AI Transparency: What Does it Mean for Your Business?

As part of the three-part event series on AI ethics and risk management jointly hosted by UCL's Centre for Digital Innovation (UCL's CDI) and Holistic AI, Holistic AI's COO and co-founder Dr Emre Kazim was joined by CMS Partner Charles Kerrigan to discuss AI transparency and what it means for businesses.

This paper summarises the key takeaways from the panel – chaired by Graça Carvalho, Director of UCL's CDI – with the view to stimulating further discussion and thought-leadership. It will begin by outlining what is meant by AI transparency before discussing how AI can be made more transparent and the implications of this for both businesses and users of AI systems.

About the Centre for Digital Innovation

UCL's CDI is a joint initiative between UCL and AWS to produce evidence-based and commercially sustainable technological innovations with a focus on Healthcare and Education. The Centre supports and accelerates UCL's technology spinouts and ecosystem of partners, users, and customers, to deliver digital innovation to solve global issues primarily in the fields of healthcare and education. This support covers distributed infrastructure design, data stewardship, and digital governance, which are fundamental components of a trustworthy digitally enabled ecosystem.

Within the Centre is the Responsible Innovation Lab, a self-sustained, free-standing initiative that will become the de-facto the go-to place for informed, structured, and engineered standards on Responsible Digital Innovation. The CDI's Responsible Innovation Lab aims to facilitate experimentation and has accountability and operationalisation of research and ideas as its core values. The Lab aims to foster public trust, community building, and communication, with the goal of providing thought-leadership in responsible digital innovation through experimentation. By engaging with the scientific community and industry, the Lab sets best practices that can be implement by third parties.

Based around the four pillars of privacy, fairness, transparency, and auditability, the Lab encourages a polycentric approach to responsible digital innovation. As part of this, the Lab regularly hosts workshops, conferences, and panel discussions, recently hosting a three-part event series with Holistic AI, centred around the pillars of transparency, fairness, and privacy.

About Holistic AI

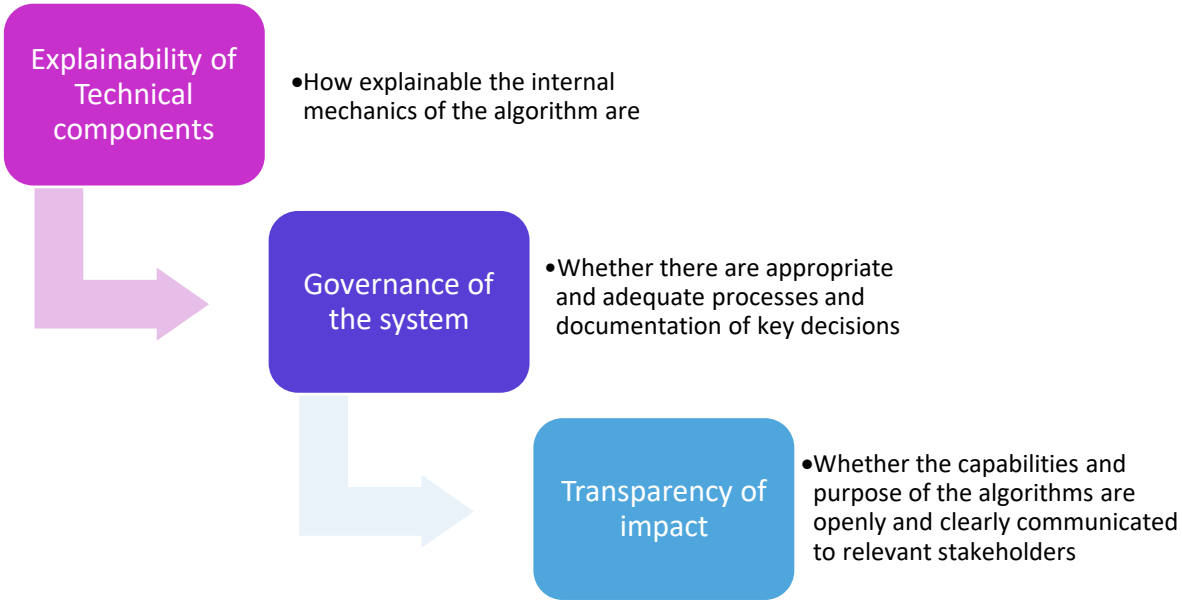
Holistic AI is an AI Risk Management platform that enables businesses to catalogue their AI systems and identify and mitigate risks associated with them. The company was founded by Dr Emre Kazim and Dr Adriano Koshiyama who are thought leaders in AI ethics and risk management and have published over 50 papers in this space. Having emerged from the Department of Computer Science at UCL, Holistic AI has now completed over 1000 risk mitigations and is trusted by global brands.

What is AI Transparency?

Artificial intelligence (AI) is a broad term that describes algorithmic systems that are programmed to achieve human-defined objectives. The outputs of these systems can include content such as images, predictions, recommendations, or decisions, and they can be used to support or replace human decision-making and activities. Many of these systems are considered to be black-box, where the internals of the model are either not known or are not interpretable to humans. In such a case, the model can be said to lack transparency.

AI transparency – one of the core pillars of the Responsible Digital Innovation Lab – is an umbrella term that encompasses concepts such as explainable AI (XAI) and interpretability and is a key concern within the field of AI ethics (and other related fields such as trustworthy AI and responsible AI). Broadly, it comprises three levels:

- Explainability of the technical components – how explainable the internal mechanics of the algorithm are
- Governance of the system – whether there are appropriate and adequate processes and documentation of key decisions
- Transparency of impact – whether the capabilities and purpose of the algorithms are openly and clearly communicated to relevant stakeholders

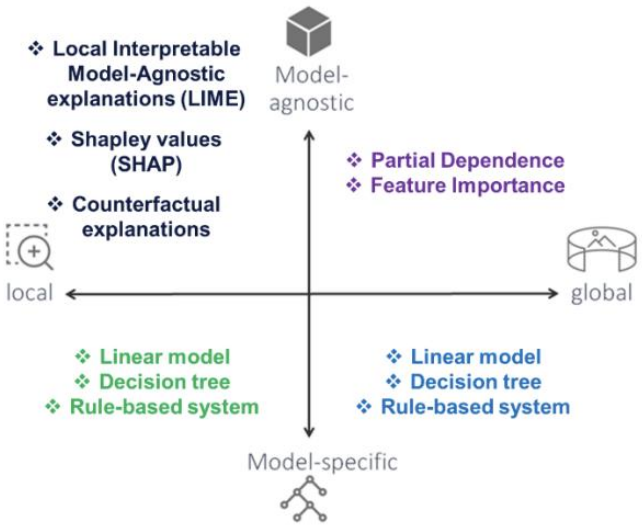


Explainability of Technical Components

Explainability of the technical components of the system refers to being able to explain what is happening within an AI system and is based on four types of explanations: model-specific and agnostic, global and local.

- Model-specific explainability – a model has explainability built into its design and development
- Model-agnostic explainability – a mathematical technique is applied to the outputs of any algorithm to provide an interpretation of the decision drivers of the model
- Global-level explainability – understanding the algorithm's behaviour at a high/dataset/populational level, something that is typically done by researchers and designers of the algorithm

- Local-level explainability – understanding the algorithm's behaviour at a low/subset/individual level, typically those being targeted by an algorithm



Governance of the System

Governance of a system includes establishing and implementing protocols for documenting decisions made about a system from the early stages of development to deployment, and for any updates made to the system.

Governance can also include establishing accountability for the outputs of a system and including this within any relevant contracts of documentation. For example, contracts should specify whether liability for any harm or losses is with the supplier or vendor of a system, the entity deploying a system, or the specific designers and developers of the system. Not only does this encourage greater due diligence if a particular party can be held liable for a system, but it can also be used for insurance purposes and to recover any losses that result from the deployment or use of the system.

Outside of documentation and accountability, governance of a system can also refer to the regulation and legislation that govern the use of the system and internal policies within organisations in terms of the creation, procurement, and use of AI systems.

Transparency of Impact

The third level of transparency concerns communicating the capabilities and purpose of an AI system to relevant stakeholders, both those who are directly and indirectly affected. Communications should be issued within a timely manner and should be clear, accurate, and conspicuous.

To make the impact of systems more transparent, information about the type of data points that the algorithm will use, and the source of the data should be communicated to those affected. Communications should also indicate to users that they are interacting with an AI system, what form the outputs of the system take, and how the outputs will be used. Particularly when a system is found to be biased towards particular groups, information should also be communicated about how the system performs for particular categories and

whether particular groups might experience negative outcomes if they interact with the system.

How Can AI be Made More Transparent?

AI systems can be made more transparent in relation to all three levels of transparency. The first step for doing this is to ask questions – can what is happening inside the algorithm be explained? Are there any processes within the design, development, or deployment of the algorithm that are not well-documented? Who is accountable for the model and any harms or losses? Is the accountable entity clearly identified in all relevant documentation and contracts? Have users been told they are interacting with an AI system? Are the key features and outcomes of the system clearly and accurately communicated to relevant parties? Based on the answers to these questions, steps can then be taken to implement appropriate strategies.

Technical Components

For example, tools to make the technical components more transparent can be applied to both the model-specific and model-agnostic parts of the system. Here, model-specific approaches can provide global and local explanations by design, while model-agnostic procedures act as a post-hoc 'wrapper' around an algorithm, with some techniques only focusing on local explanations (e.g. LIME) or global explanations (e.g. Partial Dependency plots).

Stage/method	Technical solution
In-processing/ Model-specific	• Rule-based explanations: decision trees, rule-induction methods • Model's coefficients: linear regression, linear discriminant analysis • Nearest prototype: k-nearest-neighbour, Naïve-Bayes
Post-processing/ Model-agnostic	• Surrogate explanations: LIME (Ribeiro et al., 2016), Explainable Boosting Machines (Nori et al., 2019), PIRL (Puiutta et al., 2020) • Perturbation: Gradient-based Attribution Methods (Ancona et al., 2017), Permutation Importance (Breiman, 2001), SHAP (Lundberg and Lee, 2017) • Simulation analysis (what-if?): counterfactual explanations and algorithmic recourse (Wachter, 2017; Karimi et al., 2020)

Governance

Governance transparency can be improved by ensuring that there are clear policies and processes for documenting decisions made about algorithms and that they are followed by the relevant people. These processes and documentation should be reviewed and updated as necessary, but at least annually, to ensure that they are still suitable to the specifications of the system.

Legal documents should also clearly identify the entities that are responsible for the system's outputs and any harms and losses to ensure that there is accountability and a feasible way to

recover any losses or bring about legal action. Indeed, courts have taken the position that AI systems themselves cannot be listed as a creator since there is usually at least one human behind the system that ultimately owns the rights to the content. Similarly, the liability for an AI system cannot lie with the system itself, *there must be an entity that assumes responsibility for an AI system.*

Companies should also be aware of any relevant legislation that applies to the AI system that they are using. For example, in the US, New York City has mandated bias audits of automated employment decision tools and providing information to employees and candidates about the system's specifications. Likewise, in Illinois, employers must inform job candidates of the use of AI-based video interviews and the characteristics the system uses to make its decision. Elsewhere, Spain has introduced transparency requirements for platform-based services and the EU's AI Act will impose transparency requirements for the users of certain AI systems. Outside of transparency requirements, other laws are tackling multiple risk verticals including bias or discrimination, particularly in the insurance and HR tech sectors.

System Impact

To increase the transparency of the impact of a system, procedures should be in place for communicating with relevant stakeholders. This can include adding a notice to the appropriate webpage or contacting stakeholders directly, for example. Notices should be written for the target audience (i.e. no technical jargon in communication meant for laymen) and should be accurate. Communications should be issued before a system is interacted with where possible, or as soon as possible after interaction with the system. Communications could outline whether particular groups might be subject to bias by the system, whether it performs accurately, and any privacy-relevant information.

Why do we Need AI Transparency?

A major motivation for increasing the transparency of AI systems is demystifying the systems to give users and other stakeholders more confidence. Knowing the decisions, a system makes and how it makes them can also give individuals more agency over their decisions, allowing them to give *informed* consent to interacting with a system.

As well as this, transparency can also have several business benefits. Firstly, by cataloguing all of the systems being used across a business, steps can be taken to ensure that algorithms are being deployed where they need to be and that simple processes are not being overcomplicated by using complex algorithms to do minor tasks.

Secondly, if legal action is brought against businesses and they can explain how their system works and why it came to the decisions it did, this can help to resolve the issue quickly to ensure that appropriate action can be taken when necessary. An applied example of this is the action that was brought against Apple for their Apple Card, which reportedly gave a much higher credit limit to a man compared to his wife, despite her having a higher credit score. However, Goldman Sachs, the provider of the card, was able to justify why the model came to the decision that it did, meaning that they were cleared of illegal activity, highlighting the importance of explainable AI.

Ultimately, the overarching goal of transparency is to establish an ecosystem of trust around the use of AI. In particular, transparency is with a view to establishing the trust of citizens or users of systems, especially the communities that are at the most risk of harm by AI systems. For example, underrepresented demographic groups can be particularly vulnerable to biased systems due to the small sample sizes available for testing for discrimination. Particular industries can also be at greater risk of job displacement from automation. By making AI more transparent, these communities can make more informed decisions about their interaction with and adoption of these tools and can build greater confidence around the impact of these systems.

Summary

- ➔ AI transparency is an umbrella term that is concerned with the explainability of the technical components, governance, and impact of AI systems
- ➔ Technical components can be explained using a model-specific or model-agnostic approach at a global or local level
- ➔ Technical explainability can be built into the model or added using post-hoc techniques
- ➔ Governance concerns documentation, establishing accountability and liability, and compliance with relevant regulations and legislation
- ➔ Transparency can be increased by ensuring that there are well-defined and widely abided by governance practices and processes
- ➔ Transparency in terms of impact occurs by clearly and openly communicating system specifications and implications with relevant stakeholders
- ➔ Communications should let users know that they are interacting with an AI system, how the system makes a decision, what the output is, and how it will be used
- ➔ Transparency can help to increase public trust in AI, ensure legal compliance, and help businesses survey their inventory of algorithms

Author: Airlie Hilliard – Senior Researcher, Holistic AI