**Making AI More Transparent**

*A pragmatic solution to a critical problem*

by Mark Roeder

**The Challenge**

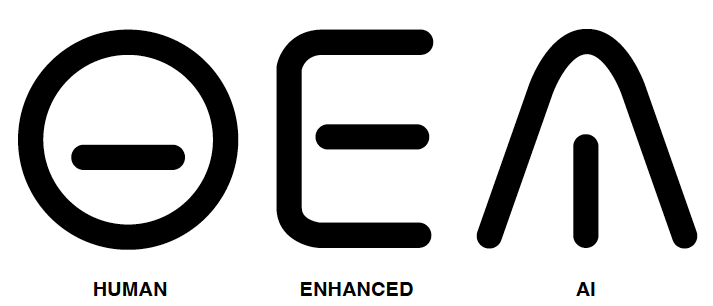
It is becoming increasingly difficult to tell whether the media content we are exposed to is created by a human or an AI (artificial intelligence). This lack of transparency has profound consequences, because we need to know who – or what - we are dealing with in order to make informed decisions. The problem can be particularly acute during elections because the public could be influenced by hyper-realistic looking, AI-created content that impersonates people or distorts reality. This could directly affect the election outcome. AI’s can also be inaccurate (or ’hallucinate’) – a problem that is compounded when consumers are not aware that AI is being used.

Indeed, the rapid evolution of AI generated content will soon permeate nearly every aspect of our lives, by blurring the lines between the real and virtual – between the authentic and artifice. This will inevitably undermine our sense of the shared reality on which societies are built.

**Solution**

This paper proposes a simple, effective solution to this critical and urgent challenge – *Source Labelling –* that identifies whether content is Human, Enhanced or generated by AI.

The labels look like this:



These labels are self-explanatory, and signal whether the content we experience is created by Humans; is Enhanced by AI, or totally generated by AI. They are designed in a logographic form that works across cultures, including the Anglo-American, European and Asian languages. Coloured versions are available – Green for Human, Orange for Enhanced and Blue for AI.

* ***Human*** – means that the content is generated solely by humans, and is not the result of an AI generative program. This Human content may have utilised technological tools, such as spell checkers or basic image augmentation, but such usage is always under the direct control of a human being, and the tool does not act ‘generatively’.
* ***Enhanced*** – means that the content or interaction is partially assisted by an AI. Examples include a human created music track that is complemented by some AI elements; an image that has been manipulated by an AI, or an article that is partially written by an AI. Importantly, the level of Enhancement must be specified below the E symbol, in percentage terms, ranging from 10% enhancement up to 90% enhancement. Above 90% the content is considered to be an AI generated content.
* ***AI*** – any content or interaction that utilises over 90% AI is considered to be ‘full AI’. For example, an article produced by, say, Chat GPT is considered to be an AI article and will receive the A symbol. Note: The AI limit is designated at 90% and above to deter AI content producers from producing work at, say, 99%, and claiming it is Enhanced.

These symbols (source labels) would appear at the beginning of any content on TV, internet, video and social media – especially during sensitive periods such as an election campaign, a public health crisis or a national emergency. Sonic versions would appear for auditory applications.

**Specific applications**

These Source Labels can be used in numerous situations:

* **Education –** Teachers can specify to their students whether an assignment is to be produced on their own (i.e., Human), or enhanced by AI, or completely produced by AI. In the latter case, the students will be assessed largely on the quality of their ‘prompts’ to the AI, which will become an increasingly important skill.
* **Social media** – Users of social media (i.e., Instagram, Tick Tok and Facebook) will know the sources of the content they are exposed to.
* **Film and Television** – Viewers will be able to discern the source of the content they are watching. Some viewers may prefer to view only human generated content because they perceive it to be more ‘authentic’ and socially beneficial. Others may be happy with pure AI material or something in between (Enhanced).
* **Creative Arts** – Consumers will know whether an artwork, photograph or music track is produced by a human, enhanced or AI generated. Conversely, some artists might highlight the human-generated nature of their works (i.e., hand crafted) and use this label as a point of differentiation. Other artists might draw heavily on AI to create their works and publicise this fact. The point is, that that Source Labelling makes it clear to everyone how the art is created, which fosters more informed appreciation of the works and their value. This clarity will also help protect artists from the widespread exploitation of their works that is currently happening.
* **Medical** – Patients will know whether the advice they are receiving has been influenced or generated by AI. Some patients may prefer a more human-centric, intuitive approach that reflects the traditional nature of the doctor patient relationship. Others, may be happy with a purely AI generated diagnosis that draws on a vast database of thousands of cases. The objective is to provide greater transparency which will, in turn, build more trust in the patient-doctor relationship. This approach is the opposite of ‘black box medicine’ where the patient is left in the dark about the inputs that generated their diagnosis or treatment plan.
* **Law** – The legal profession is already being transformed by AI, which is able to instantaneously sift through vast archives of statutes, case law and precedent, in order to create a reasoned argument, or simply prepare a will or conveyancing form. This revolution has obvious benefits. The downside is that people involved in legal matters (clients, judges, clerks etc) may not be aware of the extent to which AI has influenced the process. This will become especially important in the future when some courts may be adjudicated by AIs rather than human judges, and these AIs may not be sensitive to the subtle levels of moral nuance that a human judge may consider.

**Implementation**

The application of the three source labels (H, E and AI) should be mandatory, and enforced via legislation. Realistically, however, the implementation of this program will vary by industry sector and region, and be affected by various vested interests, some of which may be reluctant to have transparency regulation. Hence, it may be necessary to introduce the program on a ‘voluntary’ basis at first (i.e., content providers can opt-in), in order to gain public and industry acceptance before moving to a ’mandatory’ policy.

In effect, there will likely be three phases:

|  |  |  |
| --- | --- | --- |
| **Strategy** - Develop a comprehensive strategy and implementation for the AI Source Labelling program, based on input from experts and key stakeholders.  **Tools** - Create a set of tools for content providers to use.  **Education** - Devise an education program for the public and content providers on AI source labelling.  **Regulation** - Commence development on a long-term regulatory framework and the technical resources that underpin it (i.e., algorithms that detect and measure the level of AI content)  **Expertise** - Seek relevant expert partners to create this framework  **Funding** – Determine funding requirements for the program and consider sources. | **Invitation** – Encourage content providers to label the level of AI content in their outputs by displaying the source label icons.  **Tools** - Provide the tools (icons and processes) to content providers to enable them to offer AI transparency.  **Education** - Begin a public education program on AI transparency and the need for source labelling.  **Stakeholder engagement** - Engage with stakeholders on the longer-term need for a mandatory enforcement approach, and seek their buy-in. | **Legislation** - Draft legislation that describes the scope and implementation requirements of AI transparency (i.e., the AI Transparency Act)  **Certification process** – To evolve from a voluntary to a mandatory regime, it will be necessary to create a certification model for AI transparency, that requires content providers to meet certain criteria in order to claim that their content is Human, Enhanced or AI. This certifiable process will be managed, verified and enforced by the relevant regulatory authorities (see below), and supported by algorithmic technologies. Organisations like the ISO (International Organisation for Standardisation) may be helpful in this regard.  **Responsibility** - Determine who will be responsible for enforcing the mandatory approach, in each industry sector (i.e., which regulators).  **Education** – Educate the public and content providers on the mandatory program (why it is needed and how it will work) |

**Deep dive information**

The three consumer-oriented source labels above (H, E and AI) will eventually be supplemented with a ‘deep dive’ portal that grants access to detailed information about the AI sources to regulators and other interested parties. This portal would include, for example, data on the large language models (LLMs), training processes, algorithms, biases and safeguards built in to the AI. This ‘deep dive’ portal would be analogous to ‘nutrition labels’ on foods, which provide comprehensive lists of all the ingredients and chemicals used in their manufacture.

**Not for profit venture**

This initiative is not a commercially driven venture. It has been created in response to an urgent and critical problem, which is the lack of transparency in AI generated content.

Regarding copyright of the source labelling icons shown above (i.e., Human, Enhanced and AI), it is intended that these will be made available at no cost to any content provider that is willing to accurately label the sources of their content. However, as we move to the mandatory phase of source labelling, the use of these labels will be certified and licensed by the relevant regulators, as explained earlier.

**Keeping it simple**

As more people and organisations become involved in this process, there will be a tendency for the labelling approach to become more complicated. We must avoid this. The public ultimately needs to know whether the content they are experiencing is Human, Enhanced or AI generated. The simplest solution is often the best one.