investigation of potentially huge impact, requiring the next step ahead in trans-disciplinary integration of AI, data science, social sciences, psychology, network science, and complex systems.

3.1. Wisdom of crowds?

AI-powered recommendation mechanisms strive to make individual users/customers more "intelligent":

- Hidden assumption: a crowd of intelligent individuals, powered by machines and algorithms, is an intelligent crowd.

Unfortunately, this is not necessarily true:

- A crowd of intelligent individuals may be very stupid;
- This is due to the aggregated effect of interactions occurring in the complex networks surrounding us;
- Examples: online marketplaces, online public conversation, traffic.

Surowiecki[15] in his book *The Wisdom of Crowds* provides many promising and intuitive examples of a wise — and accurate — crowd. Nevertheless, he also presented situations (such as rational bubbles) in which the crowd produced very bad judgements, and argued that in these types of situations their cognition or cooperation failed because (in one way or another) the members of the crowd were too conscious of the opinions of others and began to emulate each other and conform, rather than think differently.

One would expect that, with the advent of AI, the **crowd**, powered now by machines and algorithms, could instead be an intelligent crowd and make intelligent choices and take intelligent decisions. Alas, that's not always the case. Due to the aggregated effect of interactions occurring in the complex networks surrounding us, a crowd of intelligent individuals may still be stupid. There are several examples in our everyday life, in online marketplaces, in online public conversation, in mobility and traffic.

Surowiecki states that there are three conditions for a group to be intelligent: **diversity**, **independence**, and **decentralisation**. He stresses the need for diversity within a crowd to ensure multiple viewpoints. In a wise crowd, "one person's opinion should remain **independent** of those around them," and individuals must "make their own opinion based on their **individual knowledge**." Instead, in our complex social systems, individuals influence one another, and are influenced by the media they choose. As a consequence, the wisdom of the crowd effect often vanishes.

[15] Surowiecki 2004.

3.2. Social networks, search engines, online marketplaces

Recommendation mechanisms:

- Suggest most popular choices (peers, contents, products) for people like you (based on matching profiles);
- These mechanisms are very effective for platforms to increase clicks.

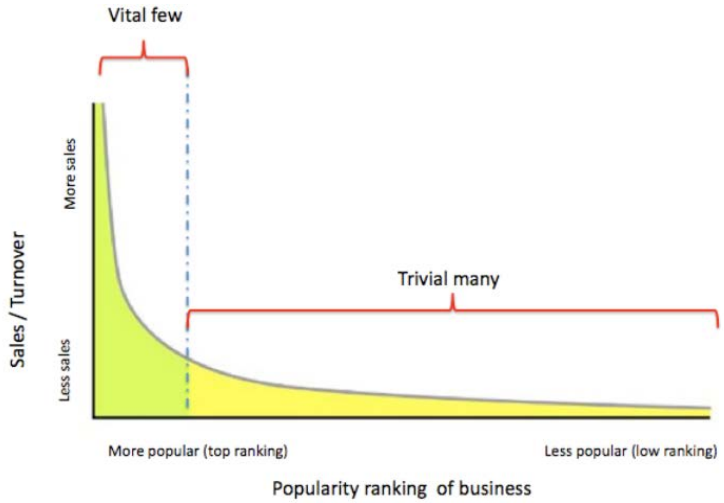
Network effects:

- Emergence of hubs and extreme inequality;
- Popular peers, contents, products get more and more popular;
- Rich get richer, to the detriment of the vast majority;
- Diversity is artificially impoverished.

The flow of information reaching us via the online media platforms is optimised not by the information content or relevance but by popularity and proximity to the target. This is typically performed in order to maximize platform usage. The recommendation algorithms suggest the interlocutors and the contents at which we get exposed, based on matching profiles, promoting the most popular choices for people similar to us.

As a result, we notice the appearance of the "**rich get richer**" phenomenon: popular peers, contents and products get more and more popular. As shown in the picture below, in the online marketplaces this would mean that a few businesses get the biggest share of the market while many have to share the rest. These businesses become the hubs of the network, gathering most of users' purchases to the detriment of the vast majority. In the social media, that would mean that certain popular peers or contents gather all the user's attention and they, on their turn, become the hubs of the social network.

Figure 2: Popularity ranking of business



Source: Source: A. Kripaa, The Long Tail and Cloud-based ERP Solutions, Ramco Systems blog, accessed online at https://blogs.ramco.com/the-long-tail-and-cloud-based-erp-solutions on April 21, 2020

A vital few, most popular products seize the more sales while many are way less popular with less sales.

We need to design novel social AI mechanisms for online marketplaces, search engines and social networks, focused on mitigating the existing inequality introduced from the current generation of recommendation systems. We need mechanisms for helping individuals acquiring access to diverse content, different people and opinions, to avoid that diversity is artificially impoverished.

Recommendations for like-minded peers and contents:

- Effective for platforms to increase clicks;
- Effective for advertisers? Questionable, due to selection effect stronger than real influence;
- Dangerous for users: artificial reinforcement of own opinions and preferences, increased polarisation, radicalisation, bubbles.

Network effects:

- Reduction of diversity of opinions;
- Threat to pluralistic debate;
- Threat to the wisdom of the crowd (ultimately, to democracy).

3.3. Opinion diffusion on online media platforms

For centuries, people shaped own convictions thanks to the exchange of opinions with other people. These relationships, based on the level of mutual trust, play a fundamental role in the process that leads us to strengthen or change our opinions. With the spread of media and, recently, of social media we are witnessing a large-scale exploitation process, especially for business purposes, of the cognitive mechanism of the **confirmation bias** – the tendency to prefer information that is close to our convictions, and induces us to interact with peers that share our preferences. How do these mechanisms alter the traditional ways of forming opinions? What impact do they have on fragmentation, polarisation and radicalisation of ideas?

Social media are increasingly used to share ideas, but also, as the report Reuters 2018 on *Digital News* shows, to read and exchange news. This means that the interlocutors and the contents with which we interact, together with the external effects in the dynamics of opinions, become susceptible to the influences deriving from the interaction mechanisms of the social media platforms in use, built with a **marketing** objective: to maximize the **number of users** and the time they spend on the platform, catching their **attention**. To achieve this, the information that reaches them is not randomly selected. There is a precise **algorithmic bias**: we are systematically exposed to news related to the topics we like most and to the opinions that are closest to ours, so as to be nudged to interact with the proposed contents and return to the platform. After all, marketing has always exploited a cognitive mechanism that is innate with humans, the confirmation bias, defined as: «a mental process that consists in searching, selecting and interpreting information in order to place greater attention, and therefore attributing greater credibility to those who confirm own beliefs or hypotheses, and vice versa, ignoring or diminishing information that contradicts them. The phenomenon is stronger in the context of topics that arouse strong emotions or that touch deeply rooted beliefs»[16]. The mechanisms

[16] Plous 1993.

of marketing, whether commercial, informational or political, leverage the confirmation bias, because each of us feels gratified when meeting subjects or contents that confirm their beliefs. Proposing to interact with people and content close to the beliefs of users and consumer preferences ensures a higher click rate, a greater probability of attracting purchases, attention, likes. Ultimately, greater revenue for the platform.

This mechanism of **confirmation bias** is exploited by online platforms, for information search and social networking and media, which employ recommendation algorithms based on AI to catalyze users' attention. As a side effect, the platform amplifies and reinforces individual bias, resulting in extreme **polarisation** of opinions and **filter bubbles** at the social level, with dramatically negative consequences on the pluralistic public debate needed to nurture democracy. In addition, access to information is often maliciously biased by either commercially or politically motivated influence agents.

The most shocking result of a research recently published in the journal *PlosOne*[17] is that the number of groups with different opinions increases, within a population, with the increasing intensity of the algorithmic bias. In other words, online platforms can favor the fragmentation of opinions causing the emergence of opposing "parties", or "bubbles", even when, in the absence of bias, the population would reach consensus - a division therefore that is totally artificial, induced by the interaction mechanism of the platform that inhibits comparison between different positions. In practice, algorithmic bias leads to an increase in **polarisation** and **radicalisation** of opinions, amplifying the already powerful cognitive mechanism that we have incorporated, the confirmation bias.

The second effect of algorithmic bias is on the **speed of opinion formation**: the changes of opinion of individuals are much slower when the bias is introduced, and therefore the population, as a whole, takes a much longer time to reach a stable situation. Even when it comes to consensus, the time to reach it becomes extremely long.

The algorithmic bias of the platforms can profoundly influence the results of **public debates** and the formation of **consensus** in society. An apparently harmless mechanism, always used in marketing, if used on a large scale has a potentially destructive "network effect", capable of creating and strengthening **"echo chambers"**, information **bubbles**, **radicalizing** public debate and **destabilising** democracy. This is a phenomenon we are witnessing with growing apprehension, registered by scholars and experts on social dynamics from all over the world[18]. This polarisation effect is probably even more dangerous for our democracy than fake news and the hate speech phenomenon, as the latter can be explained, in most cases, an effect of bubbles artificially reinforced by bias.

It's important to highlight however, that **fake news** is a serious phenomenon that got amplified on online media platforms. It's a phenomenon that has been present since before online media and even the Internet, but social media has been a major driver for fake news consumption. On social media, fake news is discussed and engaged in about as much as real news in spite of the fact that it is read ten times less. In a study from the IMCO committee[19], fake news are redefined as **malinformation**, i.e. the triple conjunction of human malevolence, marketing malpractice and engineered malware, and the costs of this malevolent combination are thoroughly studied. Among these, there are economic costs but also, most importantly, two major societal costs: the **integrity of information** and the **integrity of elections**. Fake news thus creates challenges for both the sovereignty of Member States

[17] Sirbu et al. 2019.

[18] Schmidt et al. 2018.

[19] Divina Frau-Meigs 2018.

and for the cohesion of the European Union that are complex and may have far-reaching consequences. Fake news leads to information disorders related to the transverse effect of malinformation, i.e. the trans-border, transmedia, trans-lingual and trans-time nature of Internet networks. The direct and indirect impact of fake news, though not proven yet by various nascent research protocols, may lead to long-term counter veiling deterrence measures which will reduce freedom of speech and the freedom to receive and impart information. Thus efforts to inhibit fake news will damage **democratic processes**.

But AI technology can also help at stopping fake news at almost all stages of the news process: editing, amplification and consumption[20]. To stop fake news at **editing** and **amplification** online media platforms are a natural candidate for discussion. They can label results, disclose their selection algorithms, and last but not least identify and control the **bots** – accounts that seem like real persons to people and particularly to the platform's algorithms. In fact, bots are just programs that simulate a person using these platforms, and can determine the popularity of the news. Stopping bots could be easy if the platforms enforce the "real identity" of their users. Identifying bots and identifying fake news, can be done manually or automatically via AI algorithms that study the content and the social context of its source.

Human-centric AI has a clear social dimension that may help us design novel platforms and mechanisms for access to news and information, focused on counterbalancing our built-in confirmation bias and transparently striving to expose people to assorted opinions, intelligently. We imagine mechanisms for helping individuals and communities become informed on controversial issues by offering multiple perspectives, connecting opposing views and conflicting arguments, and fostering critical thought. For example, a robot in a group conversation can highlight information that was unavailable to the group, or suggest omitted sources of important information. Advances in person-machine interaction models based on explainable AI have the potential to reach novel cognitive trade-offs between our confirmation bias and our curiosity of novelty and diversity, making it possible for more sustainable and humanized information ecosystems to emerge.

Navigation systems offer recommendations and directions to reach a destination:

- Make perfect sense to an individual driver;
- May be meaningless and misleading collectively.

Network effects:

- Extraordinary situations, such as accidents or big events;
- Polarisation effect: recommending most popular routes, thus artificially increasing traffic and risk in certain tracks while systematically avoiding alternative tracks and areas;
- Again, diversity is compromised, with negative network effects and externalities (largely underestimated today).

[20] Žiga Turk 2018.

3.4. Mobility, traffic and online navigation services

AI is a key factor in analysing and facilitating changes in cities and citizen behaviours related to mobility. Modern cities are the perfect example of environments where traffic monitoring systems, and GPS-enabled individual mobility traces, provide useful information to recommend a route and direct a driver to reach a destination avoiding traffic jams.

From the perspective of the individual driver, this system makes perfect sense. But collectively, this can be sometimes meaningless and misleading. Anyone who has used a **car navigation** system to bypass a traffic jam knows. Each navigation system generates recommendations that make sense from an **individual** point of view, and the driver can easily understand the rationale behind the recommendations. However, the sum of decisions made by many navigation systems can have grave consequences on the **traffic system** as a whole: from the traffic jams on local alternative routes to ripples propagating through the system on a larger spatial scale, to long-term behavior changes that may lead to drivers permanently avoid certain areas (which can have a negative economic impact on disadvantaged neighbourhoods), or artificially increase the risk of accidents on highly recommended roads.

It is important to study the emergence of collective phenomena at metropolitan level in personal navigation assistance systems with different recommendation policies, with respect to different policies for navigation recommendations (such as diversity) and different collective optimisation criteria (fluidity of traffic, safety risks, environmental sustainability, urban segregation, response to emergencies, etc.).

4. CONCLUSIONS AND RECOMMENDATIONS

There is a wide consensus that AI will bring forth changes that will be much more profound than any other technological revolution in human history. Depending on the course that this revolution takes, AI will either empower our ability to make more informed choices or reduce human autonomy; expand the human experience or replace it; create new forms of human activity or make existing jobs redundant; help distribute well-being for many or increase the concentration of power and wealth in the hands of a few; expand democracy in our societies or put it in danger. Our generation carries the responsibility of shaping the AI revolution. The choices we face today are related to fundamental ethical issues about the impact of AI on society, in particular, how it affects labour, social interactions, healthcare, privacy, fairness, security, and markets.

There is a need to develop the scientific foundations and technological breakthroughs to shape the AI revolution in a direction that is beneficial to humans on both individual and societal levels and that adheres to European ethical values and social, cultural, legal and political norms. The core challenge is the development of robust, trustworthy AI systems capable of what could be described as "understanding" humans, adapting to complex real-world environments and appropriately interacting in complex social settings. The aim is to facilitate AI systems that enhance human capabilities and empower individuals and society as a whole while respecting human autonomy, dignity and self determination.

The current technological advancements and developments of AI along the mentioned lines, that can occur in the near future (in two to five years), could represent an important transformation factor for ecommerce/digital services and for the Internal Market. Novel AI platforms for ecommerce and digital services based on AI interaction mechanisms have the potential to mitigate monopolistic concentration, deliver more open and resilient markets, and better connect the diverse demands of European consumers to the diverse offers of European products and services, by fostering diversity "by-design". AI-based recommendation and interaction mechanisms may help departing from the current purely "advertisement-centric" model, focusing on interest of platforms in maximising exchange of content in order to maximise advertisement revenues, to a systemic approach where focus is on the interest of citizens in accessing and sharing of high quality contents, the interest of consumers to broaden their choices and opportunities, and the interest of advertisers in broadening their audience and customer base.

In combination with on-going advancements in peer-to-peer networks and distributed ledger architectures (block-chain), AI interaction mechanisms have the potential to promote a rapid shift towards decentralized environments for ecommerce and digital services that are centred not on the intermediary of the transactions, but rather on consumers and providers directly, aiming therefore at maximising satisfaction of consumers and providers, both individually and collectively, rather than maximising the revenue of the intermediary. Therefore AI can have an impact on the Internal Market by pursuing objectives that are meaningful for consumers and providers, instead of success measures that are functional to intermediaries, and by mitigating the gate-keeper effect of current platforms' contracts. Clearly, such an ecosystem would be also beneficial to e-government and public procurement, and the same basic principles would apply both to marketplaces and information & media digital services, targeted towards the interest of consumers and providers to share high quality contents.

Within this landscape, the European strategy for the next-generation digital services and online platforms is of utmost importance, with impacts that go far beyond consumer protection, shaping the digital society that will emerge. Coherently with the recent strategic white paper on AI released by

the European Commission (*On Artificial Intelligence - A European approach to excellence and trust*, February 2020), we recommend to develop the European provision on artificial intelligence in the area of ecommerce and digital services - in the context of reforming E-commerce directive and introducing the Digital Services Act - with the aim to ascertain effective and efficient rights, institutions and procedures in the future and to properly address and exploit the transformative impact of upcoming technological developments, to the purpose of social development. Accordingly, the proposed **recommendations to EU policy makers** follow a double line of reasoning and interventions: **topic-wise**, and **instrument-wise**.

Topic-wise, it is crucial to address and operationalise the following challenges:

- **trustworthy, explainable AI** in order to **fight novel forms of discrimination and manipulation** and **empower citizens**, and;
- **societal-aware AI** in order to **fight polarisation and inequality** and **pursue diversity and openness.**

Instrument-wise, it is important to realise that the scientific and technological landscape is not yet fully mature to address the open challenges above. Therefore a mix of policies is needed, that tackle the problem at three levels:

- a bold **EU investment in fundamental and applied research in human-centric AI;**
- a bold **EU investment in creating novel online platforms and digital services** embodying human-centric AI mechanisms (and/or in supporting the scaling of existing coherent initiatives);
- a coherent set of **EU regulations concerning AI, big data and digital services**, designed not only to seize the opportunities and mitigate risks, but also to inspire research & development in AI, big data and digital platforms towards an inclusive, equal and diverse society.

4.1. Explainable AI: fostering trust in online platforms and marketplaces

Explainable AI for high-stakes decision making. Decision making in, e.g., medicine, justice, job recruiting, etc., is essentially a sociotechnical system, where a decision maker interacts with various sources of information and decision-support tools, a process whose quality should be assessed in terms of the final, aggregated outcome — the quality of the decision — rather than assessing only the quality of the decision-support tool in isolation (e.g., in terms of its predictive accuracy and standalone precision). It is therefore important to develop tools that explain their predictions and recommendations in meaningful terms, a property rarely matched by AI systems, based on machine learning, available today.

The explanation problem for a decision-support system can be understood as "where" to place a boundary between what algorithmic details the decision maker can safely ignore and what meaningful information the decision maker should absolutely know to make an informed decision. Thus, an explanation is intertwined with trustworthiness (what to safely ignore), comprehensibility (meaningfulness of the explanations), and accountability (humans keeping the ultimate responsibility for the decision). In this context, several questions emerge: what are the most critical features for explanatory AI? Is there a general structure for explanatory AI? How does an AI system reach a specific decision, and based on what rationale or reasons does it do so? Explanations should favor a human-machine interaction via meaningful narratives expressed clearly and concisely through text and visualizations, or any other human-understandable format revealing the why, why-not, and what-if.

Even in domains where the stakes all less high, such as in personalised advertising on online platforms (social media, search engines, marketplaces), an adequate level of explainability of automated recommendation mechanisms is of great importance for two reasons. First, to **empower consumers and citizens against discrimination and manipulations;** second, to **empower watchdogs and surveillance/enforcement agencies** (for, e.g., non-discrimination, consumer protection and privacy) in performing their actions. It is already clear from several concrete cases that the outcome of automated profiling for targeted personalised advertising can, often inadvertently, discriminate minority neighbourhoods and disadvantaged communities (such as systematically denying favourable prices and conditions), as well as exploit vulnerabilities (such as targeting people with addictive conditions with online betting ads). These phenomena are deeply rooted into the logic of machine learning, prone to inherit bias that is hidden in the training data. It is therefore key that explainable AI mechanisms, directed to diverse stakeholders (such as consumers and control institutions), are part of online platforms and marketplaces in an open and inspectionable format. A form of "right of explanation" for online platform and marketplaces, expanding and adapting the provisions adopted in the GDPR, might foster innovations in the sector of online advertising that, beyond bringing transparency and trust, could also improve the effectiveness of the system for both advertisers and consumers.

4.2. Societal-aware AI: fostering diversity in online platforms and marketplaces

As discussed in this report, the current wave of recommendation mechanisms supported by AI in various online platforms (marketplaces, social media, navigation systems, etc.) is functional to a specific business model based on advertising and profiling and it is potentially dangerous not only for potential discrimination and manipulation of the individual user/customer (see Sect. 2) but also for far reaching consequences at societal scale in terms of increased polarisation and reduced diversity (Sect. 3).

Diversity is the key strength of Europe at social, cultural and economic level. Europe should therefore strive for developing AI-powered tools and platforms that embrace and promote diversity and European values at all scales:

- Online media and information ecosystems that intelligently promote pluralism and connect different viewpoints on complex controversial issues;
- Online marketplaces that intelligently promote the diversity and quality of European products and companies at all scales;
- Online platforms that intelligently use AI advances in multi-lingual translation to break the barriers among international communities, both within EU and the rest of the world, making the diversity of EU media, culture and production easily accessible (see, e.g., EBU's EuroVox project[21] towards a diversity-oriented, multi-lingual digital platform for sharing high-quality public service news and media contents towards a potential population of 1 Billion people speaking 96 languages).

In particular, novel platforms for news consumption and public conversation should, contrary to current practice, put in place algorithmic mechanisms that try to **mitigate user confirmation bias** rather than strengthen it, and find ways to expose users to contents and individuals with different opinions, helping to get broader ideas of controversial or complex issues, observing them from

[21] https://tech.ebu.ch/eurovox.

multiple points of view. A different public information ecosystem may be envisaged, aimed at **diversity,** which would also be essential to preserve and nourish our collective intelligence, or the **wisdom of the crowd**, at any scale, from communities to global citizenship. As discussed in this report and demonstrated by a long history of research studies and empirical evidence[22], a **multitude** endowed with **healthy diversity and independence of opinions** can intelligently answer difficult questions (e.g. predicting an uncertain outcome), but diversity and, therefore, **collective intelligence** is quickly undermined by excessive influence and polarisation. The bias of current platforms is, in short, a danger, and measures are needed to stop or at least **mitigate their effects**, or even reverse them, as interdisciplinary researchers are trying to suggest[23].

In the meantime, **users could be informed** of how news are conveyed on the feeds of the platforms and of the fact that this could influence their opinions, and perhaps the mechanisms implemented by the platforms could be slowly withdrawn. The consequence of the discussion in this report is that the current platforms for accessing information, both traditional and social media, based on the "free" service paid for by advertising and personalised marketing, are not (and cannot be) functional to fight polarisation and radicalisation, and to promote diversity and pluralism. The ultimate aim of the business model aimed at capturing the attention of users is, by definition, in conflict with that of **promoting diversity and pluralism**. Therefore, a **different media ecosystem**, aimed at helping citizens to deal with narratives and content other than their favorite or usual ones, should implement **smart strategies to connect opposite and alternative visions**. It should be based transparently on algorithmic bias that push towards diversity and pluralistic visions. It should be an **open public platform**, a common good, independent of the media, companies and governments, which pursues the aim of preserving socio-diversity, essential for democracy.

Coherently with this picture, it is desirable that the EU pushes vigorous research along the lines of societal-aware AI, both with a strategic plan of fundamental and applied research in the **EU "brand" of human-centric AI**, and with **ad-hoc investments in novel EU-wide platforms** such as those listed above, to be created from scratch or to be scaled leveraging promising existing initiatives. At **regulatory level**, it would be desirable to devise **sustainable requirements fostering diversity in existing commercial social media platforms and marketplaces**, such as obligations to access advertising delivery data in proper formats in order to verify the exposure of advertisers, the performance of paid vs unpaid advertising, the compliance with minimal diversification requirements. AI and data science are not neutral technologies, they can be used for good or bad purposes; it is up to us, to our democratic debate, to Europe in particular, to decide which values to instil by-design in these technologies to create a real information ecology.

[22] Surowiecki 2004.

[23] Aslay et al. 2018; Garimella et al. 2017.

REFERENCES

- L.A. Adamic – N. Glance, The political blogosphere and the 2004 US election: divided they blog, Proceedings of the 3rd International Workshop on Link discovery 2005, pp. 36-43
- C. Aslay et al., Maximising the diversity of exposure in a social network, IEEE International Conference on Data Mining 2018, pp. 863-868
- A. Caliskan Islam et al., Semantics derived automatically from language corpora necessarily contain human biases, arXiv preprint arXiv:1608.07187 2016
- R. Carter – H.V. Auken., Small firm bankruptcy, Journal of Small Business Management 2006, 44: pp. 493-512
- L. De Biase, *Homo pluralis. Essere umani nell'era tecnologica*. Codice, Torino 2016
- K. Garimella et al., Reducing controversy by connecting opposing views, Proceedings of the Tenth ACM International Conference on Web Search and Data Mining 2017, pp. 81-90
- R. Guidotti et al., A survey of methods for explaining black box models, ACM computing surveys (CSUR) 2018, pp. 1-42
- R. Hegselmann – U. Krause, Opinion dynamics and bounded confidence: models, analysis and simulation, «Journal of Artificial Societies and Social Simulation» V (2002) 3
- S. Lowry – G. Macpherson, A blot on the profession, *British medical journal* (Clinical research ed.) 1988, pp. 657
- F. Pasquale, The black box society, Harvard University Press 2015
- D. Pedreschi et al., Discrimination-aware data mining, Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining 2008, pp. 560-568
- D. Pedreschi et al., Meaningful explanations of Black Box AI decision systems. Proceedings of the 33rd AAAI Conference on Artificial Intelligence 2019, 9780-9784
- S. Plous, *The Psychology of Judgment and Decision Making*. McGraw-Hill, New York 1993
- M.T. Ribeiro et al., "Why should I trust you?" Explaining the predictions of any classifier, Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining 2016, pp. 1135-1144
- A.L. Schmidt et al., Polarisation of the vaccination debate on Facebook, National Center for Biotechnology Information 36 (2018) 25, pp. 3606-3612
- A. Sirbu et al., Algorithmic bias amplifies opinion fragmentation and polarisation: A bounded confidence model, «PLoS ONE» 14(3), 2019
- J. Surowiecki, *The wisdom of crowds*, Anchor Books, New York 2004
- Artificial Intelligence for Europe, Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and social committee and the committee of the Regions, Brussels 2018
- Divina Frau-Meigs, Societal costs of "fake news" in the Digital Single Market, Study requested by the IMCO committee, 2018

- Žiga TURK, Technology as Enabler of Fake News and a Potential Tool to Combat It, In-Depth Analysis requested by the IMCO committee, 2018
- XAI (2019-2024, ERC Advanced Grants 2018) *Science and technology for the explanation of AI decision making*. https://xai-project.eu/
- SoBigData (2015-2024, H2020-Excellent Science Research Infrastructures) *Integrated Infrastructure for Social Mining & Big Data Analytics*. A research infrastructure at the second stage of "Advanced community", aggregating 32 partners of 12 EU Countries. http://www.sobigdata.eu/
- Humane-AI (2019-2020, H2020-FETFLAG-2018-01 Coordination Action) Toward AI Systems That Augment and Empower Humans by Understanding Us, our Society and the World Around Us. https://www.humane-ai.eu/

ANNEX

The human and social dimension of AI

Prof. Dr. Dino Pedreschi, University of Pisa (I)

Position paper (also presented at the first meeting of the OECD Network of AI Experts – **ONE AI**, Paris, 26-27 February 2020).

There is a wide consensus that AI will bring forth changes that will be much more profound than any other technological revolution in human history. Depending on the course that this revolution takes, AI will either empower our ability to make more informed choices or reduce human autonomy; expand the human experience or replace it; create new forms of human activity or make existing jobs redundant; help distribute well-being for many or increase the concentration of power and wealth in the hands of a few; expand democracy in our societies or put it in danger. Our generation carries the responsibility of shaping the AI revolution. The choices we face today are related to fundamental ethical issues about the impact of AI on society, in particular, how it affects labor, social interactions, healthcare, privacy, fairness and security.

I am a founding member of the **HumanE AI** network of research centers, set forth to develop the scientific foundations and technological breakthroughs needed to shape the AI revolution in a direction that is beneficial to humans on both individual and societal levels and that adheres to European ethical values and social, cultural, legal and political norms. The core challenge is the development of robust, trustworthy AI systems capable of what could be described as "understanding" humans, adapting to complex real-world environments and appropriately interacting in complex social settings. The aim is to facilitate AI systems that enhance human capabilities and empower individuals and society as a whole while respecting human autonomy, dignity and self determination.

The research of our laboratory (KDD Lab, the Knowledge Discovery and Data Mining Lab of University of Pisa and CNR) revolves around data science and AI and their impacts on society. Two key directions are *explainable AI* and *societal-aware AI*.

Explainable AI for high-stakes decision making. Decision making in, e.g., medicine, justice, job recruiting, etc., is essentially a sociotechnical system, where a decision maker interacts with various sources of information and decision-support tools, a process whose quality should be assessed in terms of the final, aggregated outcome — the quality of the decision — rather than assessing only the quality of the decision-support tool in isolation (e.g., in terms of its predictive accuracy and standalone precision). It is therefore important to develop tools that explain their predictions and recommendations in meaningful terms, a property rarely matched by AI systems, based on machine learning, available today.

The explanation problem for a decision-support system can be understood as "where" to place a boundary between what algorithmic details the decision maker can safely ignore and what meaningful information the decision maker should absolutely know to make an informed decision. Thus, an explanation is intertwined with trustworthiness (what to safely ignore), comprehensibility (meaningfulness of the explanations), and accountability (humans keeping the ultimate responsibility for the decision). In this context, several questions emerge: what are the most critical features for explanatory AI? Is there a general structure for explanatory AI? How does an AI system reach a specific decision, and based on what rationale or reasons does it do so? Explanations should favor a human-machine interaction via meaningful narratives expressed clearly and concisely through text

and visualizations, or any other human-understandable format revealing the why, why-not, and what-if. We are working to concrete answers to these questions within the ERC grant XAI (Science and technology for the eXplanation of AI decision making), led by my colleague Fosca Giannotti.

The social dimension of AI. As increasingly complex sociotechnical systems emerge, consisting of many (explicitly or implicitly) interacting people and intelligent and autonomous systems, AI acquires an important societal dimension. A key observation is that a crowd of (interacting) intelligent individuals is not necessarily an intelligent crowd. On the contrary, it can be stupid in many cases, because of undesired, unintended network effects and emergent aggregated behavior. Examples abound in contemporary society. Anyone who has used a car navigation system to bypass a traffic jam knows. Each navigation system generates recommendations that make sense from an individual point of view, and the driver can easily understand the rationale behind the recommendations. However, the sum of decisions made by many navigation systems can have grave consequences on the traffic system as a whole: from the traffic jams on local alternative routes to ripples propagating through the system on a larger spatial scale, to long-term behavior changes that may lead to drivers permanently avoid certain areas (which can have a negative economic impact on disadvantaged neighborhoods), or artificially increase the risk of accidents on highly recommended roads. The interaction among individual choices may unfold dramatically into global challenges linked to economic inequality, environmental sustainability, and democracy. In the field of opinion formation and diffusion, a crowd of citizens using social media as a source of information is subject to the algorithmic bias of the platform's recommendation mechanisms suggesting personalised content. This bias will create echo chambers and filter bubbles, sometimes induced in an artificial way, in the sense that without the personalisation bias the crowd would reach a common shared opinion. Again, a recommender system that makes sense at an individual level may result in an undesired collective effect of information disorder and radicalisation.

Aggregated network and societal effects and of AI and their (positive or negative) impacts on society are not sufficiently discussed in the public and not sufficiently addressed by AI research, despite the striking importance to understand and predict the aggregated outcomes of sociotechnical AI-based systems and related complex social processes, as well as how to avoid their harmful effects. Such effects are a source of a whole new set of explainability, accountability, and trustworthiness issues, even assuming that we can solve those problems for an individual machine-learning-based AI system. Therefore, we cannot concentrate solely on making individual citizens or institutions more aware and capable of making informed decisions. We also need to study the emerging network effects of crowds of intelligent interacting agents, as well as the design of mechanisms for distributed collaboration that push toward the realization of the agreed set of values and objectives at collective level: sustainable mobility in cities, diversity and pluralism in the public debate, and fair distribution of economic resources. We therefore advocate the emergence of societal AI as a new field of investigation of potentially huge impact, requiring the next step ahead in trans-disciplinary integration of AI, data science, social sciences, psychology, network science, and complex systems.

Data Science and AI for social good. Our research vision is centered on leveraging data science and AI to understand complex social challenges and promote social progress. Our SoBigData research infrastructure and community, funded by the EU within H2020 from 2015 till 2024, is moving forward from a starting community of pioneers to a wide and diverse scientific movement, capable of empowering the next generation of responsible social data scientists, engaged in the grand societal challenges laid out in SoBigData's *exploratories*: Societal Debates and Online Misinformation, Sustainable Cities for Citizens, Demography, Economics & Finance 2.0, Migration Studies, Sport Data Science, and Social Impact of Artificial Intelligence.

This in-depth analysis discusses the opportunities and challenges brought by the recent and the foreseeable developments of Artificial Intelligence into online platforms and marketplaces. The paper advocates the importance to support trustworthy, explainable AI (in order to fight discrimination and manipulation, and empower citizens), and societal-aware AI (in order to fight polarisation, monopolistic concentration and excessive inequality, and pursue diversity and openness).

This document was provided by the Policy Department for Economic, Scientific and Quality of Life Policies at the request of the committee on the Internal Market and Consumer Protection (IMCO).

PE 648.791
IP/A/IMCO/2020-15
Print ISBN 978-92-846-6551-8 | doi:10.2861/2605 | QA-04-20-235-EN-C
PDF ISBN 978-92-846-6550-1 | doi:10.2861/222800 | QA-04-20-235-EN-N