Legal Aspects of Transparency, Explainability and Accountability

Doaa ABU ELYOUNES

Technicalities

The raising hand functionality will be handled through a dedicated website

1) Go to: iraisemyhand.com

2) Enter channel name: **RML2023**

3) Enter your name, and join

Keep the website running in the background and simply press on the raise hand icon any time you have a question/reaction.

Introduction

What is the difference between transparency, explainability and accountability, and what does each one of the terms mean

Clarifying the terminology

According to the UK House of Lords:

- Transparency means that experts will be able to understand how a technical system is put together, it might entail being able to access the source code of an AI system, but it will not necessarily mean that we can understand why a particular system made a particular decision.
- Explainability means that AI systems are developed in such a way that they can explain the information and logic used to arrive at their decision

Clarifying the terminology

According to Doshi-Velez et al. explanation should be able to provide one of the following:

- Human-interpretable information about the factors used in a decision and their relative weight
- An answer to a counterfactual question

The two can be mapped to two technical concepts in AI: local explanation and local counterfactual faithfulness

- Counterfactuals enhance accountability
- Explanation can increase trust in the system by providing proof that a decision was made according to a fair, robust or accepted process

Doshi-Velez, Finale, et al. "Accountability of AI under the law: The role of explanation." *arXiv preprint arXiv:1711.01134* (2017).

Due process- the judiciary as an accountability mechanism

- Degree of proof that the decision caused a legal-cognizable and redressable injury
- Administrative agencies are required to explain their decisions as a matter
- Administrative agency must provide an explanation for a certain rule, and also individual explanations
- A rule that lacks an explanation will likely be struck down as arbitrary and capricious
- The precise amount of evidence required to compel an explanation varies with the governing law

Due process- judges as the ultimate explainers

- Explanation serves an important tool for accountability from judges.
- The appeal process
- Explanations helps guiding future decision making
- While generally speaking judgments must be reasoned across the board, the breadth of the explanation vary according to the context

Outcome based explanation versus logic based explanation

- Outcome based explanation requires reasoning to the level of a specific individual
  - Example, credit scoring, the Fair Credit Reporting Act and the Equal Credit Opportunity Act
  - The law requires “adverse action notice” that must include a statement of reasons for denials of credit or other credit-based outcomes
  - Adverse action notices aim to serve three purposes: (1) to alert a consumer that an adverse action has occurred; (2) to educate the consumer about how such a result could be changed in the future; and (3) to prevent discrimination.
  - Examples of reasoning codes: “no credit file,” “length of employment,” or “income insufficient for amount of credit requested.”
  - Balance between meaningful explanation and explanation that does not overwhelm the individual
- Logic based explanation is broader than the individual case
  - It focuses on the inner working of the algorithm
  - Rule based explanation

Transparency according to the GDPR

- Specific requirements in articles 13-14 GDPR
- The right to be informed: the purpose for the collection and processing, length of retention, and who it will be shared with
- The information should be given in plane language
- Information should be given about the data controller and data protection officer
- Information about the right to access the data, to erase it, to object to the collection, and the right to data portability

Prohibition on fully automated decision making- article 22 GDPR

"(1) The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

- *Debate in the literature whether "solely automated" means outcome based or logic based explanation*
- *What do you think? In the context of data processing, which type of explanation should be required?*

White-box versus black-box approaches to explanation

White box approach, self interpretable models:

- Straight forward to understand
- Always possible to know how the input features are transformed into the output or target variable
- Examples of “white box” models are decision trees, and linear regression

Black box approach, post hoc explanation:

- Relevant for more complex models like neural networks consisting of more than three layers
- Explanation is generated after the decision has been made, and can be classified as either global or local
- Global explanations provide an overall understanding of the behavior and decision-making process of an AI model, and aim to capture patterns, general trends, and insights that apply broadly to the model’s behavior
- Local explanations focus on the decision-making process of an AI model for a specific output (e.g. “why my application for a job vacancy has been refused?

Liability Rules

- Aim to ensure that in case of damage caused by a product or service, victims get proper compensation
- They provide economic incentive to the liable party to avoid causing such damage
- Delicate balance between protecting citizens from harm and enabling businesses to innovate

The fault-based liability regime

- This is the standard liability regime
- Claimants need to prove three cumulative conditions to be eligible for damages:
  - Fault as a result of someone failing to act, violation of the law or failing to apply expected cautiousness
  - Damage- the culpable conduct of by the defendant resulted in damage
  - Causality between the fault and the damage

Is there any difficulty with this model?

Other liability regimes

Other victim friendly liability regime include:

- Fault-based liability with presumptions- shifting the burden of proof
  - For example, responsibility of a builder for damaged caused if a building collapses, unless they prove that observed appropriate care
- Strict liability regime- liability without wrongdoing, action that causes harm is all that is required
  - The victim will be compensated more easily but in a more limited manner

Three types of defect that would lead to strict liability

- A manufacturing defect, where the product was not manufactured in accordance with the manufacturer's specifications
- A design defect, where a cost-effective change to the product design could have avoided foreseeable injury
- A warning defect, where inadequate instructions led to foreseeable harm

The EU Product Liability Directive 1985

- Establishes strict liability regime where produces are liable for defective products regardless of whether the defect is their fault
- The PLD is a technology neutral instrument
- It is applied to tangible goods and it can embody digital content like IoT product
- The notion of defect focuses on consumers' safety expectations to physical harm
- Damage is defined as death, personal injury, or damage to the product
- What are the difficulties in applying liability rules to AI systems?

Difficulties applying the current laws to AI

- Involvement of multiple stakeholders
- The components of AI systems are interdependent
- Opacity- it will be difficult to understand the source of harm
- Transparency requirements could solve this though the question remain who is meant to benefit from transparency and what information would be needed
- Autonomy- it will be difficult to trace back specific action to specific human decision

New law – Revised Product Liability Directive

- The goal is to broaden the strict liability regime and apply it to advanced machinery
- It will allow compensation of damage when products like robots, drones or smart home systems are made unsafe by software updates
- It will include cases where the manufacture failed to address cyber security vulnerabilities
- Requires manufacturers to disclose more evidence to investigate claims
- It alleviates the burden of proof for victims in complex cases including those involving AI

New law- the AI Liability Directive

- It aims to ease the access to redress for victims
- It will harmonize certain rules for claims outside of the scope of the Product Liability Directive, in cases in which damage is caused due to wrongful behavior
- It will cover bridges of privacy or damages caused by safety issues
- It will for instance, make it easier to obtain compensation if someone has been discriminated in a recruitment process involving AI technology.
- If fault has been established, the “presumption of causality” will be applied
- When high risk AI is involved, victims will have “a right to access evidence from companies and suppliers”

Liability laws in other countries

- The proposed EU laws inspired action in other countries
- In Argentina, Chapter VII of their AI law proposal includes parameters for determining liability for damage or misuse of AI. Art. 19 and 20 differentiate between the liability of developers/providers and users, respectively.
- In Brazil, Art. 5 of the proposed AI law is a provision on AI liability. It limits liability to the agent's involvement in the operation of AI systems.
- In Argentina, the proposed law goes one step further and article 21 obliges developers, suppliers, and users of artificial intelligence systems to have civil liability insurance that covers damage to people and property.
- What do you think about the insurance provision? Should it be included also in the EU laws?
- How about emotional harm?

Case study- who is liable for accident caused by self driving car

- Some crashes involving autonomous or semi-autonomous systems have resulted in property damage and others have even resulted in death
- Tesla, like many other car companies, maintains that full responsibility rests with the person in the driver's seat
- Is there a difference between full autonomy and semi autonomy?
- Where the line between semi autonomy and full autonomy passes?
- How responsibility should be allocated in each case?

Intermediaries Liability

Legal provisions in the United States

47 U.S.C. § 230 of the Communication Decency Act (DCA) Protection for private blocking and screening of offensive material

The goals of the law:

- To protect innovation and to preserve the “vibrant and competitive free market that presently exists”
- To incentivize companies to create technology that allow parents to block access to objectionable or inappropriate online material.
- To encourage “true diversity of political discourse, unique opportunities for cultural development, and myriad avenues for intellectual activity”.

Section 230 c(1)

TREATMENT OF PUBLISHER OR SPEAKER

"No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider."

The rational was to protect small content providers.

Section 230 c(2)

CIVIL LIABILITY

"No provider or user of an interactive computer service shall be held liable on account of—

(A) *any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected; or*

(B) *any action taken to enable or make available to information content providers or others the technical means to restrict access to material described in paragraph (1).*"

Key cases

- **Zeran v. AOL:** Immunity cannot be eliminated via demand letters or takedown notices, in addition the protection is broader than what was given in the past to publishers.
- **Doe v. Myspace:** social media company is immune from liability to violations taking place offline, even though the connection was made through the platform.
- Intermediaries have immunity also for publishing defamatory content.

Doe v. Backpage

- Three young female victims of human trafficking sued Backpage.com for running online prostitution ads featuring them.
- Their primary claim was based on the Trafficking Victims Protection Reauthorization Act of 2008 (TVPRA), which applies to anyone who "knowingly benefits, financially or by receiving anything of value from" human trafficking (18 U.S.C. 1595(a)).
- Backpage claimed protection under section 230 as they are the publisher not the advertiser

Exception to section 230 – Fight Online Sex Trafficking Act (FOSTA)

- In 2018, FOSTA was enacted as a response to sex trafficking claims against Backpage.com, and the belief that Section 230 insulated the site from liability for facilitating sex trafficking
- FOSTA provides that the Section 230 liability shield doesn't apply when plaintiffs bring a civil claim for sex trafficking where the conduct underlying the claim constitutes a violation of the federal criminal sex trafficking ban
- Does v. Reddit: the plaintiffs sued Reddit as part of a putative class action, alleging that Reddit not only failed to stop, but actively profits from, child sexual exploitation materials
- The Ninth Circuit interpreted the statute narrowly and held that in order for the FOSTA exception to apply, plaintiffs have to show that internet platforms "knowingly benefited" from trafficking;
- The court also held that the defendant-website's own conduct, rather than the conduct of a third-party, must have violated the underlying criminal sex trafficking ban for FOSTA to apply.

Legal provisions in Europe

Electronic Commerce Directive 2000/31/EC

Article 12- "Mere conduit"

1. Where an information society service is provided that consists of the transmission in a communication network of information provided by a recipient of the service, or the provision of access to a communication network, Member States shall ensure that the service provider is not liable for the information transmitted, on condition that the provider:

(a) does not initiate the transmission;

(b) does not select the receiver of the transmission; and

(c) does not select or modify the information contained in the transmission.

- **LSG case-** service provider which only provide access to the internet without additional services such as email is immuned.

"Caching"

Electronic Commerce Directive 2000/31/EC

Article 13- "Caching"

1. Where an information society service is provided that consists of the transmission in a communication network of information provided by a recipient of the service, Member States shall ensure that the service provider is not liable for the automatic, intermediate and temporary storage of that information, performed for the sole purpose of making more efficient the information's onward transmission to other recipients of the service upon their request

Article 14- Hosting

Electronic Commerce Directive 2000/31/EC

Article 14- Hosting

1. Where an information society service is provided that consists of the storage of information provided by a recipient of the service, Member States shall ensure that the service provider is not liable for the information stored at the request of a recipient of the service, on condition that:

(a) the provider does not have actual knowledge of illegal activity or information and, as regards claims for damages, is not aware of facts or circumstances from which the illegal activity or information is apparent; or

(b) the provider, upon obtaining such knowledge or awareness, acts expeditiously to remove or to disable access to the information.

Key cases

- **Google France v. Louis Vuitton 2010:** Google's advertising services which is provided in conjunction with its search engine, may fall within the scope of Article 14.
- **Delfi AS v. Estonia:** Internet news portal cannot enjoy the immunity of liability because of the economic purpose.

L’oreal v. Ebay

- L’Oréal sent a letter to eBay, expressing concern over several unauthorized sale of its trademark goods via eBay’s European websites.
- The present infringement action primarily arose out of the sale of approximately 17 items on eBay by individual sellers from non-EU countries. Two of those items were allegedly counterfeit cosmetics bearing L’Oréal’s registered trademarks.
- eBay had some automatic filtering
- Do you think eBay can be liable for infringement?

Additional provisions

- Article 15- states cannot impose obligations on intermediaries to monitor information.
- All legal provisions cover the liability of intermediaries in their role as intermediaries and not as primary infringers.
- There are suggestions to rethink intermediaries liability.

The Digital Services Act DSA

- The interplay between a lack of knowledge or awareness of illegality remains a precondition to enjoy liability exemptions, however, the DSA encourages platforms proactive investigation of hosted content, which might trigger aforementioned knowledge or awareness.
- The DSA incorporates new regulatory “layers”, which may lead to even more challenging interpretation issues
- Four types of intermediaries: services offering network infrastructure, hosting services, online platforms, and very large online platforms (covering at least 10% of the EU population)
- The power is left in the hands of the platforms to assess the legality of content
- Enhance transparency requirements

The Digital Services Act – continued

- The “good Samaritan clause”: voluntary initiatives to combat illegal content does not lead to automatic exemption from liability
- Enhance due diligence obligations: requirement to designate a point of contact and include information on content moderation and algorithmic decision making in the terms of reference
- Notice and action versus notice and take down
- Systems have to give priority to notices submitted by “trusted flaggers”
- Systems have to ensure that recipients are informed about how recommender systems impact the way information is displayed and how users can influence that

The notice and action mechanism

- Under the DSA, intermediaries are required to have a process in place to assist with the notification of ‘illegal content’ by individuals or entities. This process must be easy to access, user-friendly, and allow for the submission of notices exclusively by electronic means
- Once a notice is submitted to an intermediary service provider, they are considered to have actual knowledge of the information they store
- The DSA also imposes an obligation on the intermediary to follow up with the reporter without undue delay
- Requirement to implement an appeals mechanism
- Implement repeat abuse policies

Thank you