Journal of
APPLIED PHILOSOPHY

Journal of Applied Philosophy
doi: 10.1111/japp.12758

Something AI Should Tell You – The Case for Labelling Synthetic Content

SARAH A. FISHER

ABSTRACT *Synthetic content, which has been produced by generative artificial intelligence, is beginning to spread through the public sphere. Increasingly, we find ourselves exposed to convincing 'deepfakes' and powerful chatbots in our online environments. How should we mitigate the emerging risks to individuals and society? This article argues that labelling synthetic content in public forums is an essential first step. While calls for labelling have already been growing in volume, no principled argument has yet been offered to justify this measure (which inevitably comes with some additional costs). Rectifying that deficit, I conduct a close examination of our epistemic and expressive interests in identifying synthetic content as such. In so doing, I develop a cumulative case for social media platforms to enforce a labelling duty. I argue that this represents an important element of good platform governance, helping to shore up the integrity of our contemporary public discourse, which takes place increasingly online.*

1. Introduction

Social media environments have provided the perfect breeding ground for powerful new generative artificial intelligence and its alarmingly convincing products. The result is that familiar problems of viral misinformation and tribal propaganda have been supersized. We find ourselves fooled by deepfakes, which threaten our very grip on reality; or taken in by chatbots, whose conversational skills equal our own. How should we reap the benefits of the underlying technologies while mitigating their risks to individuals and society? This article argues, on philosophical grounds, that labelling synthetic content on public platforms is an essential first step.

Several major technology companies are already initiating some form of labelling, broadly conceived. For example, Amazon, Google, Meta, Microsoft, and OpenAI are among those developing watermarking systems for AI-generated content (a move backed up by President Joe Biden's executive order of October 2023). Meanwhile, social media sites like TikTok, Facebook, and Instagram have begun to require their users to disclose the provenance of the synthetic content they post. But what exactly is the argument for doing so – particularly given serious concerns about the feasibility, effectiveness, and costs of labelling? Notwithstanding vague appeals to safety and propriety, this question turns out to be a difficult one to answer. Here I address it in a more principled manner than has been attempted thus far. I do so by examining our various epistemic and expressive interests in identifying synthetic content as such. I argue that, taken cumulatively, they provide sufficient justification for labelling synthetic content.

I start by defining synthetic content (Section 2). I then examine how labelling could improve our epistemic position (Section 3) and our ability to respect individual expression

(Section 4). I conclude that public platforms should enforce a duty to label synthetic content – and I briefly consider the obstacles that such a policy will need to overcome (Section 5).

2. Synthetic Content

Synthetic content is defined here as any media – image, video, audio, text, or some combination thereof – produced by generative artificial intelligence (GenAI).[1] GenAI is any digital technology which identifies patterns in (large volumes of) data and uses these patterns to produce novel outputs when prompted by users – that is, outputs which need not have appeared in that form in the input data.

To give a sense of the technologies available, myriad software applications now allow quick and easy production of synthetic (audio-)visual content. For example, Reface superimposes faces onto GIFs, while FaceMagic can swap them into images and videos, and Avatarify adds the possibility of overlaying one's own voice. The last few years have also seen a huge step change in the capabilities and applications of large language models (LLMs). For example, OpenAI's ChatGPT, Anthropic's Claude, and Google's Gemini now output convincing-sounding text in response to natural language queries.[2] The most advanced systems are multimodal: for example, Dall-E uses GPT software to make fake images in response to users' natural language prompts.

Seen from one angle, GenAI might be thought not to mark any radical departure from older technologies, such as photo editing, computer-generated imagery (CGI), or text auto-complete. These are also methods for producing manipulated content, with end products that are sometimes qualitatively similar to those of GenAI. What distinguishes GenAI is the way in which it analyses and synthesises information so independently of human control. As a result, its operations resemble processes of learning and creativity that were previously thought proprietary to human beings (and perhaps some other non-human animals). So, whereas photo-editing and CGI software are paradigmatic tools in the hands of human users, augmenting but not replacing our agency, the likes of Reface, FaceMagic, and Avatarify rely far less on anything the user does, and far more on insights acquired by the model itself, during its period of operation. Meanwhile, although it is true that LLMs deploy predictive principles reminiscent of simpler auto-complete tools, they have been extended vastly, to produce whole sets of propositions about a topic (not just a user's next word).

These features make GenAI technologies genuine candidates for (co-)authorship of content. Added to their accessibility (being cheap and easy for laypeople to use) we can expect them to play an increasingly important role in our (increasingly online) media. We must grapple with what this means for the public sphere – the nature of public discourse and its impact on politics and society. The first question to address is whether we can safely allow synthetic content to go 'incognito' on public platforms, or whether it should be distinguished from other content. This is the question I set out to answer.

3. Tracking Truth

The first issue to consider concerns the effect on audiences' processes of belief-formation of integrating human and synthetic content. We might hypothesise as follows:

> *Epistemic hypothesis*: Synthetic content must be identifiable if we are to optimise our epistemic position, such that we tend to believe truths and disbelieve falsehoods.

Is the hypothesis plausible? That depends in turn on one or more of the following conditions obtaining: (i) synthetic content is more likely than other content to be inaccurate; (ii) it is more likely to mislead us when it is inaccurate; (iii) it demands distinctive forms of epistemic engagement from us.[3] I take each point in turn.

3.1. (In)accuracy

Is synthetic content more likely than other content to be inaccurate? In many domains, I believe it is. GenAI is essentially unconstrained by reality (in a way that, say, photography is not) and it lends itself to portraying unreal phenomena.

Consider, first, that GenAI works by applying low-level patterns in past data to new problems. It does not simply regurgitate content it has ingested (and so could not guarantee accurate outputs, even if trained only on accurate data). Instead, the essentially inductive nature of the technology makes a certain amount of epistemic shakiness inevitable.

Moreover, software developers have so far made only limited attempts to inhibit the production of inaccurate content[4] – and there is probably no prospect of eliminating it entirely. As a result, the content produced by GenAI need not – and often will not – correspond to how things are in the world.

At the same time, a major appeal of the technology is its ability to produce realistic portrayals of things that never happened. These include, for example, images of non-existent individuals, audio of speeches that were not made, and videos of events that did not occur. As we saw above, the purpose of an app like FaceMagic is to allow users to visualise body parts being combined in ways that are not instantiated in reality. Whatever the reason for producing such media (which could be merely for entertainment, or for less wholesome pornographic or propagandic purposes), they are distinguished precisely by *not* being accurate representations of reality. Mistaking them for such will only worsen the audience's epistemic position, leading us to believe things that are not true. Labelling content as synthetic would help mitigate that risk by inviting us to withhold credibility concerning what is portrayed.

It might be objected that GenAI can equally well produce relatively accurate representations, as when historical events are reimagined on the basis of synthesising reliable information about what took place. The tendency for synthetic content to be inaccurate, then, is contingent on how it is used (along with the nature of its training data).

In response, I suspect that audio and visual technologies will continue to be used predominantly for portraying non-reality. However, the situation may be different for large language models. LLMs certainly do have great potential for fabrication too – whether for the production of dungeon master scripts or propagandic narratives. However, they will often produce truths in response to empirical queries (depending on the nature of the training data and the user prompt, as we will see below). Because of this, they are

beginning to be used in arenas like internet search, customer service, and education. Here, at least, developers are likely to fine-tune the models in ways that improve their accuracy; and users are likely to expect – and demand – high quality information. In other words, the pressures governing the development of LLMs for such applications increase the chances of their text being accurate; and give us a correspondingly weaker epistemic case for labelling it, relative to other kinds of synthetic content. Nevertheless, we will see that the cumulative case is still sufficient.

3.2. Convincingness

When synthetic content *is* inaccurate, it is likely to have a relatively powerful epistemic effect, as compared with other kinds of content. This is because GenAI generally produces more convincing results than other, more manual technologies.

As we saw above, synthetic audio and visual content is often highly realistic, being qualitatively indistinguishable from media produced through genuine recordings of aspects of the world. For example, it can be impossible to tell if a given image is a photograph or a product of GenAI; or if a given piece of audio is recorded or synthetic.

It might be objected here that there are some situations in which we *can* tell that we are facing inaccurate synthetic content, as when it is used for satire, parody, or entertainment. The worry, then, is that labelling content in these contexts would be too heavy-handed, spoiling the fun as it were.

In response, it seems right in principle that only synthetic content which plausibly represents reality need be labelled on epistemic grounds – that is, synthetic content which reasonable audiences could take to be accurate. So, a synthetic image of a lettuce with the face of former UK prime minister Liz Truss, for example, or a synthetic video of a cat sprouting wings and taking to the air, would not require labelling. In contrast, a synthetic image of Liz Truss wearing a lettuce suit, or a synthetic video of a cat being abused, are (sadly) plausible enough to warrant labelling. The argument from inaccuracy, then, only supports a policy of labelling *plausible* synthetic content. Having said that, I suspect that such a policy may prove impossible to implement in practice, and we may be better off labelling *all* synthetic content than *none* (I return to consider the costs of this approach in Section 5).

Crucially, we are accustomed to trusting contents with the qualitative properties of photographs, videos, and audio recordings. After all, when produced by recording technologies they are constrained to be responsive in particular mechanical ways to aspects of reality (light, in the case of photographs, or sound waves in the case of audio recordings). Even if these technologies give us only partial and indirect access to reality, we can often rely on the broad accuracy of what they portray (allowing for some artificial enhancement at the margins). This is not true for the kinds of GenAI applications described above, which are essentially untethered from reality.

The result is that, when we mistake synthetic images for photographs, or synthetic audio for real recordings, we are likely to believe in what they portray. In contrast, we are less likely to trust contents that are more obviously removed from reality, like human drawings. At most, the latter may provide testimonial evidence (akin to someone's assertion of what happened) rather than the kind of perceptual evidence that genuine photographs provide (which immunises them from insincerity or factual error).[5] It is the propensity for synthetic contents to exhibit what we might call 'realistic fantasy' which makes them so

liable to mislead. Traditional content has tended to have one or other of these two properties – either being realistic and broadly accurate (like photographs) or unrealistic and potentially inaccurate (like drawings) – but not both realistic and inaccurate together.[6]

Again, the situation is somewhat different for LLMs. There is nothing equivalent to recording technology when it comes to text – there are no means of mechanically capturing aspects of the world in written form comparable to cameras or audio recorders. Therefore, the problem of realistic fantasy does not arise in quite the same way for synthetic text as for visual or audio content. The appeal of LLMs has less to do with their ability to produce realistic portrayals as to produce these just as fluently as humans do (and without all the effort). The point, in other words, is that synthetic text is indistinguishable from human speech.

Yet LLMs too have the potential to be extremely convincing. Even at this early stage, there is emerging evidence to suggest that they already surpass most ordinary people's abilities to produce persuasive text, tailored effectively to particular audiences.[7] Thus, synthetic text, like other kinds of synthetic content, may be particularly prone to mislead us when it is inaccurate. This makes labelling a prudent course of action, which can remind us to be on our guard when the author is an LLM rather than a human. In sum, the convincingness of synthetic content provides a distinct epistemic reason for labelling it. Even in situations where it is no more likely than other kinds of content to portray non-reality, the manner in which it does so is more dangerous, making false beliefs especially hard to resist.[8]

Perhaps over time we *could* learn to resist, withholding belief whenever we encounter content, regardless of how realistic or persuasive it might be. As pointed out by Regina Rini, though, this is not something we should aim for.[9] Once we lose trust in the information ecosystem as a whole and begin to suspect that any content we encounter is fake, false, or misleading, this indirectly damages our epistemic position, since it inhibits our acquisition of true beliefs from accurate content. Moreover, it is a behavioural response that is easily exploitable by bad actors. For instance, genuine photos or recordings might be insincerely accused of being fakes, in order to maliciously cast doubt on what they accurately portray (and, perhaps, to discredit those who circulate them). These deleterious downstream consequences add further grist to our mill, since labelling synthetic content would help audiences target their scepticism towards the most epistemically risky contents, while shoring up their trust in the rest of the information environment.

3.3. Distinctiveness

In the previous two subsections, I examined why synthetic content might be more likely than other content to mislead us (either because it is more likely to be inaccurate, or because it is more likely to be convincing when inaccurate). However, one might object at this point that human content can be just as misleading. We only need to think about the frequency with which speakers lie, bullshit, withhold the truth, or offer partial representations of reality, to appreciate the epistemic risks involved in ordinary conversation. So, when it comes to text, does it really make sense to flag up the synthetic as raising special epistemic concerns?

I think it does. LLMs err in different ways, in different situations, and for different reasons than humans do. Human speakers tend to make false statements when they are

mistaken about the truth, or wish to communicate something false (i.e. when they are lying or misleading), or are altogether indifferent to a proposition's truth or falsity (i.e. when they are bullshitting).[10] As audiences, we are relatively accustomed to evaluating our interlocutors' level of expertise and likely motivations, in order to decide how much credibility to assign to their utterances. LLMs, on the other hand, are driven by quite different forces, which require us to apply a different suite of credibility checks.

Consider, for example, the much-discussed phenomenon of 'hallucination', whereby LLMs appear to invent objects, individuals, or events, and present them as real. For example, ChatGPT is known to output citations and weblinks that look plausible but fail to map to any actual sources. This behaviour is unlike human lying or misleading in that the AI has nothing identifiable as an intention to deceive. The behaviour is somewhat closer to bullshitting, in that neither truth nor falsity come into the equation at all; the AI, like the bullshitter, is merely producing speech that comes across as relevant and plausible, regardless of whether or not it is true.[11]

Usually, though, human bullshitters have some non-epistemic motivation for what they say, like impressing their audiences. By tracking such motivations, we are sometimes able to spot their bullshit. LLM hallucinations cannot be spotted in the same way. They result from a process of generating sequences of words according to statistical probabilities (calculated with respect to an enormous number of dimensions). Once we understand that this is what the algorithm is doing, we can begin to gain a sense of when and why inaccuracies appear in synthetic text. Citations are prime examples: an article's title, author, and publishing details generally cannot be derived probabilistically (unless, perhaps, the source happens to have been cited, in full, enough times in the LLM's training data). Rather, this is the kind of information that requires targeted retrieval from an authoritative source (such as a library search).

Other kinds of information might fall into the category of requiring this kind of targeted retrieval, including discrete facts about individuals and organisations. Given the current state of the technology, we would do well to avoid relying on GenAI for such information (or at least take it with a large pinch of salt).[12] Similar points could be made with respect to certain reasoning tasks, where the current cohort of LLMs fail in highly distinctive ways.[13] Thus, we need to apply different rules of epistemic engagement when it comes to LLMs.[14]

Again, then, labelling synthetic text can give us the opportunity to adjust our epistemic engagement appropriately. If we know that we are dealing with human speech, we can take into account the speaker's potentially dubious motivations. If we know that we are dealing with synthetic text, we might bear in mind its (over-)reliance on probabilistic inferences.

An immediate problem with this proposal is that few laypeople have a good sense of which epistemic checks are appropriate when dealing with synthetic text. Even worse, the workings of LLMs remain highly opaque even to their developers and other technical experts. It might be objected, then, that despite differences in human and synthetic inaccuracies, labelling won't help us much in tailoring our epistemic responses.

In response, I remain cautiously optimistic that labels can provide valuable epistemic signals, even in the short term. I suspect that we will soon become sensitive to LLMs' characteristic errors, and we will do so without needing to know the full details of their internal workings (just as we have become sensitive to many aspects of human behaviour despite having only a very limited understanding of the mind). No doubt there will always be some limitations on our abilities to track the trustworthiness of synthetic text (just as there are for human text) and thus there will be some instances where it is not epistemically

advantageous to know the provenance of the text. On the whole, though, labelling is epistemically useful.

In this section, we have seen that synthetic content is generally more likely than other content to be inaccurate (especially when we are dealing with audio and video content), to be persuasive (unless it is obvious from the context that the content is not supposed to represent reality), and to go wrong in distinctive ways (especially when we are dealing with verbal content). These three points combine to make a strong epistemic case for labelling synthetic content (perhaps excluding cases of satire, parody, or entertainment, where the content is anyway implausible). In the next section, I develop a separate but complementary line of argument, which further strengthens the case for labelling.

4. Minding Minds

Accuracy is only one feature that makes content valuable. Another is its embodiment of individual expression. The second hypothesis concerns our interest in tracking this:

> *Expressive hypothesis*: Synthetic content must be identifiable if we are to give appropriate recognition to content's expressive value.

It often matters to us whether a given artefact is the product of a human mind. For example, artworks, theories, and record-breaking attempts impress us in a different way from natural or mechanical phenomena (like sunsets, plant behaviour, digital calculations, and factory-produced consumer goods). Of particular relevance to the current discussion is how we value humans' *creativity*; and their *free expression*.

Before addressing each point, let me briefly deal with a sceptical worry lurking in the background. One might wonder what, if anything, is so special about human or human-like minds as to endow them with expressive value that machines lack. Naturally, this is a vexed philosophical issue, which raises profound questions about the nature of the mind. According to the sceptic, there is nothing special about minds. Once we know enough about how they work, we will see that they are evaluatively (and perhaps even descriptively) equivalent to suitably complex machines.[15]

While I won't attempt a complete response to the objection, I believe there are at least two *prima facie* reasons for evaluating human products differently from those of machines. First, humans are conscious beings with subjective experience, enabling them to engage with the world in a way that machines cannot (at least not yet). Second, being socially situated, human persons are accorded a normative status in our collective institutions that machines lack (at least for now).[16] We will see how these two points play into considerations about expression, which will occupy us in the remainder of this section.

4.1. Creative Expression

Imagine we were to discover that Pablo Picasso's *Guernica*, Federico Fellini's *La Dolce Vita*, or Jane Austen's *Emma* were in fact produced by GenAI. I take it this would substantially change – and probably diminish – our appreciation of the works. They could no longer be understood as the products of minds like ours, situated in a particular set of cultural traditions and societal zeitgeists. If this is right, it indicates that the provenance of works of art and literature, and not their mere existence, contributes towards their significance for

us. It is relevant that they have been conceived by people with thoughts and ideas – and executed through their effort and skill – rather than the process being an automated one. A novel written by an LLM lacks an author's imaginative labour and skilled use of literary devices. Likewise, an image produced by DeepArt lacks a visual artist's individual expression and deployment of materials, techniques, and effects. Part of the appreciation of human art, then, is an appreciation of the artist's thought processes – the artist's *mind*. The value we assign to their mindful creative expression would appear not to accrue to GenAI.

Granting this point, a piece of synthetic content resembling art may be assigned less value than a qualitatively identical piece of human art – or at the very least, may be valued in a different kind of way.[17] Labelling art-like synthetic content could stop us mistaking it for human art and engaging with it inappropriately. For example, we might refrain from attributing to its producer certain fine-grained skills, depths of feeling, or incisive cultural awareness, to be appreciated by the audience. Instead, we might focus on the content's effects on us, the imaginativeness of the user's prompt, or the complexity of the underlying technology and the corresponding skill of the developer.

The long-term effect of uncertainty about the provenance of artistic content could be to stop us from properly appreciating *human* art, failing to give due recognition to artists' emotional and imaginative engagement. One reason for labelling synthetic content, then, is to help protect our existing practices of creative expression and appreciation.[18]

It should be acknowledged, of course, that only a relatively small subset of the content we encounter on a daily basis is construable as art. Most is far more prosaic in nature, providing information, expressing opinion, or selling products. Likewise, when social media platforms prompt users to post what is on their mind, this typically elicits life updates, commentary on news items, details of events and opportunities, and so on. Such contents are generally not the kinds of creative expressions that carry artistic value (although we typically still assign them free speech value, as will be discussed below).

We might wonder, then, how far the argument from creative expression extends. On the one hand, it gives us a *pro tanto* reason to label art-like synthetic content (and arguably a larger than usual proportion of synthetic content could end up being art-like, precisely because GenAI invites users to explore imaginative and aesthetic possibilities). On the other hand, the argument from creative expression is at once weak (since the harm of appreciating art in the wrong way is less severe than other physical and psychological speech harms) and limited (giving us no reason to label synthetic content that performs more prosaic functions). Let us turn, then, to the value associated with expression *per se*.

4.2. Free Expression

Does the value we place on free expression read across from human to synthetic content? The considerations here will depend in part on how that value is grounded.[19] While instrumentalist defences of free speech appeal to its positive effects (such as getting us closer to the truth or supporting democracy), others appeal to its intrinsic fittingness to the human condition (respecting as it does the autonomy of persons).

Consider first Millian defences of free speech, which see it as essential for uncovering truth and upholding truth on the basis of reasons.[20] At first blush, the provenance of content seems entirely irrelevant to that project. As long as diverse perspectives are available for examination, it doesn't matter whether they came from humans or from GenAI; our

task is to engage with the ideas, not their authors. Indeed, GenAI gives us *more* content, which can surely only be a good thing. On this approach, there might seem to be no reason to label synthetic content: the ideas contained therein demand just the same engagement as those put forward by humans (whereas labelling them risks implying that they have a different status).

On further reflection, though, we might worry that synthetic content will end up unjustifiably skewing opinion. The way GenAI responds to statistical probabilities in the training data seems likely to favour and reinforce majority ideas, is surfacing the more heterodox ones Mill was interested in. Moreover, as the models feed off one another, the ideas they present may become homogeneous and stale. Far from facilitating a vibrant exchange of ideas, then, increasing reliance on synthetic content could actually stifle intellectual progress. The risks here seem worse than potentially implying, through labelling, that a piece of synthetic content is less worthy of serious consideration. Therefore, I believe there is a reasonable Millian case for labelling synthetic content. Doing so would help individuals seek out ideas from a wider range of human *and* machine sources.

Other defences of free speech focus on the nature of speakers themselves, most prominently by appealing to their autonomy.[21] Autonomous agents are those who can freely choose which ends to pursue. This includes their choosing what to express and how to do so. Whereas we are morally obliged to respect human autonomy, GenAI is a mere technology, incapable of pursuing its own ends (at least for now). There is therefore no autonomy-based reason for protecting its expression. If, for instance, we were to discover that a particular viewpoint (e.g. 'Brutalist architecture should be protected at all costs' or 'The UK voting system should be changed to proportional representation') were being promulgated by LLM bots, rather than actual people, we would be justified in discounting it, since bots do not get to have a voice or a vote in such matters. That gives us a perfectly straightforward reason to label their output – namely, to avoid mistakenly assigning it autonomy-based expressive value.[22]

One might wonder, though, whether synthetic content is attributable to the humans who developed the technology – and thus protected by *their* free-speech rights. I don't think this can be right. The fact that the software has been programmed by humans is insufficient reason to attribute its eventual contents to them, given the relative independence with which GenAI operates and evolves.

Something similar goes for the human users who issue more proximal prompts: again, they lack sufficient control over the content produced to be deemed its (sole) author. Even when the prompts issued are quite specific, many fine-grained decisions about the precise form of the output are still delegated to the model. Therefore, respecting the autonomy of human participants does not require synthetic content to be protected as their free speech – or not in that precise form.

Arguing from autonomy, then, synthetic content lacks the expressive value of human content. The upshot is that it ought to be labelled. That would help us engage appropriately with the ideas we encounter online, giving weight to those put forward by human agents, while treating those produced by machines as mere explorations of logical space, available for inspiration perhaps but not eligible for the same free speech protections.

The situation is complicated slightly by the fact that synthetic content is often distributed by human users rather than bots. Of course, where human users are simply sharing synthetic content, there seems no reason to remove the label from the latter (while there would equally be no reason for labelling any additional content the user overlays). On

the other hand, where a human user adopts a piece of synthetic content as their own (as when text from ChatGPT is copy-pasted into a social media post) that becomes a genuine part of their expression. As such, it is attributable to them and derives free speech value from their autonomy. In that second kind of case, there is no argument for labelling (and nor would it be feasible to enforce).

I end this section by briefly considering one more potentially deleterious downstream consequence of allowing synthetic content free rein – namely, the indirect threat to our range of free (speech) action.

4.3. Speech Acts

Our ability to perform speech acts depends on the existence of social practices governed by sets of regulatory norms.[23] For example, our ability to make assertions depends on the existence of a practice in which speakers are expected to be *sincere* (roughly, saying only what they believe to be true). Of course, speakers often abuse the practice by making insincere assertions. However, enough of us make sincere assertions enough of the time to cope with some free-riding on the practice. If that ceased to be the case, though, and the sincerity norm were to fall away altogether, we would no longer be able to maintain our current practice of assertion. Then, for example, audiences presented with utterances like 'Vaccination reduces the risk of getting seriously ill from COVID-19' or 'The election will take place on 30th November' could not expect speakers to be accurately representing reality (or even attempting to do so) but merely producing well-formed declarative sentences for some other purpose than the communication of facts. Speakers, in turn, could not expect their audiences to believe the propositional contents of such utterances. The same basic point generalises to other speech acts, like promising, asking, and arguing – each is governed by a particular set of norms, the removal of which would make the associated social practice unavailable for participation.

The worry with GenAI is this: since it does not participate in norm-governed speech practices, but is at best parasitic on them, it risks undermining those practices in precisely the way described above. While GenAI may produce images, videos, or text, it does not do so from a position of being directly embedded in societal institutions. LLMs, for example, are not subject to the same social pressures as we are and have little incentive to make truthful assertions. They not only lack the requisite human intentions for sincerity but their design makes truth an incidental property (as described in Section 2). Nor does the technology have any interest in the continued existence of human speech practices. LLMs are not free agents with interests in being able to make assertions. In these respects, they differ importantly from humans.

Crucially, integrating synthetic content with human content poses a threat to our speech practices, similar to inserting fleets of undercover bullshitters into a communicative situation. If we can't tell when text is synthetic or human, we have no reason to assume *any* speech is part of a norm-governed practice. We stop expecting sincerity from speakers, just as we stop expecting trust from audiences. Eventually, the norms are eroded to such an extent that our practice is undermined and we can no longer perform assertions. The argument generalises to other speech acts, constraining our ability to do the things we want to, like promising, asking, or arguing. By that point, our freedom of action would have been severely curtailed.

The risk described in this subsection is admittedly speculative. However, it serves to identify another strand to the expressive argument for labelling synthetic content, which warrants further consideration (including to see whether and how LLMs might be co-opted into our speech act practices).[24] The suggestion for now is that labelling would help us interpret speech in the correct light, that is, as falling within or potentially outside human speech act practices and subject to correspondingly different sets of norms.

This section has argued that labelling synthetic content is important for isolating human creativity and expression, in order to assign it appropriate value. In particular, labelling could support our appreciation of art and literature and – even more fundamentally – our respect for free expression.

5. Conclusion

I have argued that labelling synthetic content is a philosophically justified measure, which helps mitigate a variety of epistemic and expressive risks. As such, it would seem to represent a core component of good content curation on social media platforms. Indeed, this might seem platitudinous – and largely moot, now that major tech companies have already embraced labelling as a key part of the solution to the problems of GenAI. Let me end by explaining why the argument remains crucial.[25]

The first point concerns feasibility issues. Inevitably, some synthetic content will fail to be labelled. Some social media users will fail to disclose the true provenance of the content they post, either deliberately or by mistake. Furthermore, in many of those instances the provenance of their content will be hard for platforms to trace. Not all synthetic content can be expected to contain watermarks, which may have been missing from the outset (if the content has been created using a rogue application) or stripped out after the generation of the content (if the in-built watermarking system is insufficiently robust against 'jailbreaking'). As a result, swaths of convincing synthetic content may go unrecognised. Evidently, the practical difficulties of enforcing labelling duties on platform users and upstream software developers are significant. One risk is that the tech companies will simply give up on the initiative after a period of time. The arguments developed in this article give them reason to persevere.[26]

Another risk is that social media users' trust in labels ends up exceeding the latter's reliability – that is, if far more content is believed to be correctly labelled than is actually the case. The upshot could be that large amounts of fake, false, or misleading content is taken at face value – and meanwhile, perhaps, authentic and accurate content is not. (Alternatively, trust could be out of step with reliability in the opposite direction, with generally accurate labelling being met with widespread scepticism.)

This brings us to a more general point about the effectiveness of labels as guides to appropriate epistemic and interpretative engagement. As they are rolled out, an important task for empirical research will be to investigate how users actually respond to them. Will we trust them? Will we become unduly complacent about the quality of the epistemic environment? Or more pervasively sceptical? Will we engage with labelled content differently? If so, how? The answers to these questions will be crucial in assessing whether the philosophical arguments put forward above continue to justify labelling in practice.

Finally, let us turn to the costs associated with labelling. Clearly, there are financial costs for tech companies in developing and implementing labelling regimes. While the

major players may be well able to bear these, smaller ones may not. Moreover, it is always possible to ask whether their resources would be better targeted elsewhere, to tackle potentially more pressing content curation and moderation problems. I have not attempted the kind of comparative analysis that would provide a satisfying resolution to this worry.

Additionally, labels imply friction costs for social media users. In a sense, this is the point of them – labels must hijack users' attention if they are to be of any use at all. Yet having to apply labels to the content one posts, and to cognitively process those applied to others' content, inevitably diminish the user experience somewhat. As we saw earlier, the worry may be especially stark where satirical, parodic, or entertaining synthetic content must be labelled.[27] The question then becomes: to what extent are the amusing effects of the content undermined by labelling; and is that loss a price worth paying for the avoidance of potentially harmful misinterpretations? While the first part of the question is straightforwardly empirical, the argument I have put forward in this article ought to help us think through the second – normative – part.

These brief concluding remarks on the feasibility, effectiveness, and costs of labelling show that the argument is not yet won. While there is a strong case for labelling, it is not free of risks. Indeed, given the early stage of GenAI, I would suggest that the case must be revisited regularly as the technology improves (perhaps acquiring greater epistemic reliability in the process) and as better information emerges (including about the uses and effects of synthetic content). That evolving understanding must also inform the development of other risk-mitigation actions (since labelling alone is unlikely to be sufficient) which should be pursued in combination with platforms' existing content moderation practices. The epistemic and expressive issues discussed in this article will be fundamental to that further task too.

Sarah A. Fisher, Department of Political Science and School of Public Policy, University College London, London, UK. *sarah.a.fisher@ucl.ac.uk*

Acknowledgements

This work was supported by UK Research and Innovation (grant reference MR/V025600/1). I am indebted to Jeffrey Howard and Beatriz Kira for insightful comments on an earlier version, as well as to audiences at the 97th meeting of the American Philosophical Assocation - Pacific Division and a workshop on 'Social Media Corporations: Risks, Rights and Responsibilities', held in Oxford in May 2024. Finally, thank you very much to a reviewer for this journal, whose wonderfully helpful comments helped strengthen the argument significantly.

Conflict of interest

No conflicts of interest.

NOTES

1 I use 'synthetic content' rather than 'synthetic media' for two reasons. First, 'synthetic media' is often used in ways that exclude textual media. Second, one of my aims in the article is to inform real-world practices of content curation and moderation. While it might be objected that 'content curation' and 'content moderation' are misnomers, which fail to respect philosophical distinctions between utterances and contents, and contents and forces, I suspect it is too late to impose new terminology on the debate.

2 As pointed out by Shanahan *et al.*, "Role-play," these applications are technically 'dialogue agents', whose functionality extends the underlying foundation models in particular ways. For my purposes here, however, I will refer to them as LLMs.

3 It should be noted that not all content can be exhaustively categorised as accurate or inaccurate (at least, given our imperfect knowledge, and perhaps also at a deeper metaphysical level). This includes normatively controversial propositions, which will become relevant when we turn to expressive concerns in sect. 4.

4 Some large language models now provide direct links to sources, which is at least a step in the direction of verifiability.

5 See Cavedon-Taylor, "Photographically Based Knowledge," for further discussion of the distinction between photographs and drawings as epistemic sources that are perceptual and testimonial, respectively. Thank you to a reviewer for directing me to this helpful discussion.

6 What about high quality fakes produced using older technologies, such as manually doctored photographs? The argument presented here applies equally to them, and gives us equal cause for labelling. Note, though, that the complete case I develop here for labelling synthetic content is presented as a cumulative one, which depends on the suite of arguments provided and not any individual one. Therefore, it does not follow from what I say in this subsection that manually doctored content should also be labelled, all things considered. It might be, for example, that its being more effortful and costly to produce reduces the likelihood of harm to such an extent that it would be disproportionate to enforce a duty to label. I will remain neutral on this issue for now, since it is not my primary focus here.

7 See, for example, Matz *et al.*, "The Potential of Generative AI for Personalized Persuasion at Scale"; Salvi *et al.*, "On the Conversational Persuasiveness of Large Language Models: A Randomized Control Trial"; Shin and Kim, "Large Language Models Can Enhance Persuasion Through Linguistic Feature Alignment." Thank you to a reviewer for pressing me on this point.

8 Harris argues in "Beyond Belief" that residual persuasive effects can persist even once we know a portrayed event did not happen. If correct, this would suggest that mere labelling of synthetic content may not be adequate; rather, there is a case for removing (or reducing the reach of) the misleading content. Given the empirical uncertainty about subliminal effects, I restrict my focus here to arguing that synthetic content should – at a minimum – be labelled.

9 Rini, "Weaponized Skepticism."

10 Here I skate over large literatures on how to define lying, misleading, and bullshitting. For overviews, see Stokke's "Lying, Deceiving, and Misleading," and his *Lying and Insincerity*. The intricacies of these debates are not important for the points I wish to make.

11 In their "Role-play," Shanahan *et al.* suggest that we think of LLMs as engaging in *role-playing*, which may be a better framing.

12 Moreover, where a lot rides on getting such details right (for example, to avoid defamation or market manipulation), it seems unwise to rely on technologies that use brute statistical power to get to answers.

13 For an entertaining example, see Smith, "Man."

14 Looking towards future iterations of the technology, credibility may depend on how effectively developers can 'ground' the outputs, for example by cross-checking against reliable sources, or by overlaying deductive computational capacities.

15 For arguments in this vein, see Dennett, *Intentional Stance*.

16 Although see Rini, "Talking Cure," for an argument that LLMs should be accorded this kind of normative status.

17 What about art produced with the aid of non-generative technologies? Insofar as these technologies scaffold, rather than replace, the artist's imaginative or aesthetic engagement, they need not diminish the creative value. So, a poet's use of a word-processor would not diminish the value we assign to her creative expression. However, her reliance on grammar-correction software probably would, since poetry as an art form depends essentially on the (often unconventional) ways in which words are put together. No doubt there will be grey areas in between these two extremes.

18 The argument presented here is orthogonal but complementary to current attempts to safeguard artists' intellectual property by regulating the data inputs to GenAI.
19 For an overview of the current state of the debate, see Bonotti and Seglow, "Freedom of Expression"; Howard, "Freedom of Speech."
20 Mill, *On Liberty*.
21 See, for example, Baker's "Harm" and "Autonomy."
22 The same argument applies equally to other kinds of inauthentic accounts, which are not powered by GenAI but may be run by human operatives (perhaps with the aid of more established technologies). In fact, there may be good reasons for removing potentially deceptive bot accounts from online platforms altogether, rather than merely labelling their outputs. However, I will not argue for this further measure here.
23 See Austin, *How to Do Things with Words*.
24 For some related discussion, see Cappelen and Dever's *Making AI Intelligible* and their "AI with Alien Content."
25 Many thanks to a reviewer for encouraging me to do so.
26 Perhaps they will ultimately decide that labelling *human* content is a better bet than labelling *synthetic* content. It would be an interesting future project to examine what further philosophical implications that flipped approach might have.
27 A reviewer reminds me of Facebook's failed attempts to introduce 'satire' labels in 2014.

References

Austin, J. L. *How to Do Things with Words*, 2nd ed. Oxford: Oxford University Press, 1975.
Baker, C. Edwin. "Harm, Liberty and Free Speech." *Southern California Law Review* 70, no. 4 (1996–7): 979–1020.
Baker, C. Edwin. "Autonomy and Free Speech." *Constitutional Commentary* 27, no. 2 (2011): 251–280.
Bonotti, Matteo, and Jonathan Seglow. "Freedom of Expression." *Philosophy Compass* 16, no. 7 (2021): e12759.
Cappelen, Herman, and Josh Dever. *Making AI Intelligible: Philosophical Foundations*. Oxford: Oxford University Press, 2021.
Cappelen, Herman, and Josh Dever. "AI with Alien Content and Alien Metasemantics." In *The Oxford Handbook of Applied Philosophy of Language*, edited by Luvell Anderson and Ernie Lepore, 573–593. Oxford: Oxford University Press, 2024.
Cavedon-Taylor, Dan. "Photographically Based Knowledge." *Episteme* 10, no. 3 (2013): 283–297.
Dennett, Daniel C. *The Intentional Stance*. Cambridge: MIT Press, 1989.
Harris, Keith Raymond. "Beyond Belief: On Disinformation and Manipulation." *Erkenntnis* (2023). https://doi.org/10.1007/s10670-023-00710-6.
Howard, Jeffrey W. "Freedom of Speech." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman. Spring, 2024.
Matz, S. C., J. D. Teeny, S. S. Vaid, H. Peters, G. M. Harari, and M. Cerf. "The Potential of Generative AI for Personalized Persuasion at Scale." *Scientific Reports* 14 (2024): 4692.
Mill, John Stuart. *On Liberty*. Indianapolis: Hackett Publishing, 1978.
Rini, Regina. "Weaponized Skepticism: An Analysis of Social Media Deception as Applied Political Epistemology." In *Political Epistemology*, edited by Elizabeth Edenberg and Michael Hannon, 31–48. Oxford: Oxford University Press, 2021.
Rini, Regina. A Talking Cure for Autonomy Traps: How to Share Our World with Chatbots [unpublished manuscript].
Salvi, Francesco, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial arXiv:2403.14380v1 2023. [preprint].

Murray, Shanahan, Kyle McDonell, and Laria Reynolds. "Role-Play with Large Language Models." *Nature* 623 (2023): 493–98.
Shin, Minkyu, and Jin Kim. Large Language Models Can Enhance Persuasion through Linguistic Feature Alignment 2024. https://ssrn.com/abstract=4725351
Smith, Gary. A Man, a Boat, a Goat – and a Chatbot! Mind Matters, May 15, 2024. https://mindmatters.ai/2024/05/a-man-a-boat-and-a-goat-and-a-chatbot/
Stokke, Andreas. "Lying, Deceiving, and Misleading." *Philosophy Compass* 8, no. 4 (2013): 348–359.
Stokke, Andreas. *Lying and Insincerity*. Oxford: Oxford University Press, 2018.