Using ‘Probability’ as a Deepfake Detection Metric

Originally published at https://www.unite.ai/using-probability-as-a-deepfake-detection-metric/

If AI-generated video and audio get good enough, deepfake detectors based on visual artifacts or other traditional signals won’t work anymore. But given how rarely people veer away from predictable behavior, perhaps ‘probability’ could be adopted more deeply as a signal of whether a video or news rumor is likely to be true.

Opinion In the early 1990s, the respected former British footballer and TV sports commentator David Icke casually revealed on a chat show that he was ‘the son of God’ – a bizarre and unexpected revelation that would evolve over the following decades into a persistent and elaborate conspiracy theory about a secret and powerful global cabal of ‘lizard people’.

With internet adoption still some years away, and the advent of social media even further in the future, the sheer dissonance between Icke’s celebrity and the nature of his new insights had a profound impact on the British public – not least because of the complete lack of context, or any type of preparation for this massive pivot, from a well-known and well-established sports personality.

More than twenty years later, a similar and far darker strain of this societal shock occurred, when beloved charity campaigner and children’s TV host Jimmy Savile was posthumously found to be a serial and rapacious life-long sex offender who had used his wholesome public image to facilitate his crimes.

The subsequent Operation Yewtree police investigation would unearth many more UK celebrities with long histories of sexual offenses; later, the Harvey Weinstein prosecution would lead to a similar discovery of celebrity sex-offenders in the US, evolving into the #metoo movement, and accreting permanently into American culture in outings such as *The Morning Show*. ‘Shock’ news seemed to be developing a new and abrupt template – one which would eventually become adopted by deepfake attackers.

The End of ‘Traditional’ Deepfake Detection?

Even had social media and AI been around in the early nineties, no predictive system in the world could have foreseen Icke’s chat-show revelations, which (as I well remember) were not in any way foreshadowed in the years leading up to the event.

But then, had AI been around, it might have taken some time to convince a wider audience that Icke's declarations were not the product of Google Veo 3, or another of the new breed of hyper-real audio/video deepfake frameworks.

It's only in the last 6-12 months that AI deepfake methods have become effective enough to fulfill years of media doom-saying about deepfake election interference, and capable enough to generate the kind of quick-strike reputational stain that's untrue, yet hard to eradicate in an increasingly credulous culture.

To date, AI video output typically falls short of true realism, limited by technical hurdles and increasingly polarized by a widening gap between restrictive Western models and China's uncensored open source releases**.

Nonetheless, I notice increasingly in the research literature a looming concession of this cold war, for instance in the new paper *Performance Decay in Deepfake Detection*†:

> *'[We] assume that deepfake videos will continue to contain machine-learnable features which reliably distinguish them from genuine videos. As the capabilities of generative AI continue to advance rapidly, this assumption may well break down.*
>
> *'In such a scenario, watermarking and other provenance tracking methods will offer the only recourse for maintaining trust in digital media.'*

However, the same paper concedes that provenance-based solutions such as the Adobe-led Content Authenticity Initiative (and the very many smaller research offerings of the last 7-8 years) require such widespread adoption as to be unrealistic; and the paper ends on a general note of retreat, if not defeat.

If audio-visual deepfake detection methods are out-evolved by generative AI, and global adoption of an intrusive watermarking or provenance scheme falls at the diverse logistical hurdles, what common central feature could replace them as indicators of potentially faked output? Or must we resign ourselves to a world where all media is in doubt, and the Liar's Dividend prevails?

Knowledge Graphs

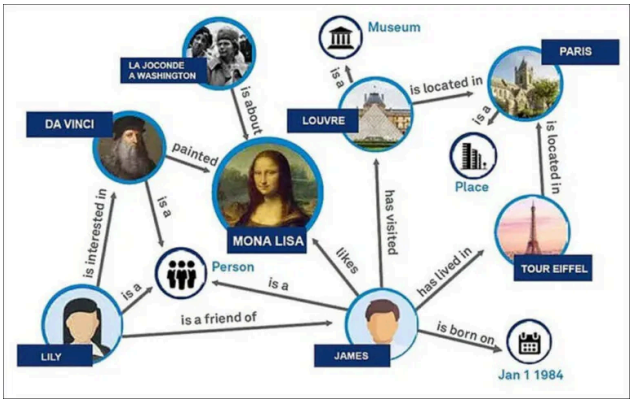
It seems time to more deeply leverage *probability* and *plausibility* of 'reported events' as a signal characteristic in deepfake detection. Further, since video and audio generative AI systems are increasingly converging, it may also be time for the separate research strands of 'fake news' (as a text-based narrative event) and fake imagery/video to similarly converge.

A *probability* deepfake metric is not the same as RAG-aided fact verification, where an AI model may bring in current web results to gain knowledge of events that occur after its own cut-off date, and/or to corroborate its claims.

Rather, it would perform predictions based on generally indicative statistical trends, derived from historical patterns that conform to a current inquiry.

In this sense, a probability method is nearer to statistical analysis than more modern threads in the current machine learning scene.

Though previously eclipsed by more modern Transformers-era approaches, knowledge graphs are making something of a comeback in the Enterprise space, and seem adapted to the potential deployment of 'probability' metrics in deepfake detection.



A simplified knowledge graph illustrating how people, places, artworks, and events can be linked through labeled relationships, enabling machines to reason over real-world entities and their connections. Source

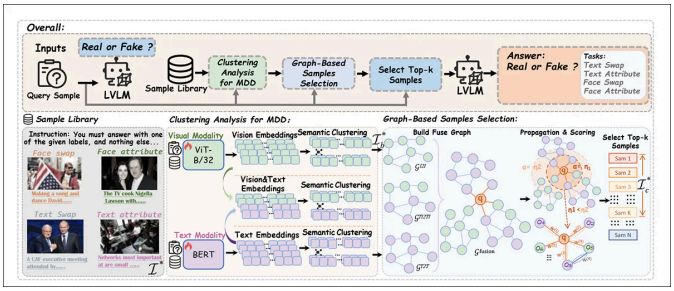
A knowledge graph is a way of organizing information by mapping real-world things such as people, companies, events, or ideas into a network of connected facts.

Each sub-entity is a node, and the links between them (edges) describe how they relate. For example, 'Microsoft' MSFT 1.75% (a node) might be linked to 'OpenAI' (another node) by an edge that says 'is a client of'. These connections are often stored in graph databases and follow a subject-predicate-object structure, such as 'Microsoft is a client of OpenAI'.

Persistent Memory

One Chinese study from September of this year proposed a training-free method that uses graph-based reasoning to detect subtle inconsistencies in multimodal deepfakes.

Instead of generating rationales or fine-tuning large models, the system retrieves image-text pairs, builds a similarity graph, and scores connections, in order to retrieve the most relevant examples,and these guide the model's judgment without the need for new training:



An overview of the GASP-ICL framework, which improves deepfake detection by combining graph-based sample selection with in-context learning, allowing a frozen vision-language model to classify image/text pairs as real or fake, without training or fine-tuning. Source

This is probably the nearest work, certainly that has crossed my path, to an 'informed' and history-aware approach to the evaluation and verification of new media output. For the most part, computer vision approaches continue to analyze *images* (including frames of videos and temporal anomalies encompassing multiple frames) while 'fake news' detection frameworks continue to emphasize text-based data, even in multimodal projects.

Feature Creep

The challenge of a predictive system of this kind is the scope of surveillance that may be necessary to make the approach fully performant – at least beyond the analysis of celebrities and public figures, for whom freely-accessible data already exists.

Probably the most similar current strand in the research is the field of pre-crime, which labels diverse multimodal intelligence signals as 'suspicious', and presents itself as a stalwart AI scarecrow in outings such as Jonathan Nolan's *Person of Interest* (2011-2016), and Steven Spielberg's *Minority Report* (2002).

While a *Person of Interest*-style omnivorous surveillance system would produce optimal results, it's unlikely at the moment that western culture could sanction the level of personal intrusiveness that China's internal networks impose on its citizens.

Therefore, in regard to potential fake news about *non*-celebrities, only governmental agencies such as the police (as well as birth & deaths registries and tax offices) would have enough pertinent historical information to inform probabilities in a graph-based workflow; and even they would need CCCP-style will, capacity, legislation and resources in order to include average citizens in their coverage and analyses (i.e., beyond banal but obligatory data points such as passport numbers and car registrations).

Probability Scoring

It seems likely that the potential effectiveness of a system of this kind would be constrained to the most obvious (current) use cases††† for deepfake content: *destabilization* (state-backed deepfakes); *celebrity and 'unknown' porn deepfakes* (which can both be considered malicious, though the latter case tends to attract deeper media concern); *fraud* (including audio/video deepfakes designed to perform 'impersonation heists'); and political character assassination.

A knowledge-based system would need a scale of probabilities for a diversity of possible events. At one end of the spectrum, common human failings such as questionable financial management, infidelity, addiction, indiscretion, etc.; at the other… revealing that you're the son of God on a live TV chat show (or events of similar scale and impact).

Even in the latter case, personal historical factors for any one individual would weight the probability outcome: a prominent political figure who has publicly equivocated in controversial matters (such as the veracity of the 1960s/70s lunar landings) to gain capital with an increasingly 'alternatively'-informed electorate, might gain additional *wildcard* status in verification routines, compared to their more staid peers.

In the case of celebrity porn, there is adequate real-world context (i.e., the 2012 celebrity photo leaks, among other – fairly rare – incidents) to generate a moderate Liar's Dividend, in certain contexts; but since these outlier incidents tend to operate as exceptions proving the rule, most of the current crop of diffusion-based celebrity porn videos would be deemed extremely 'improbable' (though this does not solve the issue of the appropriation of people's identities for such purposes).

In terms of national disruption, there is a considerable wealth of statistical data that can aid in assessing the probabilities of 'disastrous' reports. Even in ancient history, apparently 'out-of-the-blue' events such as the eruption of the unidentified volcano Vesuvius in 79ad were presaged, if you had paid enough attention; and besides the availability of a plethora of government and NGO-backed feeds, AI's evolving capacity to extract structure from raw data can provide additional historical context for probability scoring.

Conclusion

Even a well-implemented predictive system of this kind could not account for random chance, acts of God, freak occurrences, or for malicious events concocted away from all oversight.

Further, the sheer volume and depth of data needed, in order to provide coverage also for non-famous people, would be a political stumbling point – at least, for the moment.

However, the choices seem to be narrowing; vision-based analysis is poised to fail in the face of improved generative AI, while verification and provenance schemes carry an impeding burden of technical debt, and friction against adoption. This makes solutions such as the Content Authenticity Initiative, and the unfulfilled Metaphysic.ai face-copyrighting system Metaphysic Pro, challenging to popularize.

In their broadest usage, RAG-based systems can only determine if an authority source backs up an unverified claim; and since many big (true) news stories emerge without prior context, a lack of substantiation from authority sources is not necessarily meaningful.

Their value may prove greater if they can form part of a larger data ecosystem concerned with the one thing that most current forms of AI find challenging – historical context.

* *Not to be confused with the early autoencoder outings that debuted in 2017 and would eventually be supplanted by superior approaches.*

† https://arxiv.org/abs/2511.07009

** *Which can usually run freely on more powerful home PCs, instead of only being available via gate-kept APIs such as ChatGPT and the Veo series.*

††† *Omitting legitimate entertainment uses, such as professional visual effects in movie and TV productions.*

First published Thursday, November 13, 2025

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More