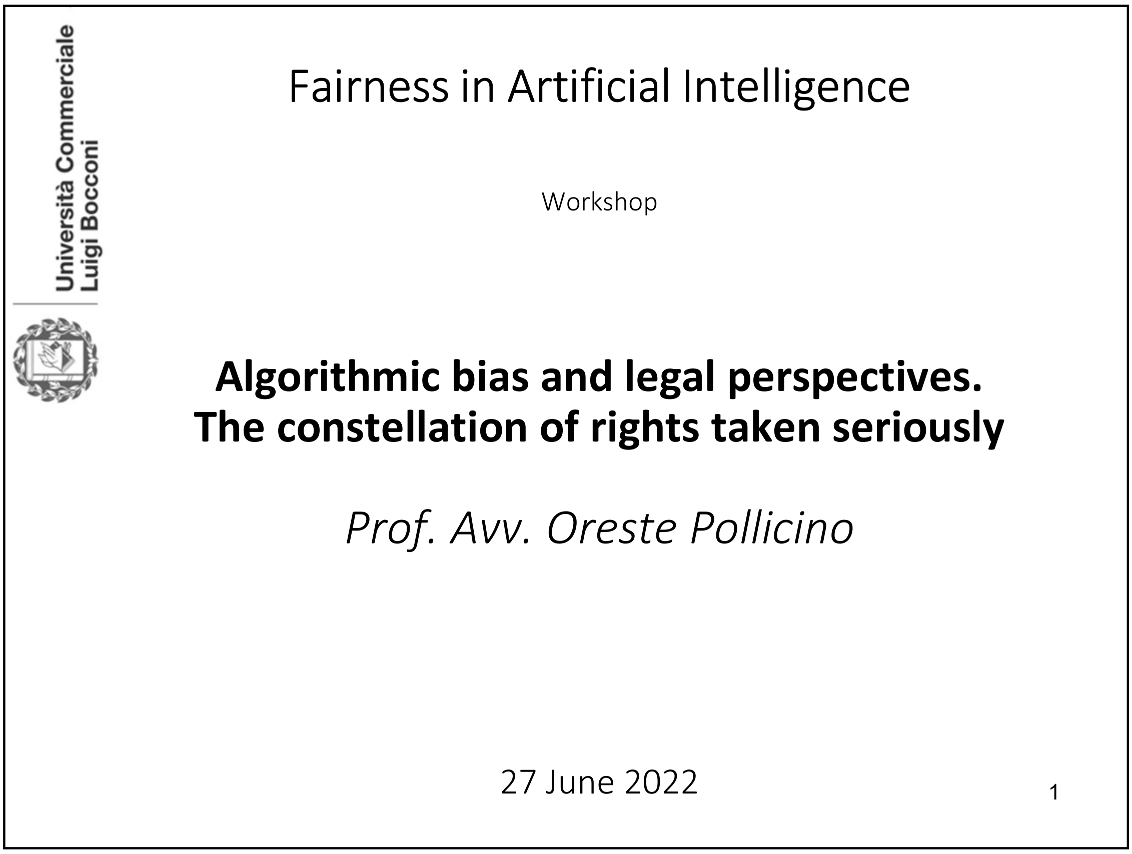


Università Commerciale Luigi Bocconi

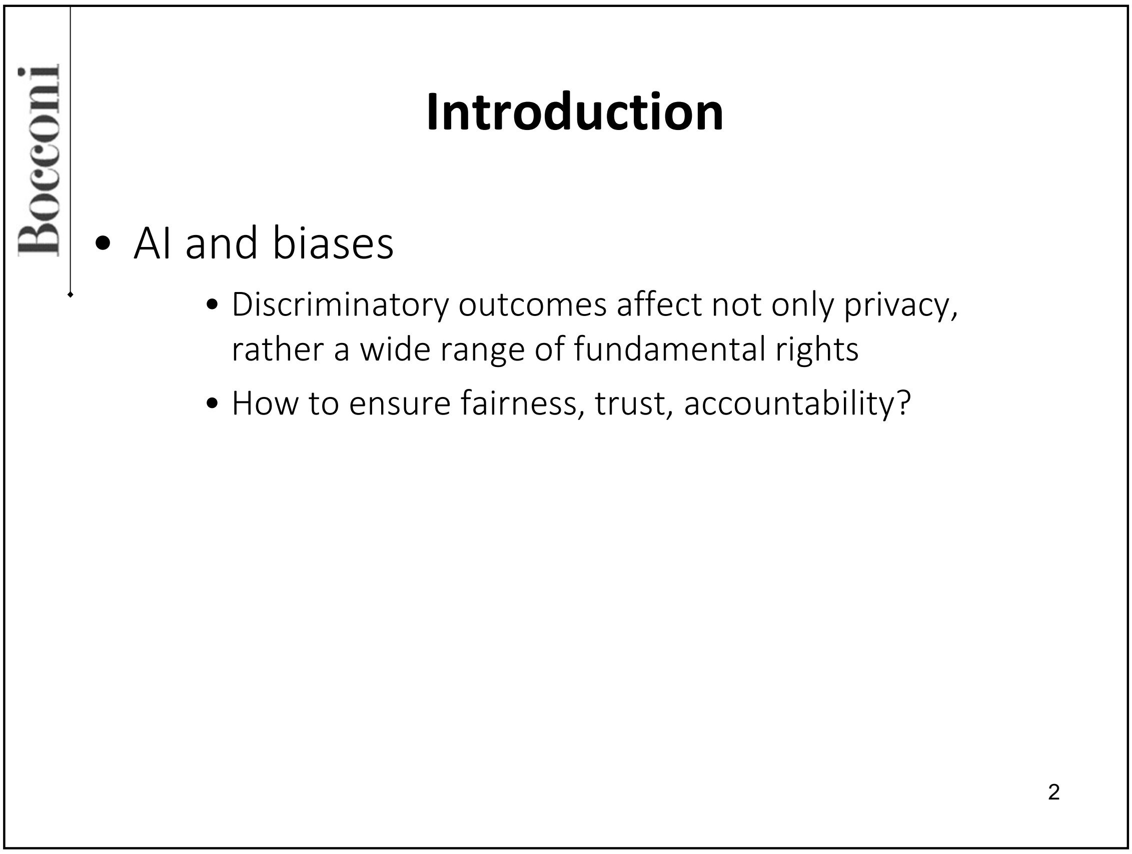
# Fairness in Artificial Intelligence

Workshop

## Algorithmic bias and legal perspectives. The constellation of rights taken seriously

*Prof. Avv. Oreste Pollicino*

27 June 2022

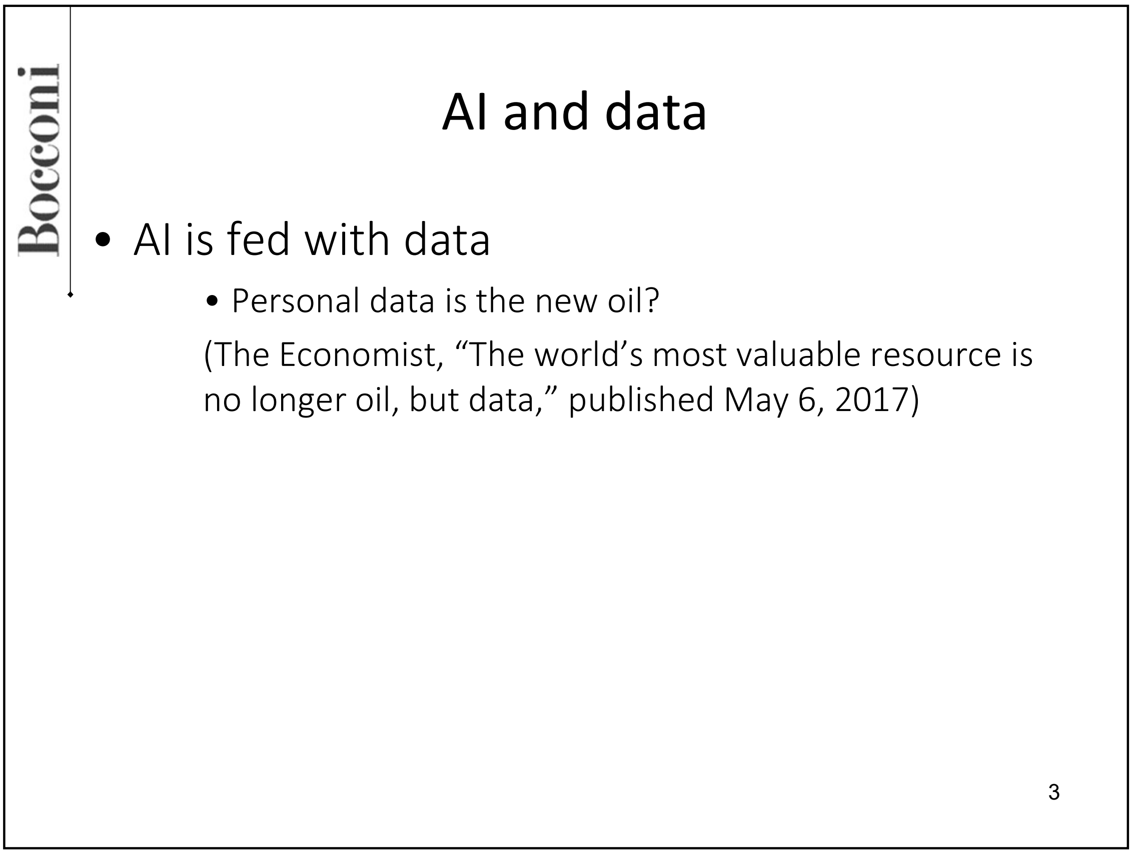


## Introduction

- AI and biases
  - Discriminatory outcomes affect not only privacy, rather a wide range of fundamental rights
  - How to ensure fairness, trust, accountability?

2

The increased use of algorithms to automate decision-making has sparked deep concern that such automated choices may produce discriminatory outcomes. The law has become increasingly interested in issues related to algorithmic biases and decisions, particularly from the perspectives of the collection, use, and processing of personal data. However, technological progress is, on closer inspection, putting the law in a corner from which the jurist is forced to question how AI systems integrate with the rationale of the norms for which they were intended. All without creating a context that can be to the detriment of the citizens themselves. An aspect that seems to capture the lawyer's attention is the risk that the algorithm can produce (and sometimes also reproduce) the social, racial, and gender biases in its decisions.

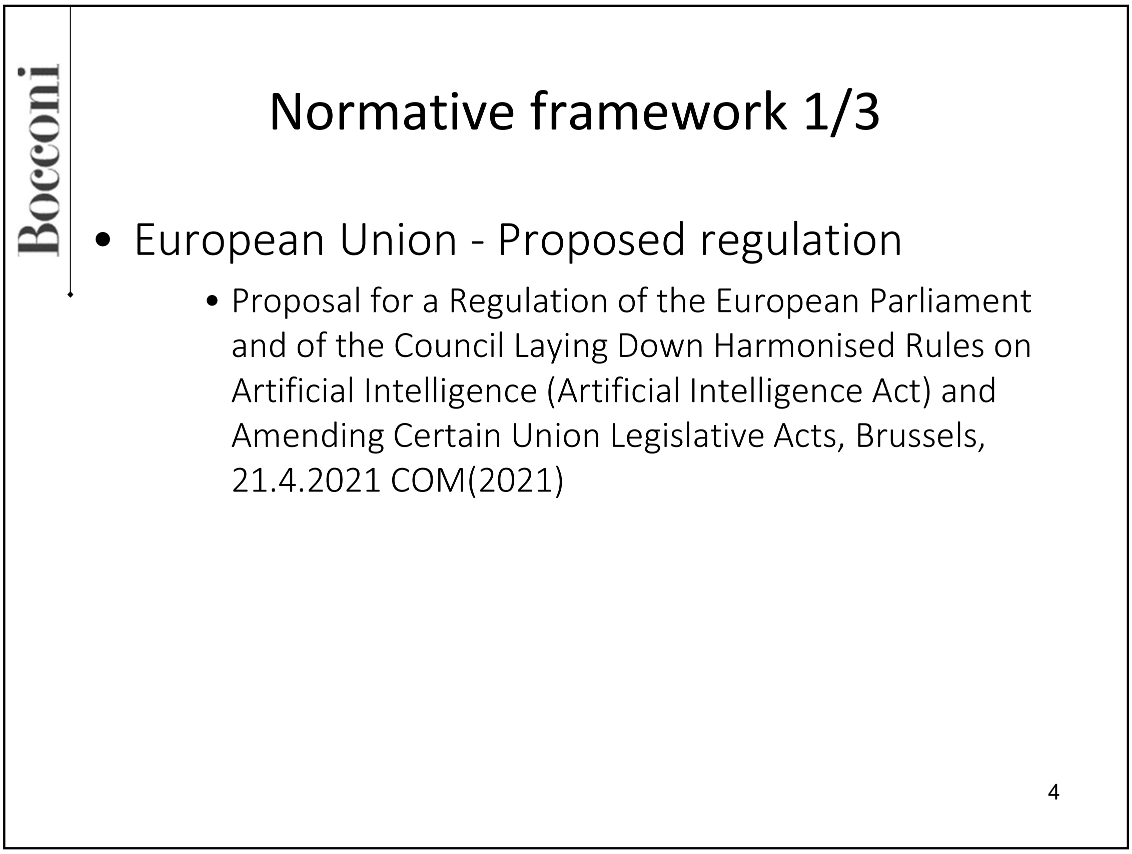


## AI and data

- AI is fed with data
  - Personal data is the new oil?

  (The Economist, “The world’s most valuable resource is no longer oil, but data,” published May 6, 2017)

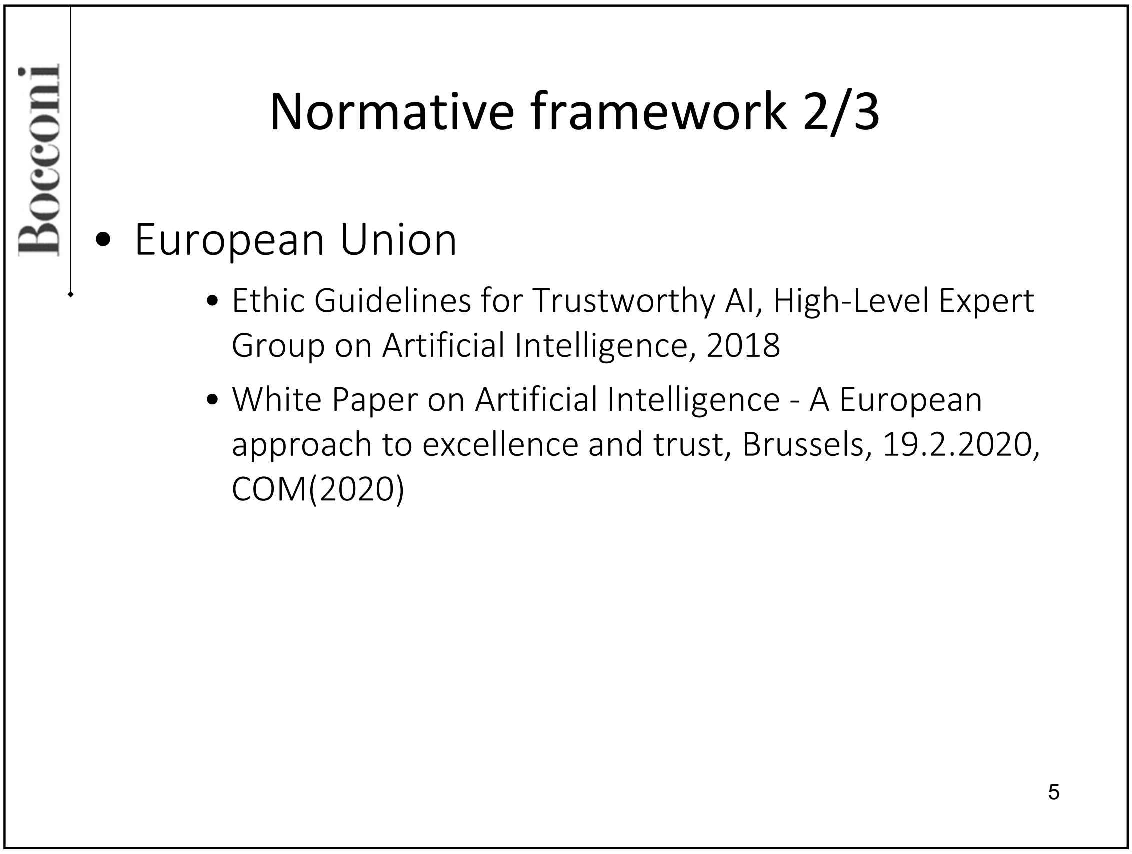
«Data is the new oil» is one of the most used expressions in the field of data protection. However, it is a very wrong assumption. Oil is an exhaustible resource, scarcely available on the planet and increasingly the subject of conflict, and difficult to acquire. Data are all the opposite and, mainly, are progressively growing in quantity and quality, 18 partly due to the use of tools that have decreased the distance between the online and offline worlds,  such as IoTs, and facial recognition technologies. While the potential of oil is in sharp decline, the prospect of data, on the other hand, is growing up and seems to be potentially perpetual.



## Normative framework 1/3

- European Union - Proposed regulation
  - Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, Brussels, 21.4.2021 COM(2021)

Around the world, regulatory proposals are emerging to regulate artificial intelligence. Particularly, the European Union, since last April 2021, has been working on the European approach to AI with the proposal for a Regulation known as AI Act. This regulation build upon GDPR data governance and map AI systems into four risk categories. The lowest risk categories self-regulate with transparency obligations. The highest risk categories require first-party or third-party assessments enforced by national authorities.



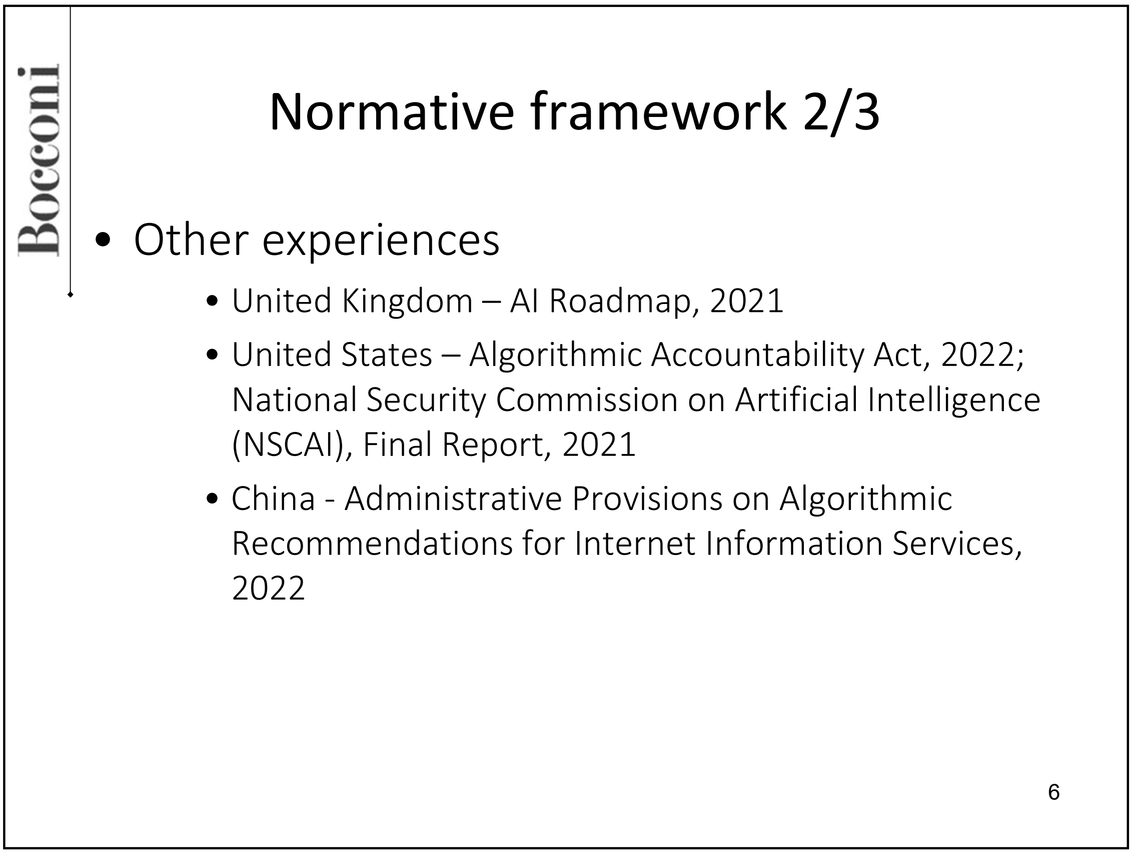
## Normative framework 2/3

- European Union
  - Ethic Guidelines for Trustworthy AI, High-Level Expert Group on Artificial Intelligence, 2018
  - White Paper on Artificial Intelligence - A European approach to excellence and trust, Brussels, 19.2.2020, COM(2020)

5

Also, the Commission and the EU High-level expert group on AI both stress that the allocation of functions across humans and AI systems should follow human-centric design principles and leave meaningful opportunities for human choice. This has further been highlighted in the proposed EU regulation of AI, through a risk-based approach requiring increased levels of human oversight in higher-risk systems. In a decision-making environment, introducing human oversight of AI-based or algorithmic work processes results in various forms of semi-automated, or hybrid decision-making.

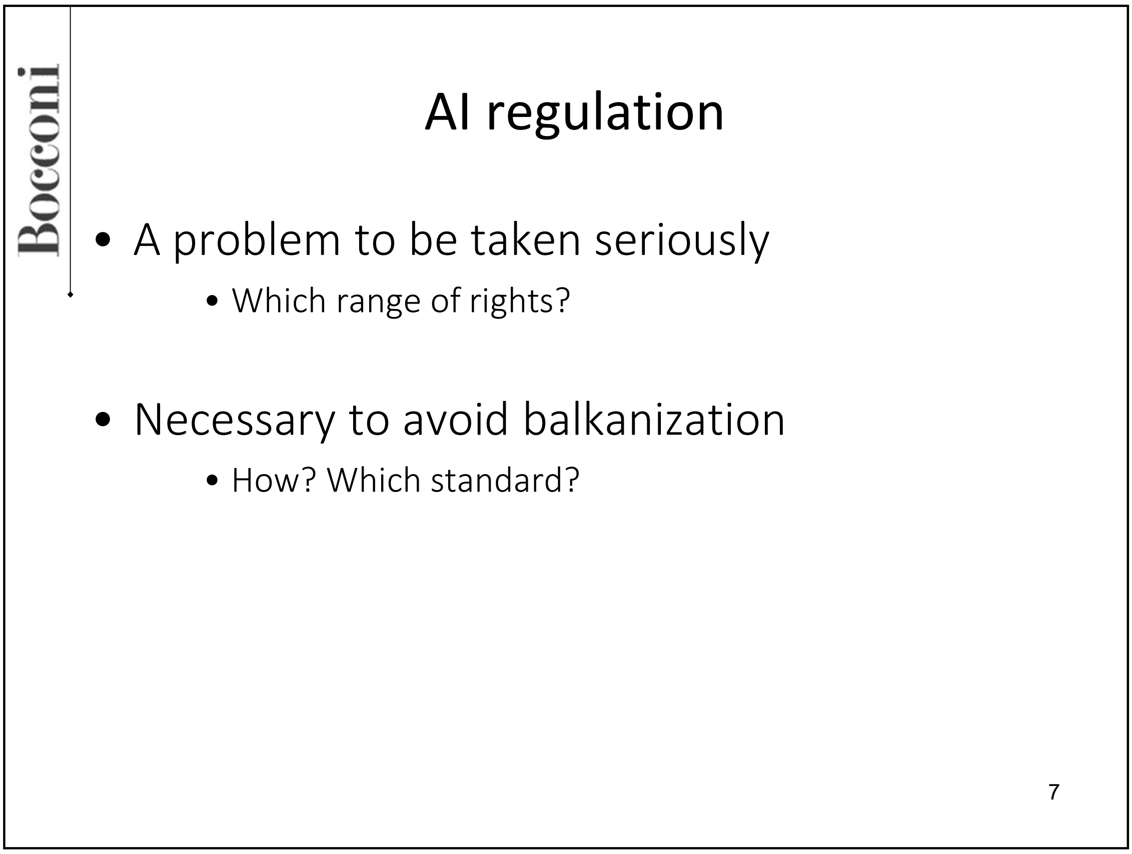
The impetus to implement hybrid decision-making may vary. In some cases, it may be driven by ambitions of increased efficiency where reducing human discretion is a specific goal that cannot fully be realized due to technical or legal constraints. The need for human contextual analysis is well known in other areas, such as online moderation. Still, the sheer scope of the task facing moderators and external pressures calls for further automation. However, in many cases, keeping a human in the loop is a deliberate attempt to maintain human agency and accountability and provide legal safeguards and quality control.



## Normative framework 2/3

- Other experiences
  - United Kingdom – AI Roadmap, 2021
  - United States – Algorithmic Accountability Act, 2022; National Security Commission on Artificial Intelligence (NSCAI), Final Report, 2021
  - China - Administrative Provisions on Algorithmic Recommendations for Internet Information Services, 2022

The United Kingdom AI Council published a roadmap that outlines a sector-specific audit-led regulatory environment, along with principles for the governance of AI systems including open data, AI audits, and FAIR (Findable, Accessible, Interoperable, Reusable) principles. In the United States, two Senators proposed an Algorithmic Accountability Act in 2022 to make responsible corporates that use algorithms and repower the features of the risk assessment tools. The U.S. National Security Commission in the AI report 2021 outlined a market-led regulatory environment, with government focus areas of robust and reliable AI, human-AI teaming, and a standards-led approach to testing, evaluation, and validation. China’s AI development plan is emphasizing the societal responsibility; companies chosen by the Chinese state to be AI champions follow national strategic aims, and state institutions determine the ethical, privacy, and trust frameworks around AI.



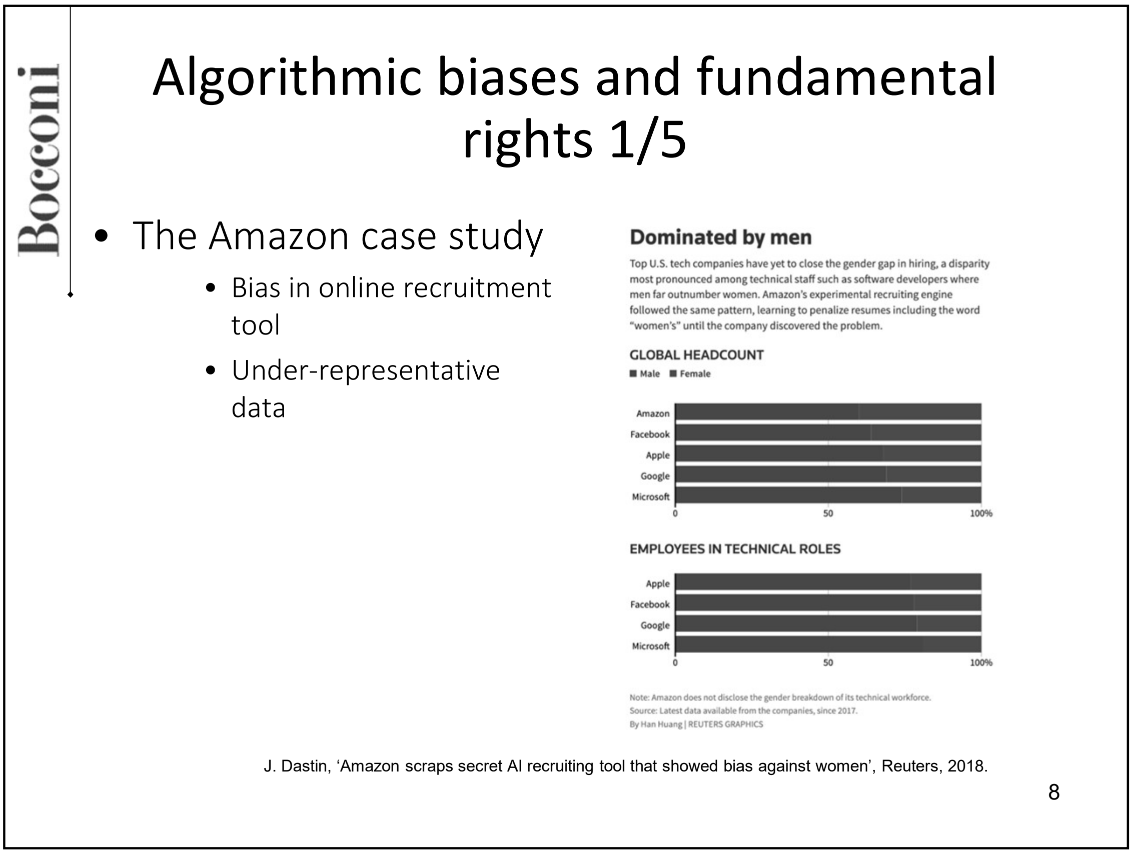
# AI regulation

- A problem to be taken seriously
  - Which range of rights?
- Necessary to avoid balkanization
  - How? Which standard?

To regulate AI means many different things: it may mean individuating legal regime for robotics, introducing regulation of the civil or criminal liability of AI systems, providing limits on the use of algorithms, establishing rules for specific AI applications, such as facial recognition systems, and so on. Add to this the difficulty of identifying the ideal level for such regulation: which is unlikely to be only national but will undoubtedly require supranational or even global coordination. The European alternative is indeed characterized by strong ethical stances around AI applications, for example limiting the autonomy of military AI systems, in direct contrast to China, where autonomy for AI-directed weapons is actively encouraged as part of its military-civil fusion strategy. However, due to the global nature of our contemporary world and, moreover, the source of the data from which is derived the machine’s training and analysis sets, the wide use of automated decision-making systems, it does not seem useful to rely on a dichotomy. Differently, it appears necessary to identify a common ground for the regulation of AI.

Another issue concerns the need to rethink the skills and abilities of a part of the workforce serving the public administration: when and if AI becomes an indispensable tool of work and a technology destined to permeate public action and service delivery, thought needs to be given to the appropriateness of recruiting staff who can govern such new tools, since one of the fundamental conditions for an anthropo-centric development of AI is meta-autonomy, that is, the maintenance of a balance between the human and machine components, rendered through the guarantee of “human in the loop”. Although numerous initiatives to regionally regulate artificial intelligence and its biases in outcomes are emerging in various parts of the globe, two aspects still seem to be overlooked.

In fact, from an *ex ante* perspective, a lot has been said about privacy and data protection, but accountability is a topic that has been scarcely developed in the practice, with a range of different approaches that vary from Europe to the US; secondly, considering the *ex post* targetability of the discriminatory automated outcomes, it is scarcely addressed the problem of empowering individuals with procedural rights that can tackle biased decisions. Thus, this presentation will consider these two and some other aspects that are touched, sometimes without efficacy, by some of the legislative initiatives arising all over the world, in order to identify possible grounds for a paradigm protective of innovation, technology, and fundamental rights that do not reproduce a balkanized model.



Bocconi

# Algorithmic biases and fundamental rights 1/5

- The Amazon case study
  - Bias in online recruitment tool
  - Under-representative data

**Dominated by men**

Top U.S. tech companies have yet to close the gender gap in hiring, a disparity most pronounced among technical staff such as software developers where men far outnumber women. Amazon's experimental recruiting engine followed the same pattern, learning to penalize resumes including the word "women's" until the company discovered the problem.

Note: Amazon does not disclose the gender breakdown of its technical workforce.
Source: Latest data available from the companies, since 2017.
By Han Huang | REUTERS GRAPHICS

J. Dastin, 'Amazon scraps secret AI recruiting tool that showed bias against women', Reuters, 2018.

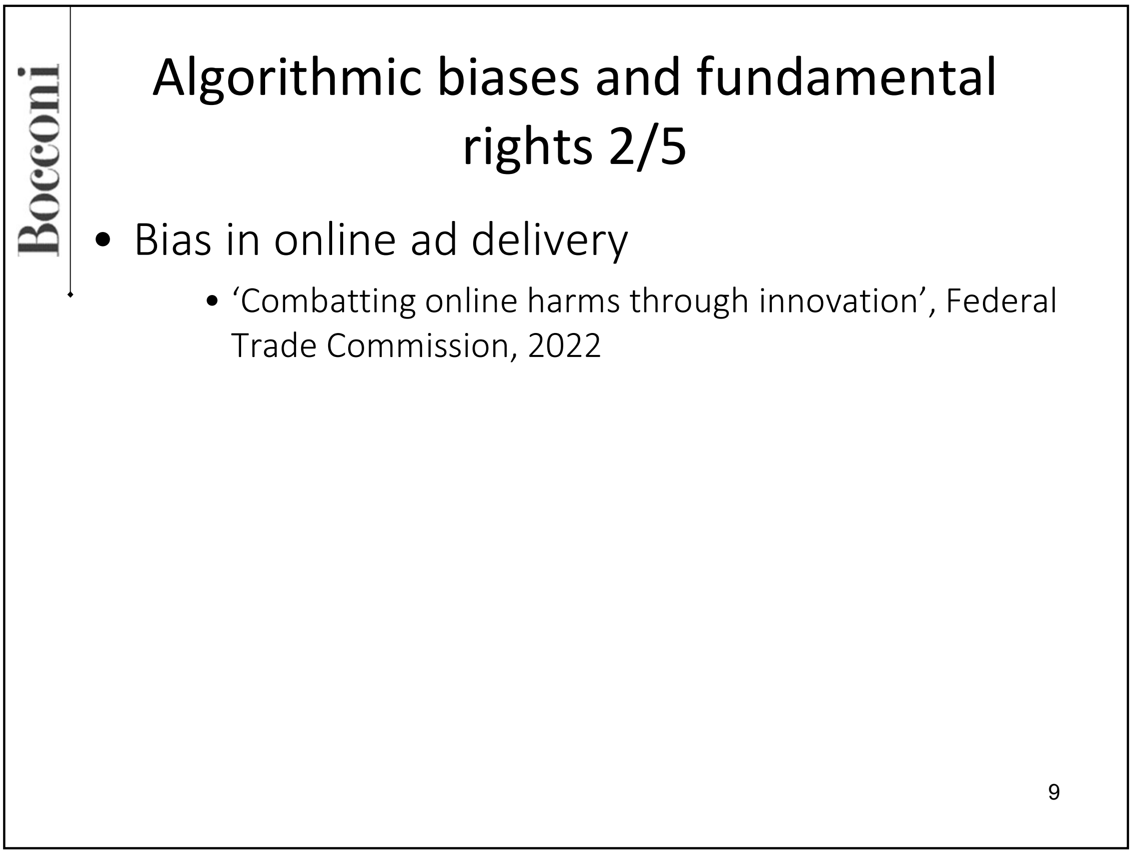
8

In the pre-algorithm world, humans and organizations made decisions in hiring, advertising, criminal sentencing, and lending. These decisions were often governed by federal, state, and local laws that regulated the decision-making processes in terms of fairness, transparency, and equity. Today, some of these decisions are entirely made or influenced by machines whose scale and statistical rigor promise unprecedented efficiencies. Algorithms are harnessing volumes of macro and micro-data to influence decisions affecting people in a range of tasks, from making movie recommendations to helping banks determine the creditworthiness of individuals.

Before understanding the legal reaction, it is believed useful to consider come case studies.

1. Amazon

Amazon, whose global workforce is 60 percent male and where men hold 74 percent of the company's managerial positions, recently discontinued the use of a recruiting algorithm after discovering gender bias. The data that engineers used to create the algorithm were derived from the resumes submitted to Amazon over a 10-year period, which were predominantly from white males. The algorithm was taught to recognize word patterns in the resumes, rather than relevant skill sets, and these data were benchmarked against the company's predominantly male engineering department to determine an applicant's fit. As a result, the AI software penalized any resume that contained the word "women's" in the text and downgraded the resumes of women who attended women’s colleges, resulting in gender bias. Although Amazon scrubbed the data of the particular references that appeared to discriminate against female candidates, there was no guarantee that the algorithm could not find other ways to sort and rank male candidates higher so it was scrapped by the company.

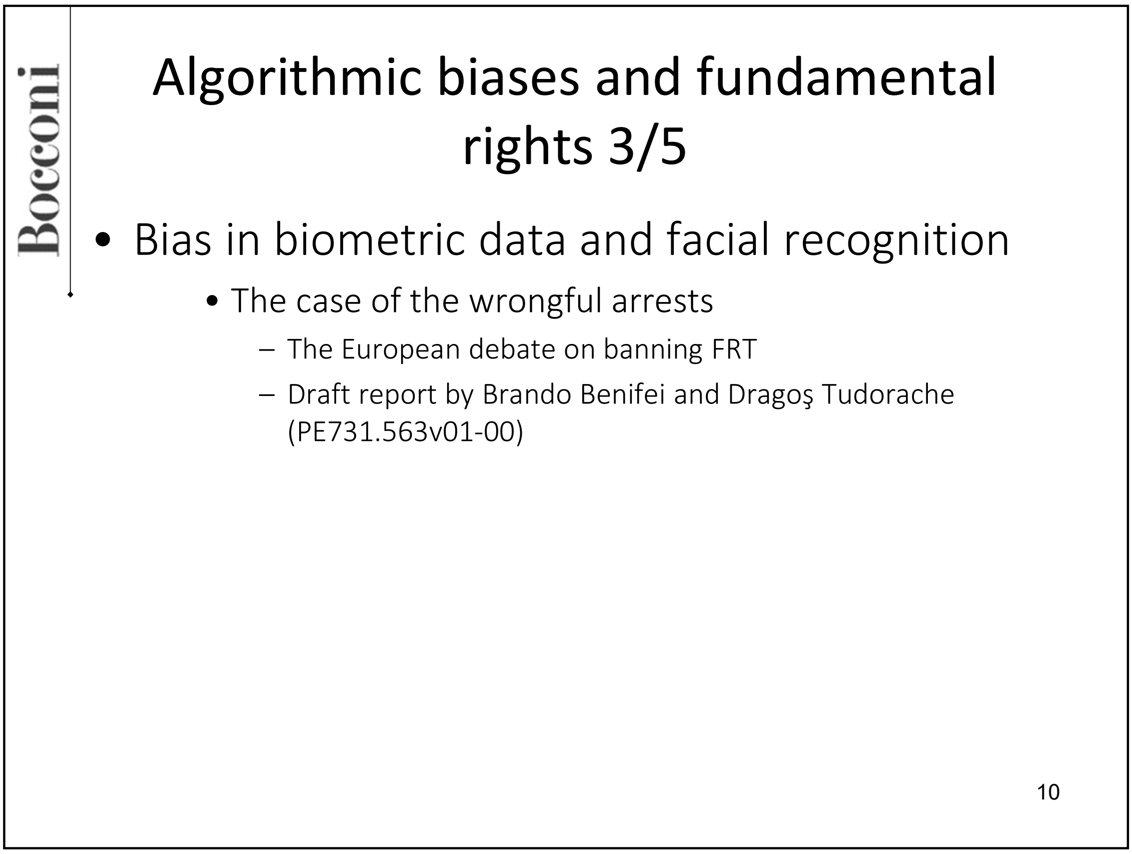


## Algorithmic biases and fundamental rights 2/5

- Bias in online ad delivery
  - ‘Combatting online harms through innovation’, Federal Trade Commission, 2022

9

Secondly, bias also influences online ads. In particular, a study highlighted how online search queries for African American names were more likely to return ads to that person from a service that renders arrest records, as compared to the ad results for white names. The same differential treatment occurred in the micro-targeting of higher-interest credit cards and other financial products when the computer inferred that the subjects were African Americans, despite having similar backgrounds to whites.

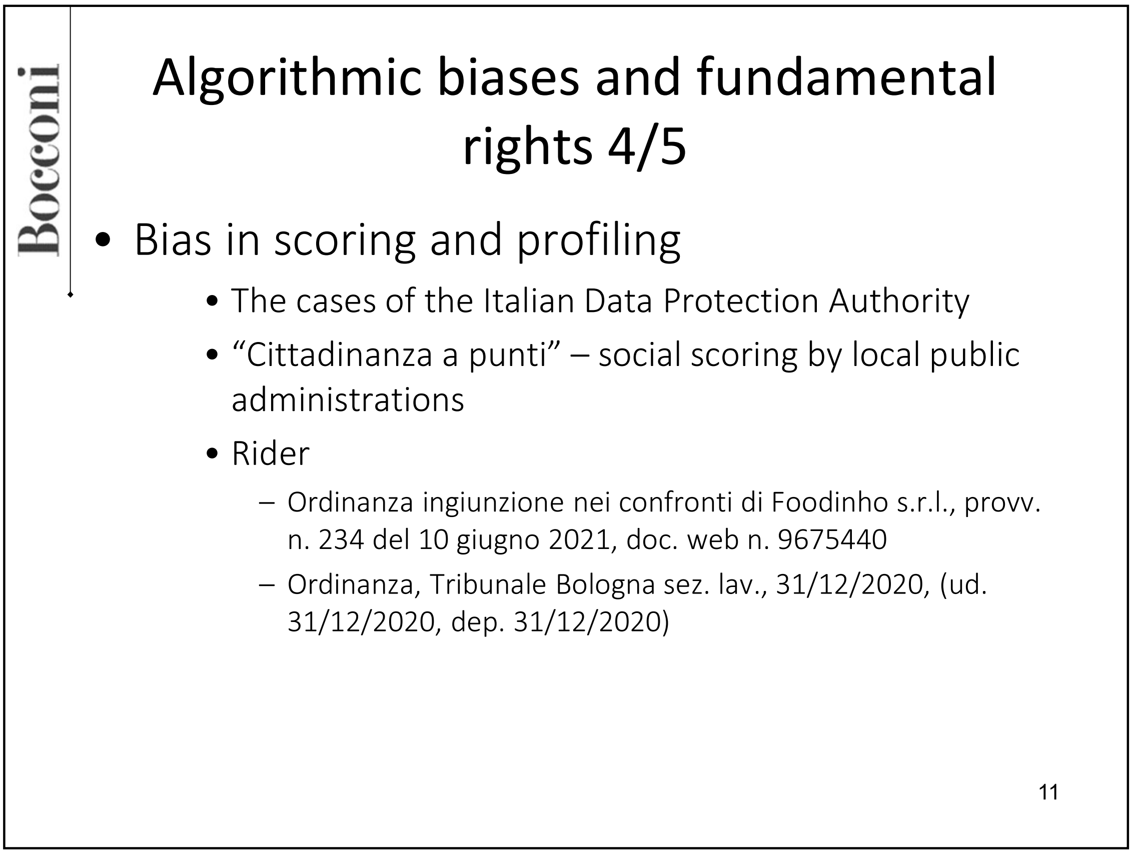


# Algorithmic biases and fundamental rights 3/5

- Bias in biometric data and facial recognition
  - The case of the wrongful arrests
    - The European debate on banning FRT
    - Draft report by Brando Benifei and Dragoş Tudorache (PE731.563v01-00)

Thirdly, another sector highly affected by algorithmic bias regards the employment of facial recognition technology for purposes of surveillance and law enforcement. Many have been cases of wrongful identification of individuals that were wrongfully arrested on the basis of the algorithmic output. These systems, trained on highly sensitive data, such as the biometric ones, fail to recognize darker-skinned complexions. Many studies are demonstrating how most facial recognition training data sets are estimated to be more than 75 percent male and more than 80 percent white.

Amendment by Benifei and Tudorache: https://www.europarl.europa.eu/doceo/document/CJ40-AM-732802_EN.pdf: «the use is admitted only for the purpose of law enforcement must therefore be prohibited, with the exception of border control and in the context of the fight against terrorism», the amendment to recital 19.



## Algorithmic biases and fundamental rights 4/5

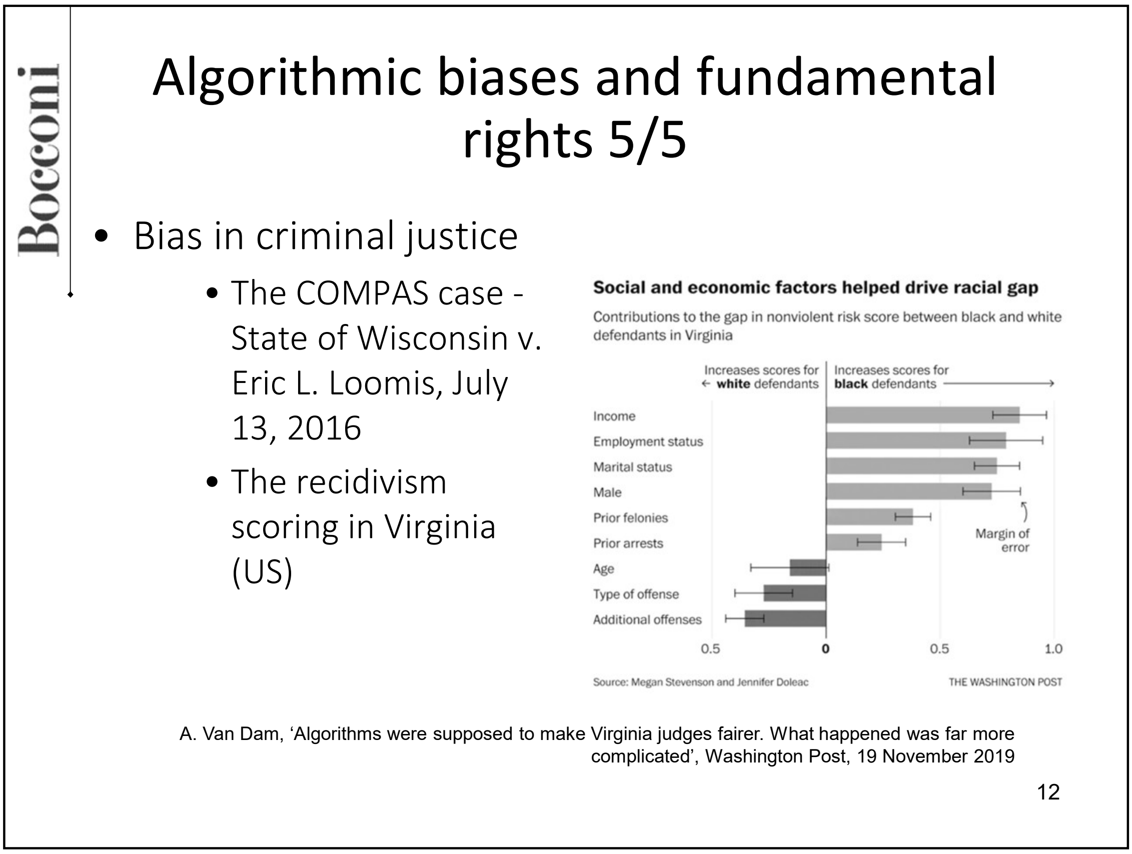
- Bias in scoring and profiling
  - The cases of the Italian Data Protection Authority
  - “Cittadinanza a punti” – social scoring by local public administrations
  - Rider
    - Ordinanza ingiunzione nei confronti di Foodinho s.r.l., provv. n. 234 del 10 giugno 2021, doc. web n. 9675440
    - Ordinanza, Tribunale Bologna sez. lav., 31/12/2020, (ud. 31/12/2020, dep. 31/12/2020)

11

Another two examples of biases and discrimination are concerning two latest initiatives of the Italian Data Protection Authority.

1) The first one concerns “Il Progetto Pollicino”, a project pursued by some local public administration in Italy. These are initiatives aimed at enhancing the virtuous behaviors of model citizens regarding the environment, the IRS, sports, and other areas. The projects provide for the assignment of scores even with regard to data collections conferred voluntarily by data subjects, and the Garante had to intervene because of the risks related to profiling mechanisms that involve “scored citizenship” and which may result in negative legal consequences on the rights and freedoms of data subjects, including the most vulnerable. Precisely, such automated social scoring processes may run counter to the founding principles of the GDPR, starting with respect for human dignity. But not only that, said profiling mechanisms could lead to a point ranking of citizens risks infringing on the fundamental rights and freedoms of especially the most fragile and vulnerable citizens, and therefore most in need of effective protection, by implementing the use of automated public decision-making in areas of the world that are still too gray.
2) Another topic that interested a lot the Data Protection Authority regards the **riders**. The Authority found a number of serious wrongdoings, particularly regarding the algorithms used to manage workers. For example, the company had not adequately informed workers about the operation of the system and did not ensure guarantees about the accuracy and correctness of the results of the algorithmic systems used to evaluate riders. Nor did it ensure procedures to protect the right to obtain human input, voice their opinions, and

challenge decisions made through the use of the algorithms in question, including the exclusion of some riders from work opportunities. In addition, following a collective discrimination complaint filed by Filcams CGIL, Nidil CGIL, Filt Cgil against Deliveroo Italia S.R.L, the Tribunale of Bologna delivered a decision finding that the riders were discriminated. The indirect discrimination consisted of reserving the same treatment for different situations. The Court noted that the profiling system, based on the two parameters of reliability and participation, in treating, in the same way, those who do not participate in the booked session for futile reasons and those who do not participate because they are striking (or because they are sick, have a disability, or assist a handicapped person or a sick minor, etc.) concretely discriminates against the latter, possibly marginalizing them from the priority group and thus significantly reducing their future opportunities to access work.



# Algorithmic biases and fundamental rights 5/5

- Bias in criminal justice
  - The COMPAS case - State of Wisconsin v. Eric L. Loomis, July 13, 2016
  - The recidivism scoring in Virginia (US)

**Social and economic factors helped drive racial gap**

Contributions to the gap in nonviolent risk score between black and white defendants in Virginia

Source: Megan Stevenson and Jennifer Doleac

THE WASHINGTON POST

A. Van Dam, 'Algorithms were supposed to make Virginia judges fairer. What happened was far more complicated', Washington Post, 19 November 2019

1. Lastly, another sector that was highly influenced by the use of automated decisions was judicial reasoning and the employment of algorithms in criminal justice to calculate the risk of recidivism. In a controversial 2016 ruling, the Wisconsin Supreme Court (State of Wisconsin v. Eric L. Loomis, July 13, 2016) ruled on the appeal of Mr. Eric L. Loomis, whose six-year prison sentence had been imposed by the La Crosse Circuit Court. In determining the sentence, the judges had considered the results developed by the COMPAS (Correctional offender management profiling for alternative sanctions) program owned by the Northpointe (now Equivant) company, according to which Loomis was to be identified as a high-risk recidivist. But let us take a step back. In 2013 when Eric L. Loomis was stopped by police while driving a car used to commit a shooting in the state of Wisconsin, USA. He is charged with five counts, all of which are repeat offenses. At that point, Loomis decided to appeal the sentence, complaining that the use of the software results had not guaranteed him a fair trial. However, the Wisconsin Supreme Court ruled against the man, arguing that the decision would have been the same anyway, even without using COMPAS. However, some researchers  pointed out in their analysis of COMPAS that black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism while white defendants were more likely than black defendants to be mistakenly flagged as low risk. In short, according to the COMPAS software, Hispanic or black people were at higher risk of recidivism than white people, who, in contrast, were more likely than black defendants to be flagged as low-risk individuals. It should be noted, however, that the software made calculations based on factors such as age, gender, and criminal history. No

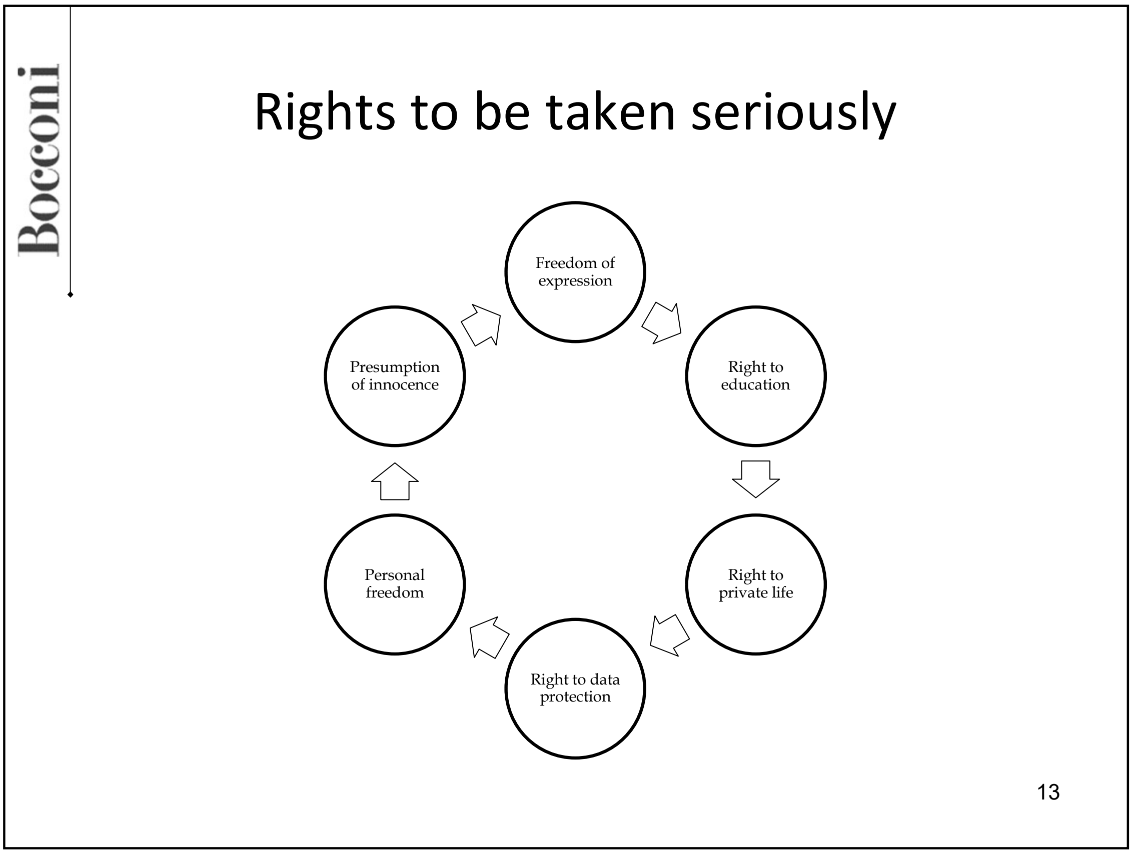
reference, therefore, to the subject's ethnic background, yet, as scholars noted, although the data used by COMPAS do not include an individual's race, other aspects of the data may be correlated with the race, which can lead to racial disparities in the predictions.

After accepting the man's guilty plea, the court ordered a Presentence Investigation Report (PSI) i.e., a report on the subject's personal history useful for the purpose of determining the severity of the sentence. Also used in the PSI is a software called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), a tool designed to predict, among others, the risk of recidivism. Loomis was sentenced to serve six years in prison. The La Crosse Circuit Court, in determining the sentence, had weighed, among other factors, the results of COMPAS, which presented the individual as a high-risk individual for the community.

2. This aspect is crucial if we consider AI's use in judicial and public administration decisions. It is not problematic tout court the use of such systems; on the contrary, it is the blind trust in the automated decisions that, bypassing human reasoning, can result in the crystallization of an output. An output that is not *per se* the result of judicial reasoning but of a machine, with the result of depowering the individual of the power to appeal: one of the crucial rights protected by democratic states.

In the state of Virginia, US, the judiciary started to deploy algorithms with the aim to remove some of the guesswork from judges' sentencing decisions by assigning a simple risk score to defendants. The metrics included data such as offense type, age, prior convictions, and employment status. Larceny scores higher than drug offenses, men score higher than women, and unmarried folks score higher than their married peers. Judges were supposed to use risk scores to identify felons who were least likely to re-offend and either give them shorter jail sentences or send them to a program such as probation or substance-abuse treatment. Therefore, a study conducted by scholars considered together with the result of judges and algorithmic assessment. They underscored that the judges followed the algorithm's suggestions a bit less than half of the time. However, the effect was that people whom the algorithm deemed high-risk received longer sentences than they would have, and candidates assessed as low risk got shorter ones. The two adjustments offset each other, so the overall numbers didn't change, but the interaction between algorithmic sentencing recommendations and judges' discretion nonetheless produced perverse effects. The latter one appeared when considering age as a metric: this is one of the most heavily weighted factors on the risk assessment too. Hence, the judges were more merciful toward young defendants than the algorithm recommended.

These results are not concerning from a point of view of fairness, but also from a critical angle: namely, the reliability of the humans on algorithmic results. As a matter of fact, the technological evolution that has affected public administrations has not been limited to the use of information or communication technologies but has recently gone further by using algorithms within administrative processes.



## Rights to be taken seriously

Freedom of expression

Right to education

Right to private life

Right to data protection

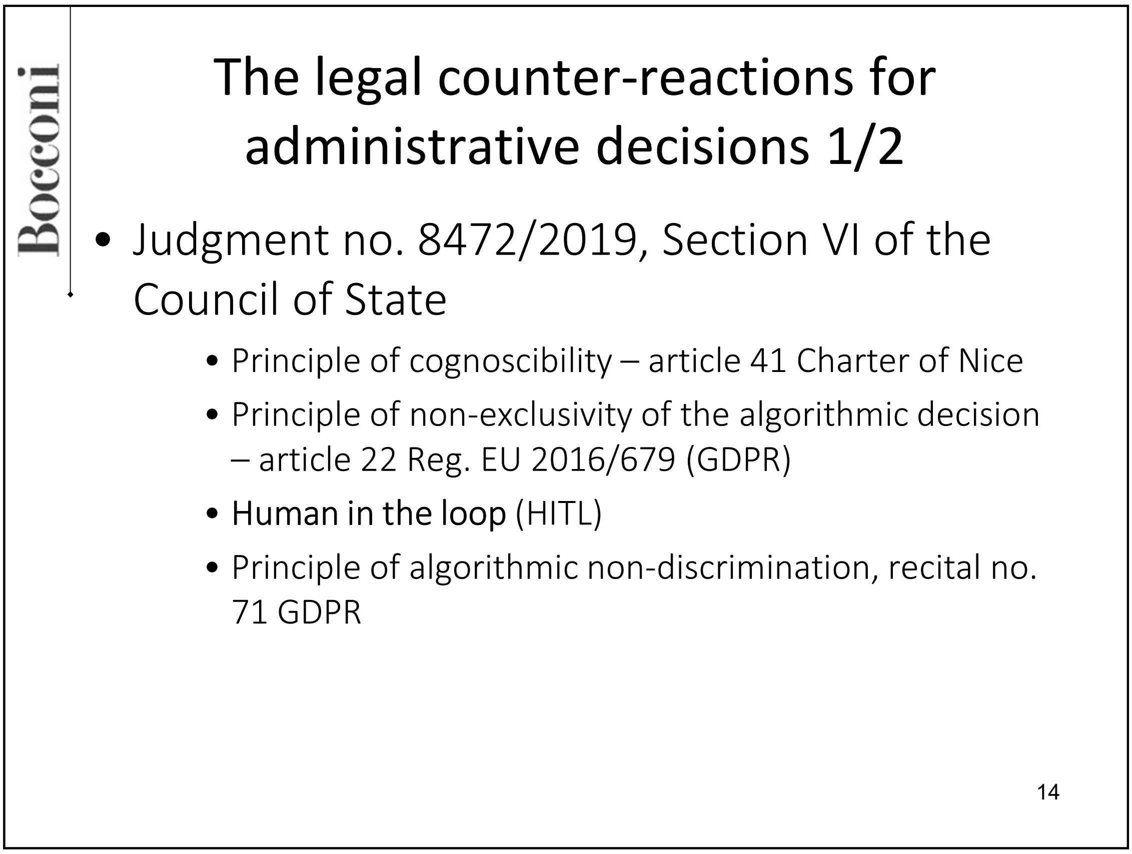
Personal freedom

Presumption of innocence

The examples previously mentioned are noteworthy not only because of the specific technology used to assess recidivism. They are also witnessing a particular feeling of trust in technology that is not always felt on the other side of the Ocean. In Europe, the approach is highly oriented on fundamental rights, and the use of technology is conceived only in the framework of the risks that the former can endure without significantly harming individuals. On the other side, in U.S. it is witnessed a different approach, more oriented toward the output rather than the input. As the COMPAS case showed, the judges evaluated the results instead of looking at how the results were achieved. In other words, the focus was not on the data, how representative they were, and how – even if the racial metric was not explicitly included – even sensitive information can be detected by the algorithm and can lead to a biased result. While it is intuitively appealing to think that an algorithm can be blind to sensitive attributes, this is not always the case. Critics have pointed out that an algorithm may classify information based on online proxies for the sensitive attributes, yielding a bias against a group even without making decisions directly based on one's membership in that group. Thus, it is possible that an algorithm that is completely blind to a sensitive attribute could actually produce the same outcome as one that uses the attribute in a discriminatory manner.

In Europe, too, therefore, the debate about the integration of these systems with the work of public administration and judicial decision-making is wide-ranging. Then, the issues that have recently risen to the attention of the judges do not so much concern the use of information technology supports in the material drafting phase of acts, in particular, administrative ones, but the legitimacy of tools capable of determining the very content of acts. It is, therefore, a different “paradigm of decision-making” regarding which while on the one hand there is unanimous recognition of the benefit in terms of usefulness and effectiveness of the use of AI in support of a human activity, disagreement arises with respect to the possible substitution of AI for human decision-making. In this respect, the Council of State, in Italy, nurtured a trend that is paradigmatic of the European approach, concerning precisely the use of automated systems in administrative decisions This is often called an anti-classification criterion that the algorithm cannot classify based on membership in the protected or sensitive classes.

Online proxies are factors used in the scoring process of an algorithm which are mere stand-ins for protected groups, such as zip code as proxies for race, or height and weight as proxies for gender. They are often linked to algorithms, and they produce both errors and discriminatory outcomes, such as instances where a zip code is used to determine digital lending decisions or one’s race triggers a disparate outcome.



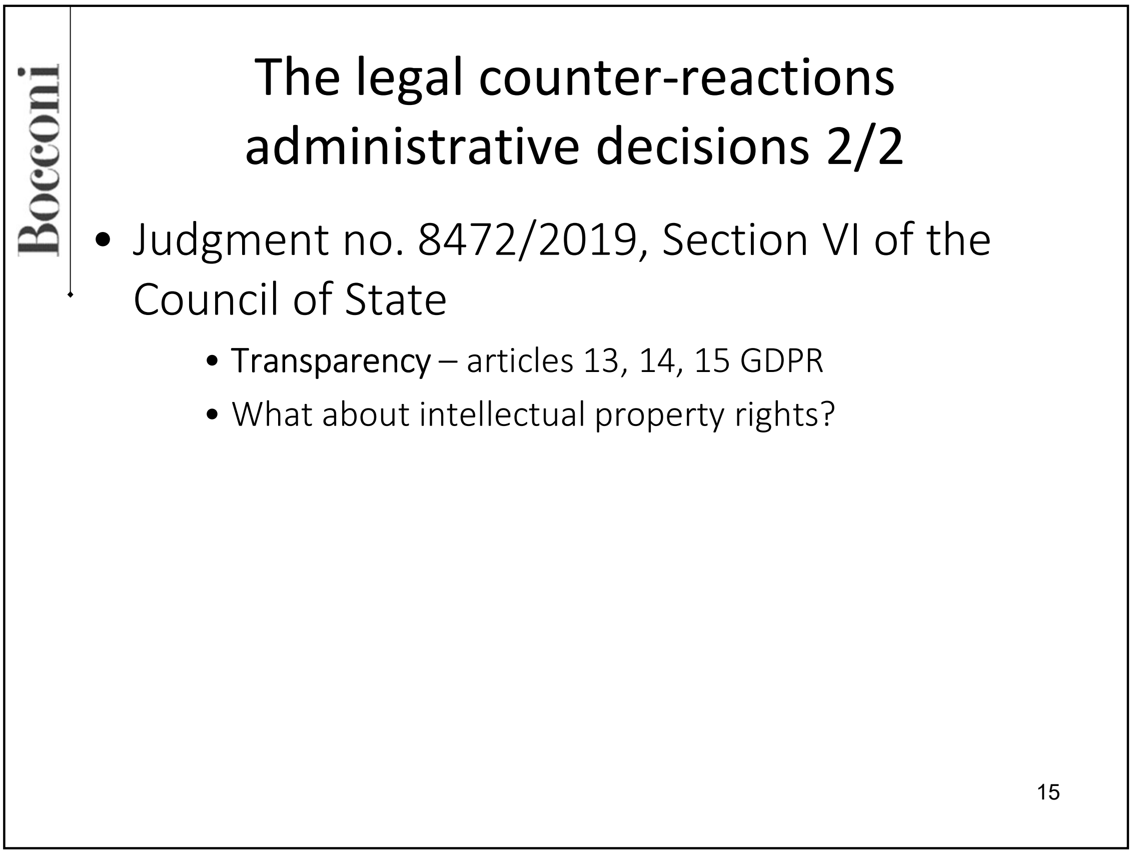
## The legal counter-reactions for administrative decisions 1/2

- Judgment no. 8472/2019, Section VI of the Council of State
  - Principle of cognoscibility – article 41 Charter of Nice
  - Principle of non-exclusivity of the algorithmic decision – article 22 Reg. EU 2016/679 (GDPR)
  - Human in the loop (HITL)
  - Principle of algorithmic non-discrimination, recital no. 71 GDPR

In this respect, the Council of State, in Italy, nurtured a trend that is paradigmatic of the European approach, concerning precisely the use of automated systems in administrative decisions. Particularly, in the Judgment no. 8472/2019, Section VI of the Council of State had the opportunity to expand the set of principles regarding the use of algorithms in administrative activity posited by previous case law, opening up, albeit under two conditions (knowability of the algorithm and the accountability for the decision to the administrative body), the possibility of employing AI also in the context of the discretionary activity of the Administration. The new element introduced by this ruling is the focus on the aspects related to the protection of citizens' fundamental freedoms and, in particular, to the protection of personal data, until that moment all but omitted by administrative judges.

The characteristics of most AI technologies, including opacity (black box effect), complexity, unpredictability, and partially autonomous behavior, make it difficult to verify compliance with rules of existing EU law for the protection of fundamental rights. Those characteristics can also hamper the effective enforcement of law and policy within societies. Lacking the means to verify how a given decision was taken with the involvement of AI, makes it difficult for authorities and individuals to assure whether the relevant rules were respected. Guaranteeing effective access to justice in situations where such decisions may negatively affect individuals and entities, is indeed a challenge. For this reason, according to the Council of State, it is crucial to grant the implementation of such systems only by considering specific safeguards, such as: the principle of cognoscibility, *ex* art. 41 of the

Charter of Nice, which in the field of algorithms is complemented by the principle of comprehensibility, namely, the possibility of receiving meaningful information about the logic used in algorithmic decisions; the principle of non-exclusivity of the algorithmic decision, referable to Article 22 of EU Regulation No. 2016/679 (General Data Protection Regulation), by virtue of which, decisions concerning natural persons must not be based solely on an automated process, if they are likely to produce legal effects that concern or significantly affect the persons to whom they pertain, there always having to be a human contribution for this purpose, in accordance with the so-called model - HITL (human in the loop); the principle of algorithmic non-discrimination, referred to in recital No. 71 of the GDPR, in the mind of which it is appropriate that, in order to ensure the security of personal data and prevent discriminatory effects against natural persons, the data controller should use all the most appropriate measures, from mathematical or statistical procedures to appropriate technical and organizational measures, in order to ensure that data are processed in accordance with the principles of the GDPR, in particular, rectifying factors leading to data inaccuracies and minimizing the risk of errors.

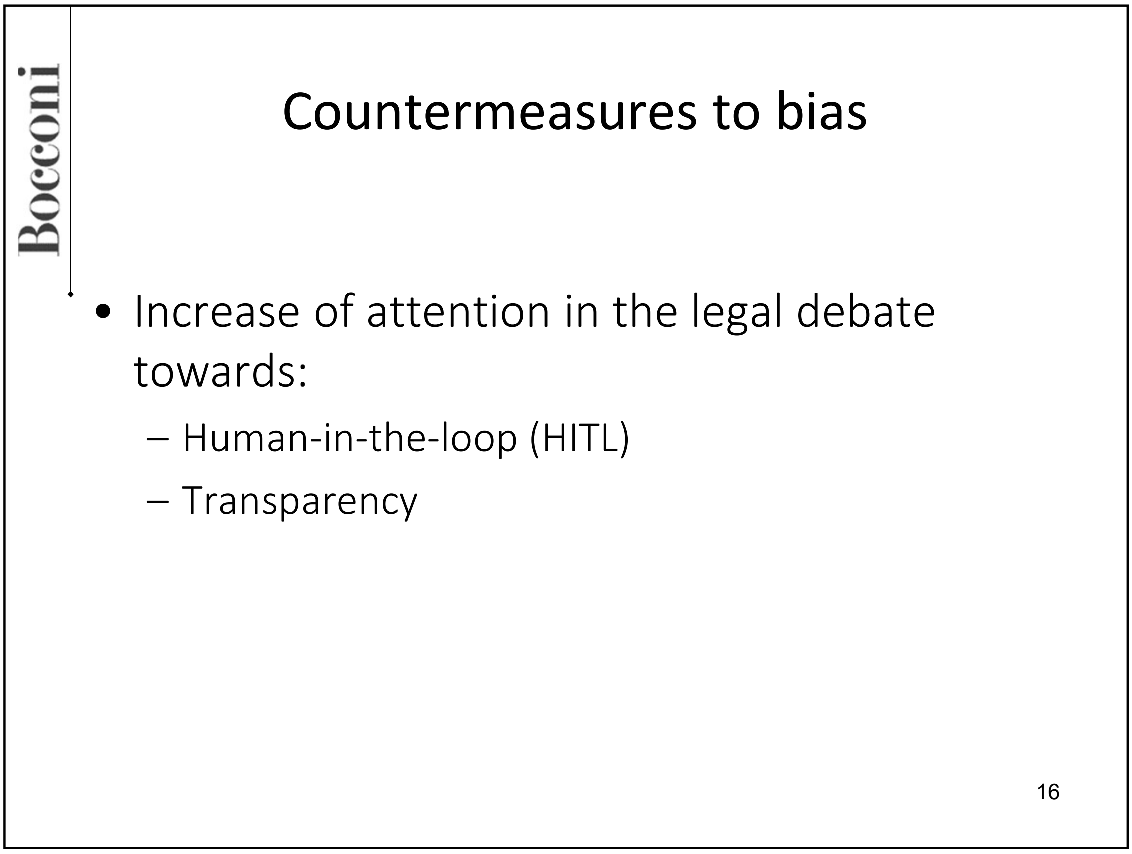


## The legal counter-reactions administrative decisions 2/2

- Judgment no. 8472/2019, Section VI of the Council of State
  - Transparency – articles 13, 14, 15 GDPR
  - What about intellectual property rights?

The issues underlying the automated decisions are manifold: from transparency to the duty to state reasons to the protection of industrial property and public-private actors' interactions. In this regard, according to the Council of State, the confidentiality of the producing companies (or the rules on intellectual property) must yield in the face of the requirements of transparency of the algorithmic decision-making process as well as of the public administration reasoning, cannot assume relevance to the contrary.  It is, after all, Articles 13 and 14 of European Regulation 679/2016 that stipulate that, when faced with an automated procedure, the data subject must be able to know the significant information about the logic utilized, as well as the importance and the consequences envisaged by such processing for the data subject. And this principle of knowability must be understood in terms of the principle of comprehensibility. This guarantee is complemented by Article 15 of the same regulation, which not only provides an actionable right in the data subject (and not merely an obligation in the data controller) but also to acquire information throughout the automated procedure and even after the decision has been adopted. The right of access to information is then accompanied by the other essential guarantee in the face of an automated decision affecting his or her interests, namely, the person's right not to be subjected to a fully automated decision without human involvement, that is, the right to supervision by a human. This means not only that the liability of the decision in the hands of a human must be ensured, but that this human must be able to verify the logicality and legitimacy of the algorithmic decision.

However, the question must be asked: are these guidelines sufficient? What model applies to a global dimension of the digital environment, data origin, and AI outputs? Which paradigm is desirable to emerge: a liberal, output-focused one or a more guarded, risk-oriented one?



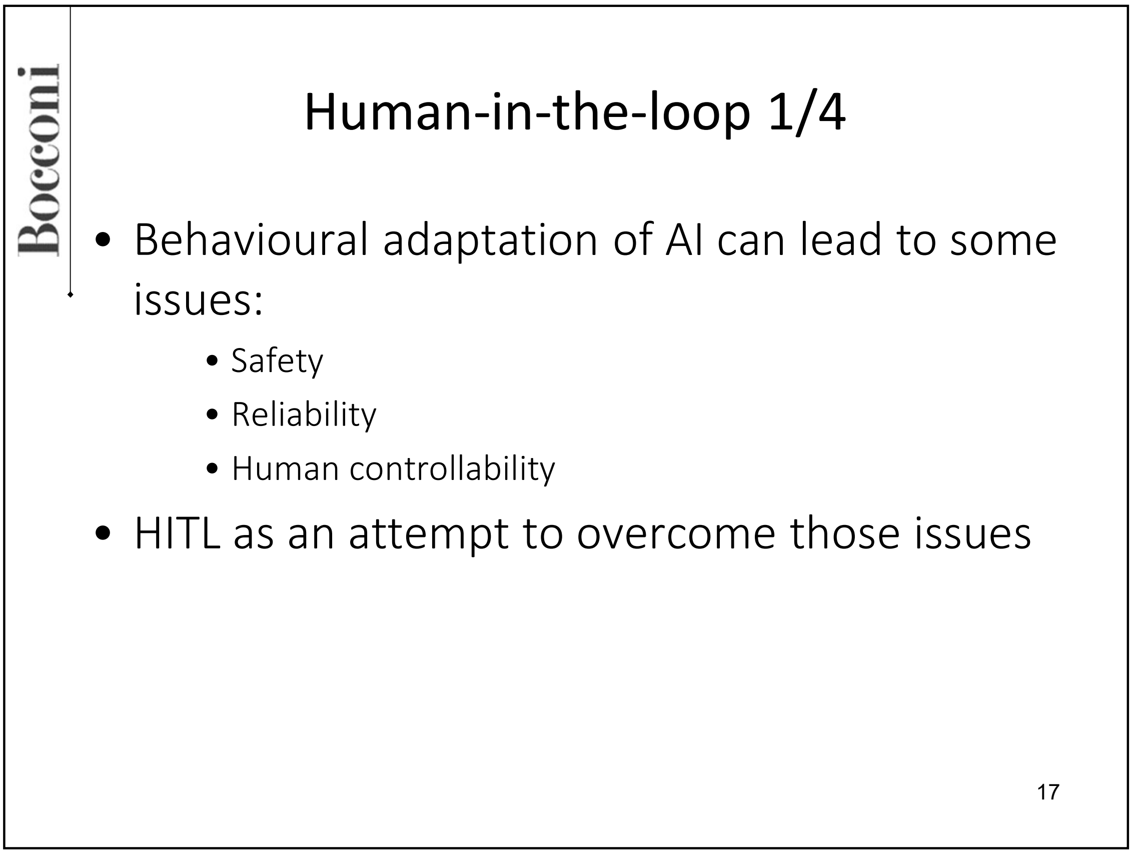
## Countermeasures to bias

- Increase of attention in the legal debate towards:
  - Human-in-the-loop (HITL)
  - Transparency

16

AI systems employ pre-trained learning algorithms that adapt to their users and environment, with learning either pre-trained or allowed to adapt during deployment. Sometimes, the fact that AI can adapt its behavior can cause a perceived risk for safety, reliability, and human controllability. For this reason, legal researchers, as well as policy and lawmakers, have increasingly turned their attention toward the need to guarantee the presence of a “human-in-the-loop”. This remedy aims to make sure that an automated decision has not been made exclusively on the basis of obscure algorithmic processes and that a human being can be held accountable for the decision taken.

Let’s focus before on HITL and then on transparency requirements.

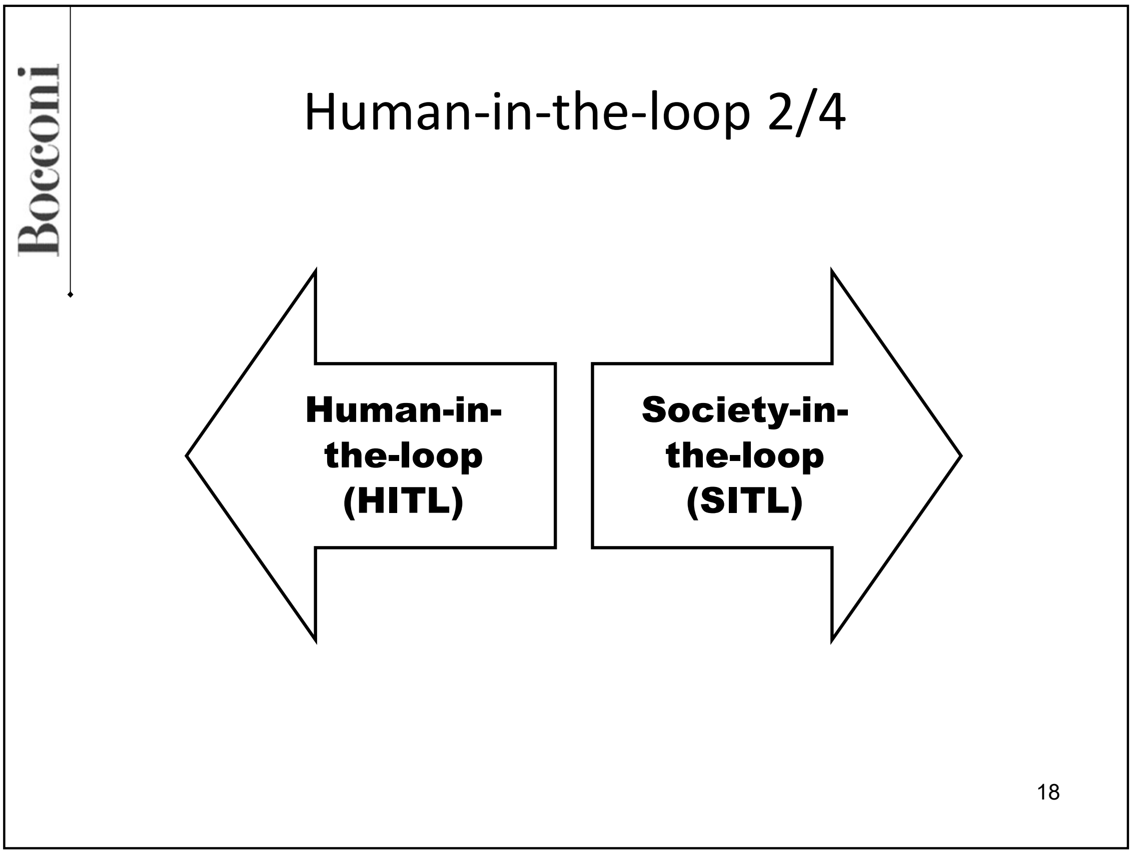


## Human-in-the-loop 1/4

- Behavioural adaptation of AI can lead to some issues:
  - Safety
  - Reliability
  - Human controllability
- HITL as an attempt to overcome those issues

17

Human-in-the-Loop (HITL) systems are grounded in the belief that human-machine teams offer superior results, building trust by inserting human oversight into the AI life cycle. One example is when humans mark false positives in email spam filters. HITL enhances trust in AI by optimizing performance, augmenting data, and increasing safety. It enhances trust by providing transparency and accountability since, unlike many deep learning systems, humans can explain their decisions in natural language. However, if perceived as a top-down oversight by experts, HITL is unlikely to address public trust deficits. Society-in-the-Loop (SITL) seeks broader consensus by extending HITL methods to larger demographics. A growing trend is to add humans into deep learning development and training cycles. Research into HITL is much more evenly spread across the European, U.S., and Chinese regions than work on safe and reliable AI. The European region does differentiate itself with a stronger focus on HITL to promote ethical AI and responsible innovation, as opposed to the U.S. and China, where there is a tighter focus on using HITL to increase AI performance.



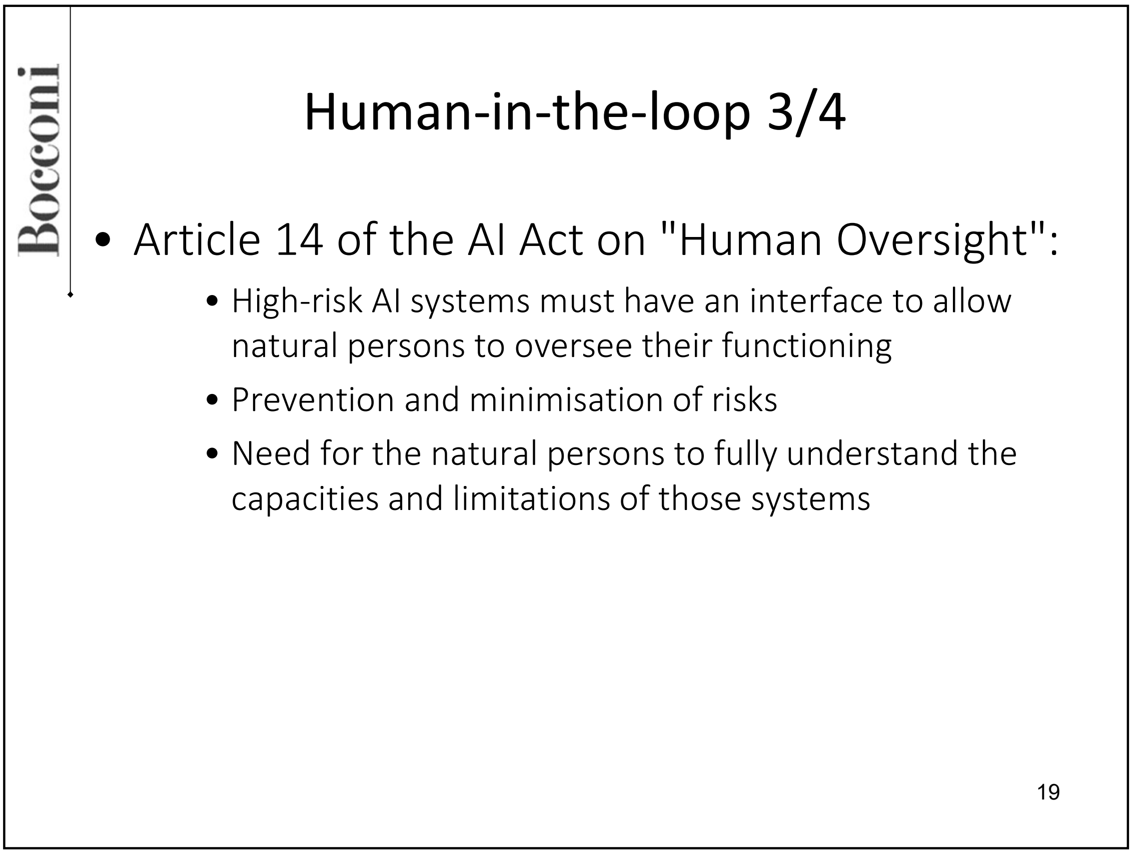
## Human-in-the-loop 2/4

Human-in-the-loop (HITL)

Society-in-the-loop (SITL)

18

Hybrid decision-making can thus be said to operate in-between somewhat counterbalancing ambitions, where the wish for effectivization and automation may require a reduction of human discretion at the same time as legal requirements of maintaining human oversight and agency may necessitate such discretion. Hybrid decision-making comprises a range of systems, including those systems where human agents retain full decision-making autonomy but rely on algorithmic or automated aspects of information gathering, as well as the range of recommendation or recommender systems. It also comprehends those systems where humans are included as a primarily rubber-stamping mechanism, with only nominal control or responsibility for decisions. Also worthwhile is relating hybrid decision-making to the degree to which the decision-making is overseen through humans-in-the-loop (HITL), humans-on-the-loop (HOTL), or humans-in-command (HIC), a terminology both commonly used in research and included in the EU ethics guidelines for trustworthy AI. According to the EU high-level expert group definitions, HITL requires capability for human intervention in every decision cycle of the system. At the same time, HOTL instead aims at human intervention through the design and monitoring of the system.

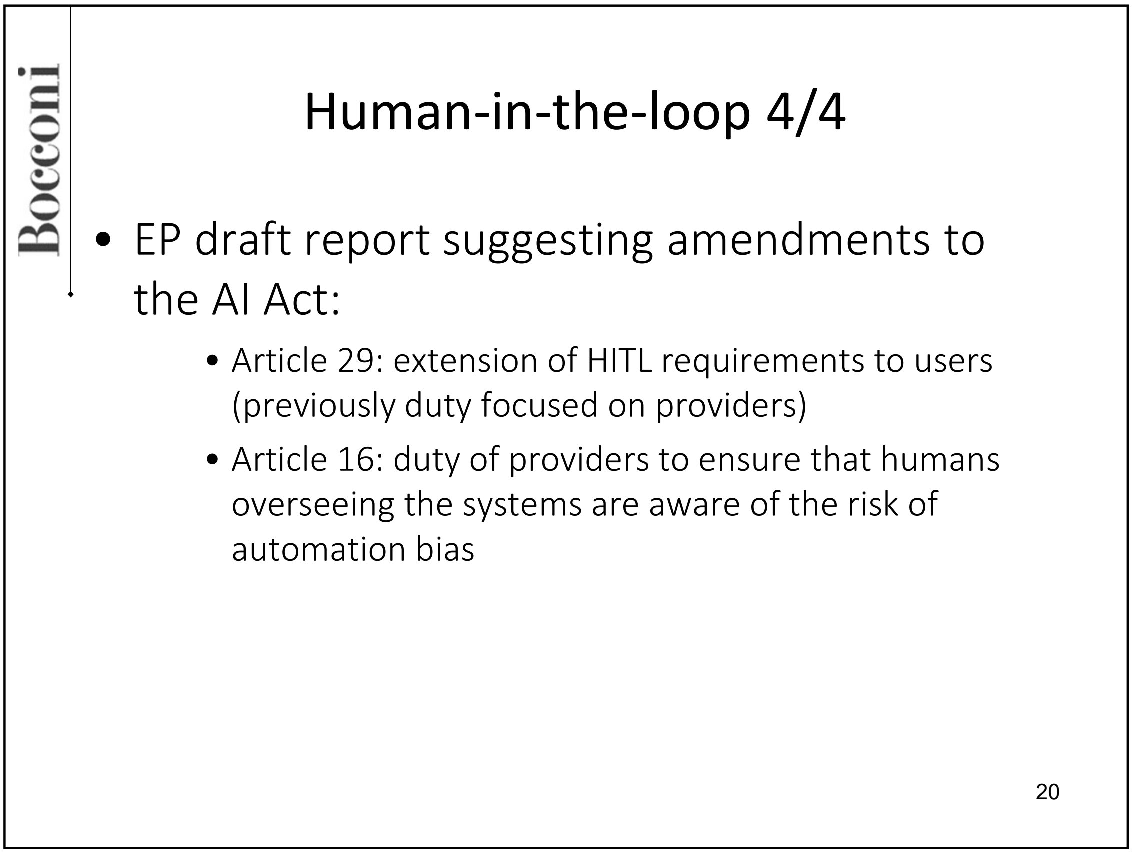


## Human-in-the-loop 3/4

- Article 14 of the AI Act on "Human Oversight":
  - High-risk AI systems must have an interface to allow natural persons to oversee their functioning
  - Prevention and minimisation of risks
  - Need for the natural persons to fully understand the capacities and limitations of those systems

19

The EU AI Act proposal introduced the principle of human oversight for regulating high-risk AI systems. Indeed, Article 14 is specifically dedicated to this purpose, providing that high-risk AI systems «shall be designed and developed in such a way, including with appropriate human-machine interface tools, that natural persons can effectively oversee them during the period in which the AI system is in use». This human oversight shall aim to prevent or minimize the risks to health, safety, or fundamental rights that may emerge when a high-risk AI system is used. The individuals to whom human oversight is assigned will have to be able to: fully understand the capacities and limitations of the high-risk AI system and be able to monitor its operation duly so that signs of anomalies, dysfunctions, and unexpected performance can be detected and addressed as soon as possible; remain aware of the likely tendency of automatically relying or over-relying on the output produced by a high-risk AI system ("automation bias"); be able to correctly interpret the high-risk AI system's production, taking into account, in particular, the characteristics of the system and the interpretation tools and methods available; be able to decide, for any specific situation, not to use the high-risk AI system or otherwise disregard, override or reverse the output of the high-risk AI systems; be able to intervene on the operation of the high-risk AI system or interrupt the system through a "stop" button or a similar procedure.

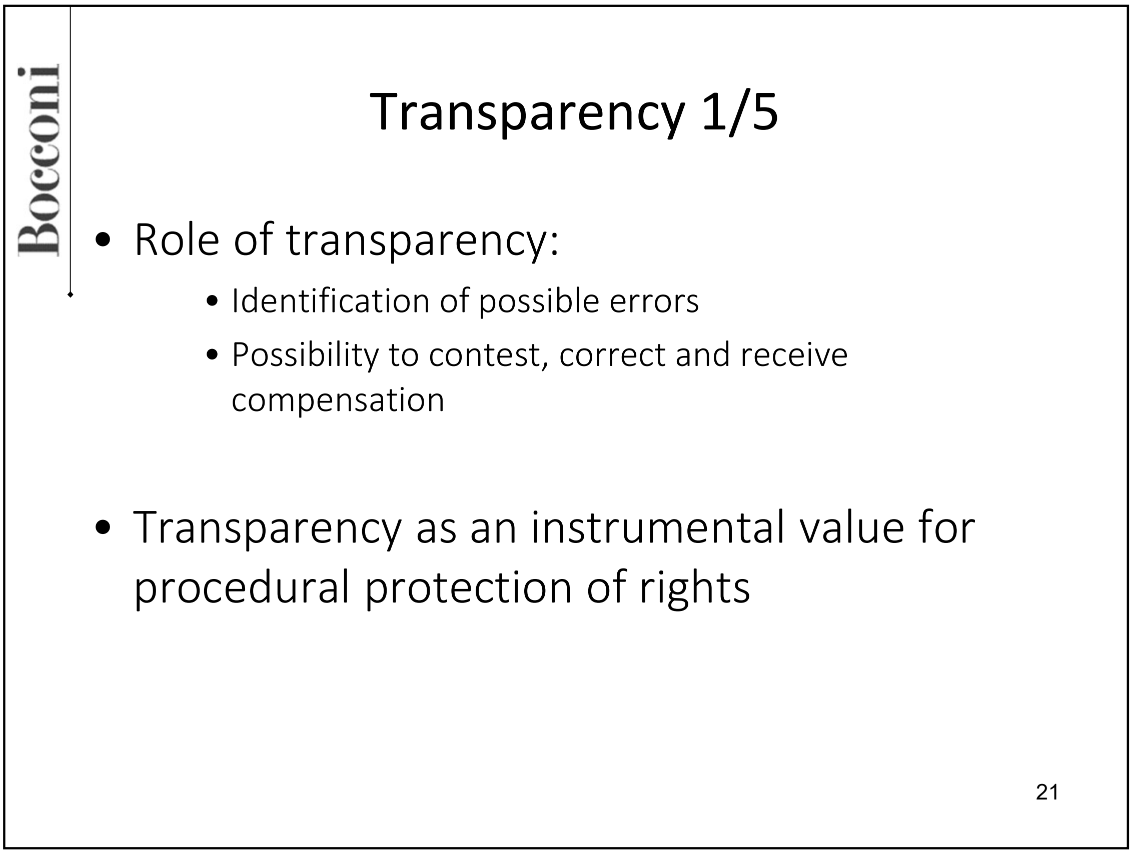


## Human-in-the-loop 4/4

- EP draft report suggesting amendments to the AI Act:
  - Article 29: extension of HITL requirements to users (previously duty focused on providers)
  - Article 16: duty of providers to ensure that humans overseeing the systems are aware of the risk of automation bias

Besides, the role of HITL in the AI Act has become even more relevant in the European Parliament’s Draft Report, suggesting amendments to the Act. In particular, additional provisions have been inserted within Article 29, explicitly dedicated to the duties of users of high-risk AI systems. Most notably, the suggested amendments require that users of those systems comply as well with the human oversight requirements laid down by the Regulation: this is an important point because, previously, the human oversight duties were mainly set at the stage of the system’s development and therefore, mainly concerned the activity of high-risk AI systems providers. Moreover, users will have to ensure that the natural persons assigned to this task are «competent, properly qualified, and trained and have the necessary resources in order to ensure the effective supervision of the system in accordance with Article 14». A similar duty, besides, has also been introduced concerning providers: indeed, providers, too, under the new Article 16(1)(aa), will have to «ensure that natural persons to whom human oversight of high-risk AI systems is assigned are expressly made aware and remain aware of the risk of automation bias».

https://www.europarl.europa.eu/doceo/document/CJ40-PR-731563_EN.pdf.

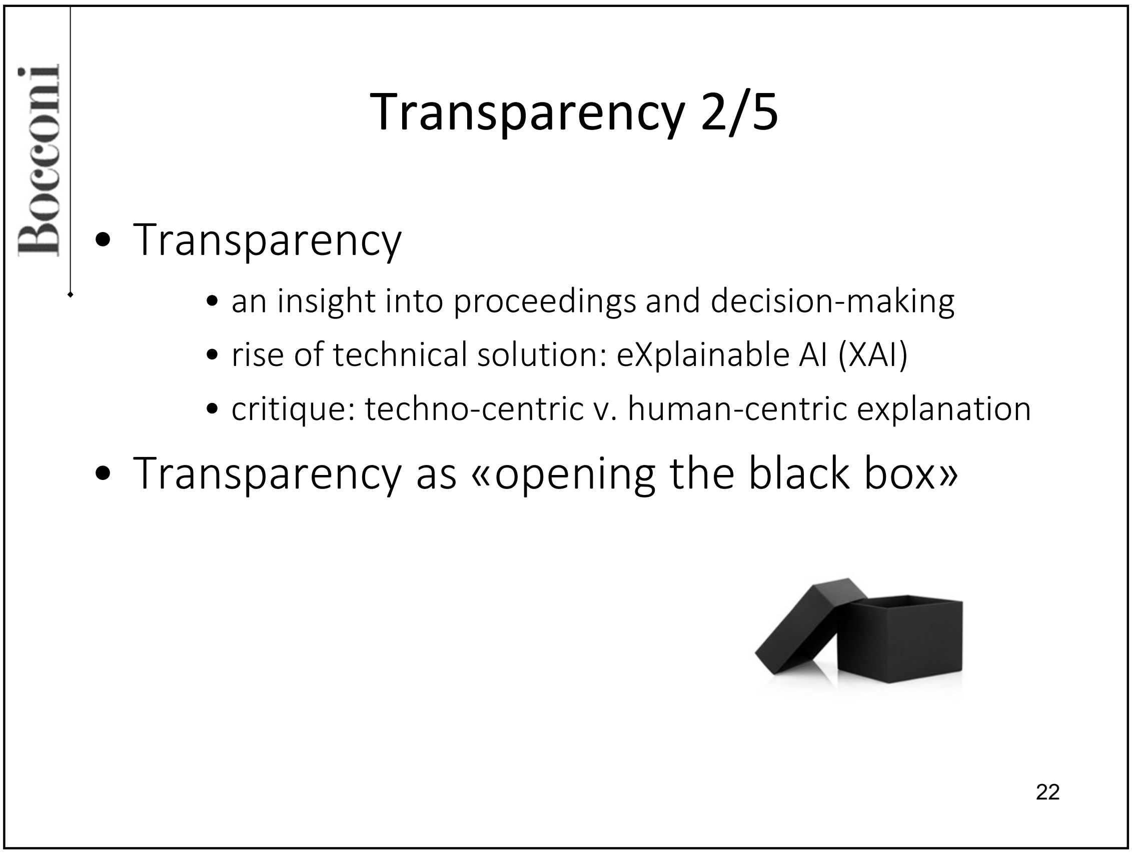


## Transparency 1/5

- Role of transparency:
  - Identification of possible errors
  - Possibility to contest, correct and receive compensation
- Transparency as an instrumental value for procedural protection of rights

21

This duty to give reasons is – aside from being a transparency obligation – thus an important means to facilitate accountability and individual access to justice. Transparency is not an end, but a necessary condition for trust and exercise of procedural rights. Humans require a narrative form of explanation which opposes the binary nature of AI systems' outputs: as human beings, we are not at ease vis-à-vis decisions made through a decisional process we cannot explain nor understand The lack of transparency and explanations means that it is more difficult for individuals to challenge the basis of automated decisions. Transparency is necessary to identify possible errors within and to have the possibility to contest, correct and receive compensation for erroneous decisions. Therefore, transparency is fundamental to ensuring the procedural protection of individuals' rights. Transparency is thus a key value in the relationships between individuals and public institutions, as well as in those between private actors. At the same time, ensuring transparency is not always easy, also because AI systems generally need opacity for a wide array of reasons of private interests (e.g., preservation of trade secrecy). Commonly, transparency is defined as the characteristic of being "easy to see through", or as the quality of openness without secrecy. In the context of law, it can be said to mean an insight into legislative, administrative or judicial proceedings.

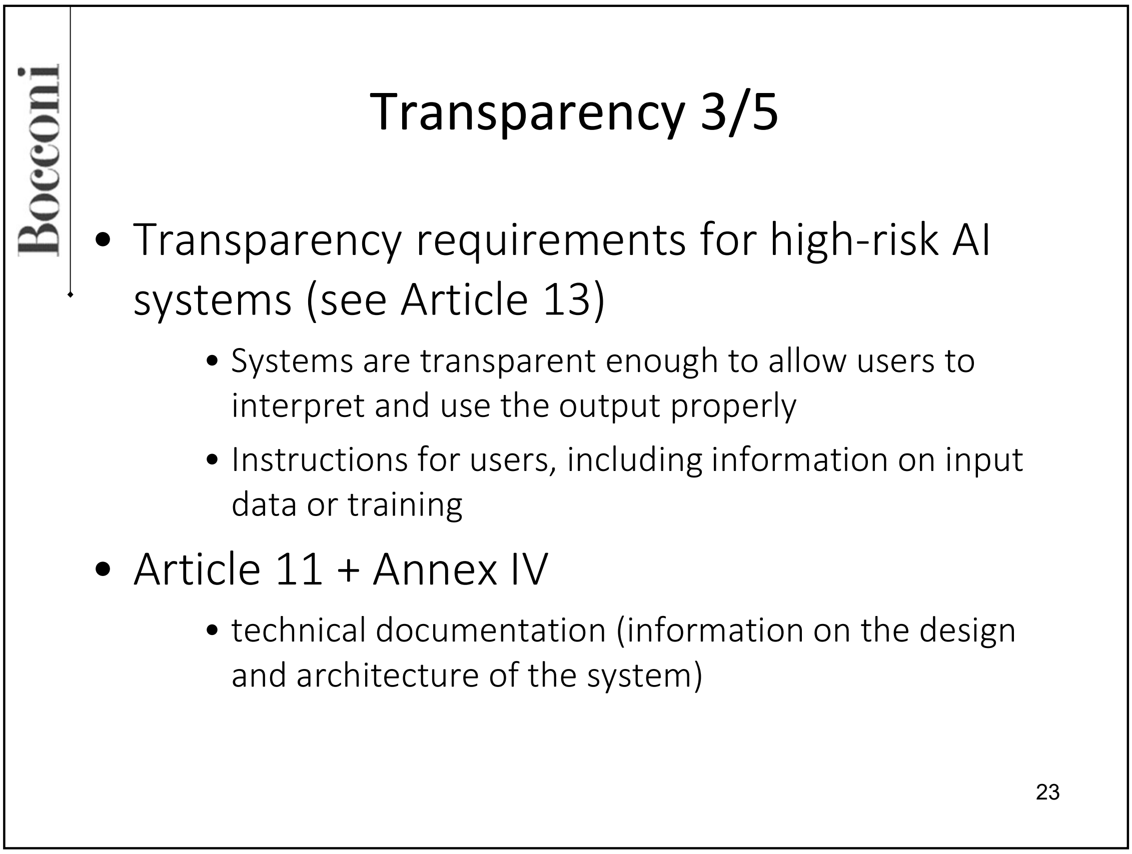


## Transparency 2/5

- Transparency
  - an insight into proceedings and decision-making
  - rise of technical solution: eXplainable AI (XAI)
  - critique: techno-centric v. human-centric explanation
- Transparency as «opening the black box»

Transparency is not an end, but a necessary condition for trust and exercise of procedural rights. Humans require a narrative form of explanation that opposes the binary nature of AI systems' outputs: as human beings, we are not at ease vis-à-vis decisions made through a decisional process we cannot explain nor understand. From a technological point of view, the search for transparency has prompted a development of an entire field of eXplainable AI (XAI) which focuses on designing tools that can enable explanations for the decisions produced by complex autonomous systems. However, XAI as a practice is not without some shortcomings. Most notably, the field has been recently critiqued for its techno-centric view, with limited concerns about the situated needs of its intended audience. Currently, there is a gap between the way dominant algorithm centered XAI approaches work and the way explanations are sought and produced by people: therefore, there has been a growing call for forms of Social Transparency (ST) capable of better responding to the needs of the population.

Many calls for transparency advocate the "opening of the black box" of AI systems, thus allowing to ensure accountability and governance. However, transparency of this kind has limits. Nonetheless, it must also be taken into account that transparency itself can have its own downsides, especially when it is used to 'game' the system to obtain goods or services unfairly. Additionally, transparency obligations are at risk of causing "inadvertent" or "strategic" opacity. In both cases, important information is concealed amidst great quantities of information provided by actors complying with transparency regulations. Sifting through such an impossible amount of information renders transparency unusable



## Transparency 3/5

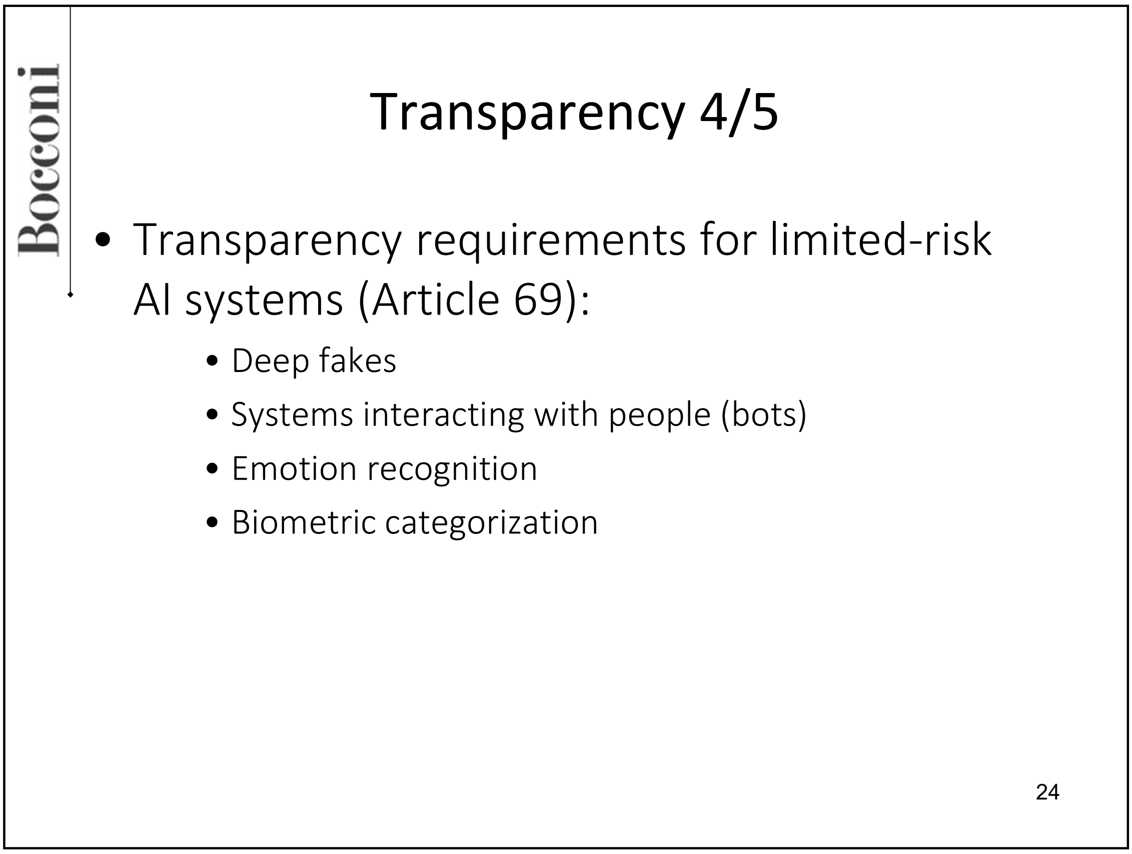
- Transparency requirements for high-risk AI systems (see Article 13)
  - Systems are transparent enough to allow users to interpret and use the output properly
  - Instructions for users, including information on input data or training
- Article 11 + Annex IV
  - technical documentation (information on the design and architecture of the system)

23

Unsurprisingly, AI transparency has been at the center of legal research and law and policy making for some years. The High-level Expert Group on AI (HLEG-AI), in 2019, published its guidelines on Trustworthy AI, identifying 7 key requirements. One of these key requirements was, indeed, transparency: according to the HLEG-AI, «the data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations». Coherently, the 2020 White Paper on AI underscored the role of transparency, highlighting that the lack of transparency «makes it difficult to identify and prove possible breaches of laws, including legal provisions that protect fundamental rights, attribute liability and meet the conditions to claim compensation».

Transparency is also central within the Union's AI Act proposal and informs some provisions concerning both high-risk and non-high-risk AI systems. As regards high-risk AI systems, Article 13 sets some important rules concerning transparency obligations for providers and the provision of relevant information to users of those systems.

First of all, high-risk AI systems must "«be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately». High-risk AI systems must also be accompanied by instructions for use that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to users: most notably, those instructions must include information on the characteristics, capabilities and limitations of performance of the high-risk AI system, including (i) its intended purpose; (ii) its level of accuracy, robustness, and cybersecurity; (iii) known or foreseeable circumstance which may lead to risks to the health and safety or fundamental rights; (iv) its performance as regards the persons or groups of persons on which the system is intended to be used; (v) specifications for the input data, or any other relevant information in terms of the training, validation and testing data sets used, taking into account the intended purpose of the AI system. Moreover, Article 11, in conjunction with Annex IV of the proposal, obliges providers to draft detailed technical documentation containing information concerning, among the rest, the elements characterizing the functioning of the AI system itself, including the “design specifications of the system” and “the description of the system architecture explaining how software components build on or feed into each other and integrate into the overall processing”.



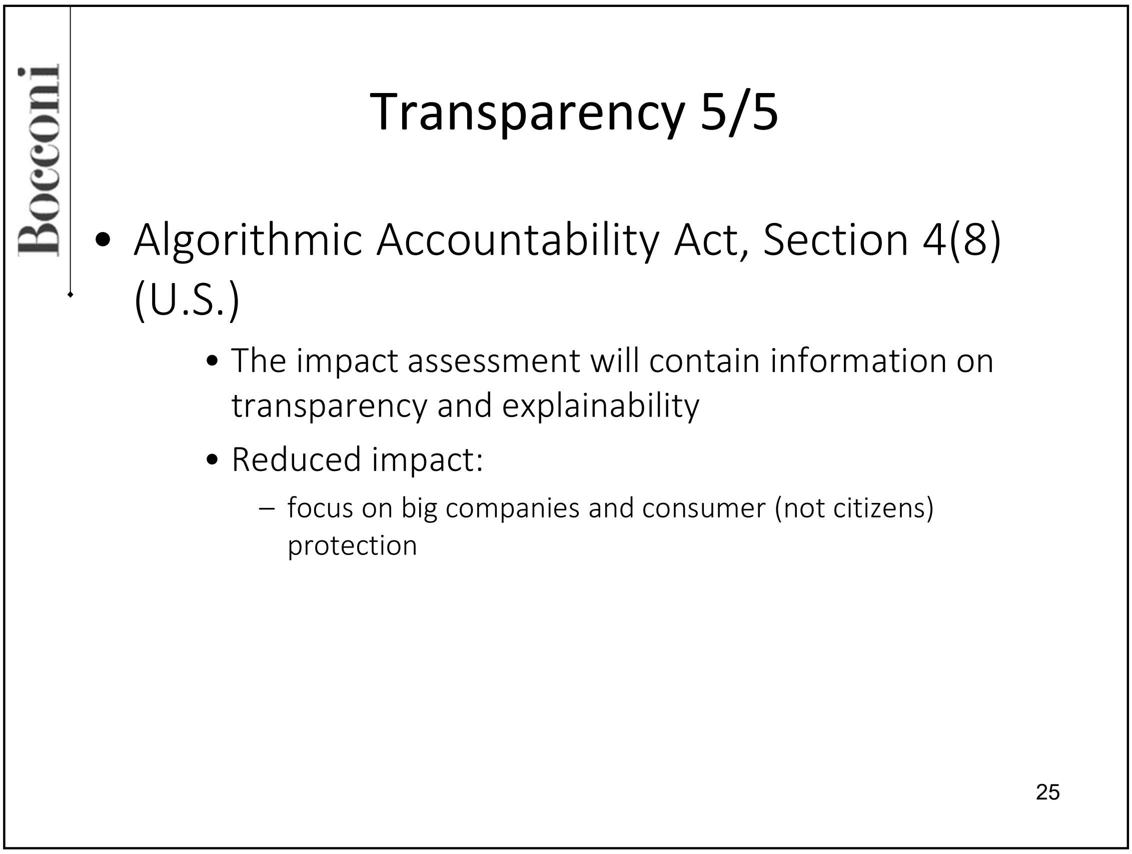
## Transparency 4/5

- Transparency requirements for limited-risk AI systems (Article 69):
  - Deep fakes
  - Systems interacting with people (bots)
  - Emotion recognition
  - Biometric categorization

The introduction of these transparency requirements has to be welcomed in that it will ensure higher standards and a possibility to investigate further the functioning (or malfunctioning) of AI systems. However, it is not always certain what information may afford sufficient transparency as required by Article 13. Moreover, it remains unclear if this is enough for transparency we want and need as citizens with rights. Most notably, information to be provided to users does not necessarily mean that individuals as third persons are always afforded necessary safeguards in the form of transparency. As with XAI, it seems that transparency, as envisaged by the AI Act, remains at a “higher” and technocratic level, so that openness is relevant for the relationships between providers and users and for the relationship between providers and public institutions: limited attention is given to transparency for the individuals actually affected by automated decision-making. Moreover, the AI Act proposal imposes strict confidentiality standards.

Non-high-risk AI systems comprise those that pose a “minimal risk” and those of a “limited risk”. As for minimal risk AI systems, Article 69 of the AI Act simply encourages the making and use of codes of conduct aiming at fostering transparency, human oversight, and robustness. Therefore, transparency is left to the self-regulation of providers and users of those systems. Instead, the category of limited risk features AI systems that pose issues precisely in terms of transparency: deep fakes, systems intended to interact with people (notably, bots), and emotion recognition and biometric categorization systems. In the case of bots, providers are required to design and develop AI systems so that natural persons are informed that they are interacting with an AI system.

Similarly, users of an emotion recognition or a biometric categorization system are to inform natural persons exposed about its operation. Finally, artificial generation or manipulation in the case of “deep fakes” should be disclosed. Overall, users of limited-risk AI systems need to be transparent about their artificial nature towards the natural persons exposed to them.

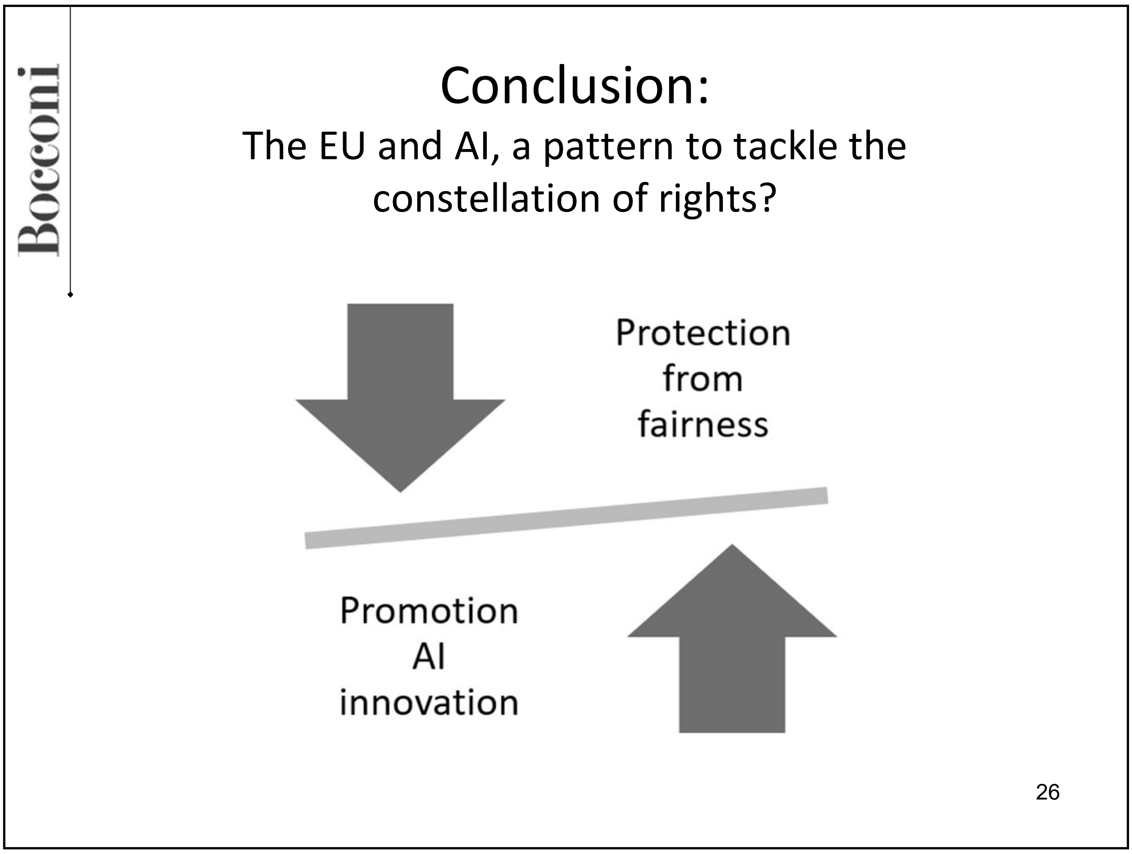


## Transparency 5/5

- Algorithmic Accountability Act, Section 4(8) (U.S.)
  - The impact assessment will contain information on transparency and explainability
  - Reduced impact:
    - focus on big companies and consumer (not citizens) protection

25

It is important to note that the goal of fostering transparency is not exclusive to the AI Act proposal alone. Indeed, on the other side of the Atlantic, the Algorithmic Accountability Act (AAA) also focuses notably on procedural regularity and transparency. The Act sets an obligation on specific actors to perform an impact assessment which must, *inter alia*, evaluate the rights of consumers, including the extent to which they are provided with clear notice that an algorithmic system or process will be used and the existence of a mechanism for opting out of 24 such use (Section 4(8)). Moreover, the impact assessment will assess the transparency and explainability of such systems or processes and the degree to which a consumer may contest, correct, or appeal a decision or opt out of such systems or processes. In so doing, an important aspect will be that concerning the information available to consumers or representatives or agents of consumers, such as any relevant factors that contribute to a particular decision, including an explanation of which contributing factors, if changed, would cause the system or process to reach a different decision (counter-factual explanation), and how such consumer, representative, or agent can access such information. Some commentators have indeed praised the AAA for its focus on procedure and transparency: however, those same commentators have also argued that the US approach, despite being the «latest milestone in a worldwide trend to complement self-regulation in the domain of automated decision-making with legislation», it is nonetheless too modest if compared to the Union's AI Act, partly because it only applies to "large companies" and partly because it refers simply to the rights of consumers, rather than to the rights of consumers.



# Conclusion:

## The EU and AI, a pattern to tackle the constellation of rights?

Protection from fairness

Promotion AI innovation

When addressing digital policies and issues concerning developing technologies, the approach of the EU is in general based on the goal of pursuing a balancing between two often contradicting goals: the fostering and protection of innovation and of the Digital Single Market, on the one hand, and the mitigation of risks deriving from the implementation and use of those technologies. This goal, which emerges in many pieces of EU legislation including the GDPR and the AI Act, has its roots in the characteristics of European constitutionalism, in which the logic of balancing permeates the entire constitutional architecture. Against this backdrop, no right or liberty, most notably economic freedom, may be invoked as a justification to destroy other individual fundamental rights. The EU's regulation ultimately attempts to regulate the digital market by striking the optimal balance between innovation and protection of constitutional and democratic values.
AI can represent a fundamental tool to improve our lives and our world, but specific attention must always be given to its collateral effects which, especially when it comes to biases and unfairness, must be counterbalanced through the resort to legal principles and strategies such as the human-in-the-loop and transparency.