UNDERSTANDING ARTIFICIAL INTELLIGENCE

AI Labelling & Watermarking – Explained

This explainer on AI labelling and watermarking is part of CCIA Europe's 'Understanding AI' series, which aims to inform EU policymakers and the wider public about important concepts related to artificial intelligence (AI) and the EU regulatory framework.

What are labelling and watermarking?

Labelling and watermarking are two techniques used in AI. However, they are deployed at different stages and for different purposes.

Labelling involves providing clear and accurate descriptions of the data used to train AI systems. It helps the AI understand and recognise patterns in the data, making it more effective in generating content aligned with the intended outcomes. For example, in training an AI to recognise images of dogs, labelling would involve assigning a label of 'dog' or 'not dog' to each image in the data set.

Watermarking is the process of embedding subtle, unique, and imperceptible markers or symbols into generated content. These watermarks can be used to recognise and detect AI-generated content, such as articles or images. Watermarks can also serve as a digital signature, allowing content creators to claim ownership and identify misuse.

In summary, labelling is used at the training stage of an AI system to improve the quality of the input data, while watermarking can be used on the output data generated by an AI system to track and identify specific content.

How are these used in practice?

In practice, when training AI models the **labelling of training data mainly allows for the classification of images** (such as cars or animals). The labelling of text, on the other hand, helps to **identify a sentiment or specific keywords**. For the training of AI systems intended to recognise speech, labelling can also include transcribing audio recordings or identifying specific noises (such as traffic or planes in the background) in an audio input file.

Beyond copyright protection and fraud prevention, **watermarking can be used to trace AI-generated content**. This can enable the identification of AI-generated text or images, as well as the AI system that generated the content.

What are the implications for AI policy?

While labelling training data is a rather straightforward process and has relevance in the context of data governance, watermarking is a more complicated process and has more wide-ranging implications.

In the EU, the proposed **AI Act** **contains two different sets of requirements for labelling and watermarking, though both are referred to as "labelling" for some reason** – which only adds to the confusion. The AI Act subjects high-risk AI systems to data and data governance requirements, notably by requesting that training data be subject to appropriate governance measures, including – but not limited to – labelling training data.

The AI Act also sets out a number of transparency requirements for AI systems interacting with humans, such as chatbots. EU Member States have clarified that this requirement would mean that users of generative AI **have to disclose that the content has been artificially created or manipulated** by labelling output and disclosing its artificial origin.

The European Parliament went a step further, calling for users of generative AI to be **required to label any artificial content that possibly could be construed as being created by a human** in a way that is clearly visible to other users or recipients.

Yet when it comes to watermarking, this is easier said than done. In addition to the **technical challenge of watermarking audio and visual content**, the Parliament's position would also oblige users to watermark or otherwise 'label' AI-generated text.

What are sensible rules?

It is very **important that the new EU rules clearly distinguish between the labelling of training data** – at the input stage of the AI development process – **and the labelling of output data through watermarking**. Both processes have very different purposes, technical requirements, and policy implications.