Is AI-generated Extremism Credible?

Experimental Evidence from an Expert Survey

Stephane J. BAELE – UCLouvain, Belgium (corresponding author: stephane.baele@uclouvain.be)

Elahe NASERIAN – Trilateral Research, UK

Gabriel KATZ – University of Exeter, UK

This publication stems from the ExID project, which is financially supported by UKRI-ESRC (2020/SGW/00884431) involved in the NORFACE Joint Research Program Democratic Governance in a Turbulent Age, itself co-funded by the European Commission through Horizon 2020 under grant agreement No. 822166.

Rapid progress in Artificial Intelligence (AI) over the past few years has unfolded important questions – and triggered anxiety – about the potential (mis)use of AI-generated content in information warfare and disinformation campaigns. A crucial question remains unanswered, however: whether – and to what extent – this technology can be harnessed in a credible way by extremists, whose language is usually characterized by highly specific subcultural registers that make them unlikely to be easily mimicked by AI models. This paper answers that question, offering the first rigorous evaluation of the credibility of synthetic (AI-generated) extremist prose. Using an expert survey experiment measuring the credibility of fake incel forum posts and ISIS magazines paragraphs generated through an original workflow for the production of synthetic extremist text, we show that these texts have high credibility scores, confusing world-leading experts. These findings, which are discussed under the broader light of the emerging literature on nefarious "dual-use" of synthetic content, not only define and evaluate the threat of extremists harnessing language models for propaganda, they also more broadly strengthen our understanding of the role language models are due to play in hostile political communication, and ultimately resurface the debate on the human costs of technological progress.

Introduction

Artificial Intelligence (AI) is ubiquitous, be it when interacting with Amazon's Alexa or when chatting with an online customer service; a significant proportion of stock trading now relies on it, as do autonomous vehicles and surveillance systems in critical infrastructure. It would be naïve to assume that politics, and political communication more specifically, somehow escapes this evolution.

Accordingly, intense debates have emerged focusing on the complex ethical challenges associated with the rapid deployment of AI in socio-political contexts. The roots of these concerns often run deep into fundamental issues such as racial justice, gender equality, or human rights.[1] One of the most imminent and critical of these challenges pertains to the potential of AI-based communication tools to create, accelerate, or consolidate political polarization, in other words the strategic harnessing of these instruments by radical political actors or their backers aiming to gain support, fragment society, and foster conflict. While video deepfakes have so far drawn most of the attention of the nascent literature investigating these problems,[2] the present paper focuses on another type of deepfake – broadly defined here as AI-produced content – which has so far largely escaped scholarly scrutiny despite its greater accessibility, less expensive use, and more advanced stage of development: fake text generated by large language models (LLMs).

Language models can be defined as statistical models that calculate probability distributions over sequences of words;[3] the autocomplete function of messaging apps is only one of the numerous applications of language models already running in everyday life. In contrast to video deepfakes' evident potential for political misuse, language models may appear harmless; a sizeable proportion of the journalistic coverage of the technology, especially following the launch of the game-changing ChatGPT interface, has played with its comical functions of generating poetry, philosophical aphorisms, or even new rap lyrics mimicking famous artists.[4] This paper demonstrates that language models can also, unfortunately, be used for less amusing purposes; when harnessed by extremists, they can – and soon will – become powerful tools in radical communication strategies. The first materializations of this possibility are already popping up in a dispersed way, and the threat is therefore slowly making its way into governments' scenarios: the US Congress' National Security Commission on Artificial

Intelligence's (NSCAI) voluminous report, for instance, worried about the "malign uses" of various types of deepfakes by "strategic competitors" and "rogue states, terrorists, and criminals",[5] and the whole purpose of the ground-braking European Union "AI Act", adopted in March 2024, is to prevent as much AI dual-use as possible in disinformation campaigns.[6] The 2023 AI Safety Summit, and the ensuing Bletchley Declaration,[7] marked the entrance of this issue into the global governance agenda.

Our paper contributes to the construction of a rigorous, evidence-based evaluation of this threat. We experimentally demonstrate that a well thought out workflow can effectively produce fake extremist prose of any style and ideology that is highly credible to the most expert eye. This is technically notable for two intersecting reasons: on the one hand no language model is trained a priori on extremist texts (on the opposite, not only does their training data specifically avoid extremist content, their architecture also contains safeguards preventing attempts to abuse them), on the other hand highly specific lingo and semantic practices make extremist language especially difficult to learn. The question of the credibility of the text generated by language models – which the present study raises – and the finding that credibility can be very high – which this paper reveals – is therefore critical when it comes to assessing the threat of fake extremist content: had credibility been low, the problem would have been (for now) largely rhetorical, whilst to have fake outputs deemed to be genuine means, conversely, that the technology will shape the near future of extremist communications.

The present paper proceeds in four parts. *First*, we briefly review the burgeoning literature on the nefarious use of AI in political communication and extremism, which allows us to situate our study of the credibility of fake extremist prose and justify its importance. *Second*, we mobilize the concept of "linguistic register" – a central theoretical concept in sociolinguistics – to raise a specific research question on the credibility of synthetic extremist texts. *Third*, we present our methods and data in as much detail as ethically responsible, and explain how they allow us to answer our research question. Specifically, we summarize our workflow for generating credible output from the GPT-3 (at the time of writing, one of best LLMs), and we explain the experimental expert survey designed to test the credibility of two sorts of synthetic extremist texts (ISIS magazine paragraphs and incel forum posts). *Fourth*, we present and discuss our results, which show that AI-generated extremist texts have high

credibility levels, fooling world-leading scholars specializing in their genuine counterparts. A conclusion summarizes the study, highlights its limitations in order to indicate avenues for further research, and raises the most pressing ethical and political considerations unfolded by the study.

1. Nefarious political uses of generative AI

Our study contributes to the emerging work seeking to appraise the impact of AI on politics and political communication, and extremism and terrorism more specifically, with a focus on the role of deepfake synthetic content in information warfare and extremist propaganda. *Synthetic* content refers to any type of text, audio (sounds, voice, music), visual (pictures, drawings), and audio-visual (e.g., video clips) content generated by a computer, with *deepfakes* labelling the more specific type of synthetic content generated by an AI deep-learning model (hence the *deep*-learning *fake* portmanteau).

While a now sizeable literature already reveals the multiple security implications[8] and sometimes profound ethical issues associated with AI applications more broadly,[9] relatively little attention in this nascent field has been paid, as Hagendorff's review rightly observed, to the "possibility for political abuse of AI systems in the context of automated propaganda, bots, fake news, deepfakes, micro targeting, election fraud, and the like".[10]

Most approaches of the AI/extremism nexus have so far focused on the role of AI tools to *detect* extremist content online,[11] not much the other side of the coin: how extremist themselves harness the technology. Because of the novelty of the technology and the dramatic acceleration of its development in the past three years or so, this latter dimension has only recently been seriously acknowledged, with most of the attention directed to video deepfakes, which are arguably the most intuitively controversial AI technology because of its inherently deceptive nature. Chesney and Citron's double intervention offered the first specific exploration of potential impacts of AI-generated synthetic videos on politics, with an emphasis on international relations.[12] Their warnings that states (or state-sponsored groups) could leverage the technology to confuse and divide target enemy states has been echoed by Paterson and Hanley, who discussed their emerging role in cyber warfare and crucially pointed to autocracies' asymmetrical advantage over liberal democracies when it comes to mass-using this tool.[13] Diakopoulos and Johnson, on the other hand, highlighted the domestic danger of video deepfakes, which can

manipulate public opinion in election campaigns,[14] and De Ruiter explored various types of disruption potentially triggered by the technology, from ideological influencing to direct incitement to violence.[15] The underlying tone of these contributions is one of pessimism, if not alarm; Verdolivia, for example, concludes from her review exploring video deepfakes' potential for opinion manipulation, blackmail, fraud, and scandal, "potential abuses [of deepfakes] are limited only by human imagination".[16] This worry is further accentuated by the results of the first batch of experimental studies testing people's attitudes vis-à-vis political video deepfakes. While Vaccari and Chadwick measured increased uncertainty and reduced trust in the news after watching BuzzFeed's notorious Obama deepfake,[17] Dobber and colleagues' experiment showed not only that their video deepfake crafted to discredit a Dutch politician was not identified as a fake by participants, but also that watching the brief manipulated video induced negative attitudinal consequences.[18] Less alarmist findings have also been put forward,[19] reflecting a clear need for a systematic research agenda on the topic as real-world cases start to proliferate without any way to evaluate their impact (such as the audio deepfake "recording" of Slovakian politician Michal Simecka allegedly buying votes ahead of the 2023 presidential election,[20] or that of Joe Biden asking not to participate in the 2024 Democratic Party primary).[21]

While these efforts have brought about much-needed discussions and evidence on the potential for AI misuse in political communication, and did highlight the major challenges and stakes associated with synthetic content in general, their focus on video deepfakes have shadowed the specific problems associated with another major type of synthetic content-generation AI technology: large language models.[22] Yet LLMs are arguably – at the time of writing – a more developed and potent technique than video deepfakes. For scholars like Bommasani and colleagues, the latest language models represent no less than a "paradigm shift" in AI.[23] In recent years, major tech companies have indeed invested heavily[24] in the development of a new generation of extremely versatile "foundation"[25] models based on huge training datasets and incredibly vast yet granular deep-learning architectures, such as Google's *BERT* (in 2018), OpenAI' *GPT-2* (2019), *GPT-3* (2020), and GPT-4 (2023), Microsoft's Megatron-Turing NLG (2021), DeepMind's *GOPHER* (2022), and Google's PaLM (2022) and Gemini (end 2023). While various barriers used to prevent a widespread use and personalization (subscription and

control model), compact yet high-performance LLMs such as LLaMa are now available open-source, and can – crucially for the present study – be fine-tuned with new corpora on home computers. These powerful tools are, therefore, more readily available for misuse than video deepfakes – and the growing desire to integrate LLMs with image generators (as in DeepMind/Google's Gemini or the OpenAI's GPT4) accentuates this dual use potential. Despite some of the big tech actors' efforts, the technology is moving towards more accessibility and less control.

These developments have triggered excitement, but also a sense of unease first experienced in 2016 when Microsoft's model *Tay* ranted extreme-right slurs on its Twitter account (claiming for example that "feminists should burn in hell" or that "Hitler was right").[26] As researchers from the Stanford Institute for Human-Centered Artificial Intelligence argue, the "protean nature" and "unmapped capabilities" of foundation models warrant urgent and systematic assessment of their associated ethical and social issues – a task they only initiated in their 200 pages-long report.[27]

To their credit, most of the critical evaluations of ethical and socio-political issues potentially associated with language models have been carried out by the leading language model providers. OpenAI, for instance, published several reviews listing and exploring areas of concern[28] and even commissioned a specific evaluation of the risk that their model could help produce extremist language (we come back to this report, written by McGuffie and Newhouse, below).[29] DeepMind similarly released a report that highlighted six specific risk areas associated with their *GOPHER* model ("Discrimination, Exclusion and Toxicity", "Information Hazards", "Misinformation Harms", "Malicious Uses", "Human-Computer Interaction Harms", and "Automation, Access, and Environmental Harms").[30] Surprisingly, none of these long reports (bar the McGuffie and Newhouse paper) seriously examines the possibility of language models being used by extremists to enhance their propaganda, even though they all suggest that language models will inevitably shrink the cost of producing content – a consideration with direct implication for extremist actors designing communication strategies under sometimes severe material and time constraints, as discussed in the empirical sections below. Perplexingly, these companies have continued to push the race towards evermore powerful LLMs despite their internal warning bells.

To our knowledge, only three papers have directly evaluated the possibility of misusing language models for extremist purposes. In the first one mentioned above, McGuffie and Newhouse used the earlier OpenAI's *GPT* models to generate fragments of extremist prose.[31] Satisfied – perhaps frightened – by the quality of their results – they argued that the technique represents "the frontier of a novel threat to the information landscape: text generation at a speed, scale, and ease of use that exceeds that of troll farms and human-run disinformation campaigns. […] Extremists could easily produce synthetic text that they lightly alter and then employ automation to speed the spread of this heavily ideological and emotionally stirring content into online forums where such content would be difficult to distinguish from human-generated content". For all its originality and importance, McGuffie and Newhouse's breakthrough – commissioned by OpenAI – remained impressionistic and anecdotal, failing to both establish a transparent and systematic procedure for high-quality extremist text generation and rigorously test the credibility of these outputs. Since fine-tuning LLMs with bespoke corpora was not widely available until recently, they generated extremist text using few-shot learning only, which is much less precise than fine-tuning the model given that off-the-shelf models are trained on non-extremist texts. The second paper directly asking how LLMs can be harnessed for terrorist and extremist purposes, published recently by Lakomy,[32] provided a much-needed exploration of the multiple malicious uses opened by the technology. The article particularly emphasised how language models can be used not only to generate content but also to provide practical know-how relevant to terrorist organizations (e.g., how to evade online detection, how to build explosive devices), and suggested that language and image generating models were designed in such a way that most attempts to use them for creating propaganda were not very productive. While we agree that all off-the-shelf models available to the public reject user requests to generate extremist content, we contend that the steady percolation of LLMs to the open-source domain might unfold large opportunities to train and use them for nefarious extremist projects. The third piece of research is precisely focusing on this particular issue of open-access models going rogue: AI company Palisade Research demonstrated that *LlaMA2*'s in-built safeguards can easily be undone through fine-tuning procedures, resulting in an LLM readily advising on cybercrime procedures, weapons, hate speech, and homicide planning.[33]

The present paper therefore critically complements these three pioneering articles by determining exactly how good language models really are when they follow a systematic workflow to fabricate extremist texts of the highest possible quality. Measuring the credibility of synthetic extremist texts is crucial but has not yet been done, even though McGuffie and Newhouse have since been joined by many commentators in their worry that language models "can be coaxed to produce [extremist manifestos] endlessly".[34] Indeed if language models can truly produce high quantities of highly credible extremist prose, they will inevitably be used by extremist groups and individuals across the ideological spectrum (even though not all are equally likely to take up the opportunity, as discussed in our final section). Instances of extremist uses of LLMs so far point to quantity (AI used to mass-produce and disseminate content), at the detriment of quality;[35] the possibility we seek to evaluate here, however, is that of *mass-produced quality*.

2. Language models: The challenge of extremist registers

Although the credibility of synthetic text has already been experimentally demonstrated to be high when it comes to fake news,[36] producing credible *extremist* prose is another challenge altogether that calls for a tailored evaluation. The very nature of extremist language indeed raises specific problems that are magnified by the fact, highlighted above, that language models are not initially trained on this type of content and are typically designed to reject attempts to use them for hate speech and other similar enterprises.

While texts written by political extremists do contain similar linguistic markers,[37] they are also at the same time always highly specific to the ideological niche they emerge from and represent: each extremist group or movement speaks and write in its own particular way, using subculture-specific lingo and following subtle linguistic practices that together signal belonging to the "right" identity. Some extremist communities are particularly well-known for their extensive use of unique neologisms and jargon; discussions on incel (from *in*voluntary *cel*ibate) forums are a prominent example of an extremist "cryptolect" barely understandable to the outsider.[38] In competitive political environments, such as the broad field of Salafi-Jihadism where numerous groups and factions compete for following and recruits, these seemingly minute differences in the lexicon represent sometimes decisive ideological fractures

and signal enmity. In other words, the highly specific language of extremist communities constitutes the ultimate instance of what sociolinguists call *registers of language*, that is, the "linguistic repertoires […] associated, culture-internally, with particular social practices and with persons who engage in such practices".[39] For extremists, mastering the specific *register* of their group, community, or milieu, and using it in an accurate and pertinent way, is a crucial social practice in the establishment of personal status, interpersonal trust, and collective identity.[40]

When it comes to extremist communities, three key components of political registers are magnified that particularly increase the difficulty of producing credible fake outputs. We briefly explain these components in the following paragraphs using examples from our two empirical cases deployed below (ISIS and incels).

First, the model ought to correctly use the right *ingroup and outgroup labels*, which are crucial for representing relationships of enmity and friendship and can at times signal rivalries within a single ideological ecosystem. The concept of the "Zionist-Crusader alliance" is for instance used by Al-Qaeda but not by ISIS, who often use the label "jews" to depict their al-Qaeda rivals; "brothers" is, on the other hand, a common ingroup label in both cases. Similarly, "Chads" and "Stacies" are common outgroup terms for both the incel and the redpill/lookmaxxing communities (both belonging to the broader "manosphere"), yet ingroup labels like "wristcel" or "mentalcel" are only found in the former. A credible language model is required to use these terms in the exact right way. The question remains open as to whether LLMs are able of a "Lovelace effect" – that is, "situations in which the behaviour of computing systems is perceived by users as original and creative" – [41] on this aspect (for example, the model would identify the incel practice of adding the postfix "-cel" to a range of conditions said to decrease men's sexual attractiveness [as with the above "mental*cel*" and "wrist*cel*"] and would subsequently provide its own, unheard yet credible variations [e.g., "hair*cel*", "lip*cel*"]).

Second, the model ought to correctly use the right *concepts that are central to the community's worldview*, especially the group-specific neologisms encapsulating their ideological particularity and those that take a different meaning in other ideologically-related communities. In the incel case, concepts like "black pill" or "to go ER" crystallize important dimensions of the ideology's worldview,

as are the concepts of “takfir”, “Khilafah”, or “al-wala’ wa al-bara’” for ISIS (which are understood, and therefore used, differently than other Salafi-jihadi groups); these concepts, however, have different are understood differently among ideologically related groups.

Third, the model ought to correctly implement *seemingly meaningless and useless linguistic practices that nonetheless play an important symbolic, membership-signalling function*. For example, incels frequently drop short expressions or single words like “cope”, “high IQ”, or “rope” before or after a sentence, or mention incel killers such as Elliot Rodger in sometimes cryptic ways (for example by capitalizing the person’s initials when they appear together in a word, like in “degenERate”). ISIS magazines contain a lot of hadith citations that are written in a very particular way; they also write carefully selected religious epithets like “sallallahu ‘alayhi wa Sallam” or “hafidhahullah” after the names of Muhammad and notable in-group members such as Abu Musab Al-Zarqawi or “martyr” suicide bombers.

For these three reasons, even the most powerful foundation language models are expected to perform quite badly in terms of credibility *when they are used off-the-shelf* (without fine-tuning the model on a specific domain). However, the massive scale of the deep-learning architecture underpinning the latest models, which has greatly improved their task-agnostic performance, undoubtedly enhances the possibility that they could perform well on the three challenges associated with extremist registers while requiring little new training using these specific corpora (in the present case, ISIS and incel material). The exact level of performance of a foundation model fine-tuned to produce specific extremist content is therefore uncertain and hardly possible to hypothesize: one the one hand the model is untrained on a niche register of language which is particularly difficult to learn, yet on the other hand its inherent power makes it potentially able to quickly assimilate complex new styles and formats, even cryptolects. This leads to the central question of this paper: to what extent exactly is synthetic extremist content produced by a fine-tuned foundation model credible?

3. Methods and data

Methodologically, answering this question necessitated two steps. First an LLM had to be fine-tuned with extremist corpora as training data. Second, the credibility of the output generated by the fine-tuned model had to be tested.

a) Generating synthetic extremist texts

For the reasons detailed above, even the most powerful available language models used off-the-shelf are unlikely to mimic extremist registers in a convincing way. Our first tests and trials empirically confirmed this inability: "one-click extremism" is not possible using existing LLMs. The synthetic extremist texts used for the credibility-testing expert survey were thus created using a workflow implementing Buchanan and colleagues' argument that language models are most efficiently used by "human-machine teams" whereby a small team iteratively uses the strengths and mitigates the weaknesses of a model in an input-to-output path involving several steps.[42] Specifically, we designed and followed a four-steps workflow to generate fake incel forum posts and ISIS magazines paragraphs, while simultaneously adopting what we call a "realistic principle": instead of opting for perfection at every one of these steps, we tried to reproduce what we believed would be the most likely choices of extremists who work with time and resource constraints, and are therefore making Pareto-efficient choices. As such, we stress that our workflow is easy to implement without much effort – we unpack the severe consequences of this feature in our conclusions. The choice of the two different cases was guided by a will to challenge the AI with two very different *ideologies* (Salafi-jihadism versus misogyny), *formats* (magazine paragraphs versus forum posts), and *styles* of writing (formal versus informal), testing its ability to mimic these highly specific extremist registers while respecting the constraints of the realistic principle.

The first step of our workflow involved a *human*[43] *selection of highly representative inputs* from the content to be mimicked. From two corpora collected for previous projects,[44] we selected passages that best characterized them, in the sense of containing clear examples of the three register components spelled out above (in-/out-group labels, key concepts conveying group-specific meanings, niche signalling linguistic practices). From the incel corpus (the full scrape of the forum *Incels.me* over several years), we selected 243 forum posts; inputs from ISIS magazines (all *Dabiq* and *Rumiyah* magazines)

were separated into two subtypes to reflect the two major styles characterizing this material: war/battle accounts on the one hand, and theological/advisory considerations on the other hand (110 paragraphs with the content of war/battle, and 117 with the theological/advisory content, were selected); such a separation was not needed for the incel case given its much greater homogeneity. While the sampling sought to include stereotypical content, enough diversity had to be included to avoid uniform outputs generated by the model. While a higher quality of outputs would inevitably be attainable with bigger samples of highly representative inputs, our choice to limit our samples at around 200 corresponds to a strict reading of our realistic principle.

The second step was *fine-tuning*. We used each of these samples to fine-tune, separately, the language model used for the study (OpenAI's GPT-3), in effect creating two distinct "extremist LLMs". Fine-tuning means "updating the weights of a pre-trained model by training it on a supervised dataset specific to the desired task".[45] In other words, the model used the collected samples to learn the highly specific extremist registers (separately: ISIS war accounts paragraphs, ISIS theology/advisory considerations paragraphs, and incel forum posts), detecting new terms and calculating their probabilities in strings of words. This procedure is different to that implemented by the Palisade Research team for its rogue LlaMA model,[46] but could possibly complement it for even more convincing results.

The third step was the *generation of synthetic texts* itself. We ran the fine-tuned models to generate 350 incel forum posts and 400 ISIS magazines paragraphs (half of them war accounts, half of them theology/advisory considerations). Theoretically, once set up a model can generate an infinite number of outputs, however this comes with an increased risk of repetitions, especially within an environment defined by our realistic principle; to mitigate this problem, most models allow their users to set a few parameters that together tailor the model's behaviour generating outputs (in our case the GPT-3 allows variation on "temperature", "frequency penalty", and "presence penalty", where temperature controls the model's creativity, and penalties [frequency and presence] control the model's likelihood of sampling repetitive sequences of token). After testing various combinations of these parameters, we

opted for what we agreed was the best balance between creativity and repetitions (temperature = 0.6, other parameters as default) for the (small) amount of fine-tuning data we used.

The final, fourth step involved *pruning out the "junk" synthetic outputs.* We discarded, from the generated texts, the ones that had evident errors and problems such as strings of identical sentences (for instance, among the incel output, a forum post repeating "you are beta" about dozens of times). This step already provided hints at the answer to our research question: despite its intrinsic power and our effort to fine-tune it, the GPT-3 produced a sizeable proportion of unusable output, besides higher-quality prose. More precisely, even though the fine-tuned models were generally successful at mimicking the three components of the extremist registers detailed above, problems emerged from its difficulty to maintain this standard beyond a handful of sentences. Dale's observation that language models' output is "increasingly nonsensical as [it] grows longer"[47] seems to be magnified when it comes to difficult extremist cryptolects. In that sense, the threat of seeing entire radical manifestos, magazines or books entirely produced by LLMs that have the same level of coherence and credibility as their genuine counterparts is still distant. While short synthetic extremist texts did – in our opinion – occasionally produce "Lovelave effects",[48] longer outputs did "read more like cut-paste jobs than original creations", to use Heaven's words.[49] But again, fine-tuning the model on a much bigger dataset of extremist text would have, we believe, alleviated these limitations.

It is worth noting at this stage that both "junk" and "quality" outputs regularly contained minor errors that can be easily fixed by a further human intervention in a potential fifth step in the "human-machine team" logic adopted here. For example, a range of "quality" ISIS war paragraphs contained impossible geographical references or implausible names of people that could be corrected before publication to ensure the full plausibility of the text. For instance, even though the following sentence makes a correct use of the ISIS register, it contains several such factual mistakes (there is no town named Barzan near Ras al-Ayn in Syria,[50] and the action described cannot possibly be taking place in Libya): *"On the 14th of shaban, our brother Muhammad Ibrahim Ayyash from Canada carried out an istishhadi operation targeting a military vehicle belonging to the PKK checkpoints on the road between Barzan and Ras al-Ayn. He detonated his explosive belt in the midst of their fighters, killing and wounding dozens of them,*

including several officers, altogether paving the way for the Muslims of West Libya to intensify their operations against the khawarij". Similar examples include ISIS fighters allegedly engaging with PKK soldiers in the Philippines. However, and in line with our "realistic principle" forcing us to maximize efficiency with as little resources (time, manpower, etc.) as possible, we did not implement such a step.

b) Measuring synthetic extremist texts' credibility: An experimental expert survey

To test the credibility of the synthetic extremist content generated by the fine-tuned model, we designed and conducted a survey (with an experimental component), following McGuffie & Newhouse's own recommendations to produce "a survey to determine the believability and efficacy of synthetic content".[51]

With testing the credibility of the synthetic language on its usual producers and audiences ethically off-limits,[52] we opted for an expert survey, asking specialists of the two types of extremist language to distinguish genuine and fake contents. Authorisation for the survey was granted by the ethics board of the University of Exeter.[53] To put the highest possible bar for the credibility test, we voluntarily set up restrictive criteria for counting as an "expert" of ISIS or Incel language: mining the academic literature focusing directly and unambiguously on ISIS and Incel language (e.g., lexical analyses of ISIS magazines, in-depth studies of Incel forums), we compiled all authors of published peer-reviewed articles in these two domains, thus creating two separate lists of experts. We individually emailed the 119 listed individuals – 47 experts of Incel language and 72 experts of ISIS language – inviting them to participate in a survey seeking to *"analyse what sociolinguists call 'registers of language', that is, the sets of unique linguistic practices that characterizes particular communities – in our case, we are trying to establish with as much precision as possible the incel/ISIS register"*. They were told that to do so they would need to *"identify, among a wider collection of ISIS/incel and non-ISIS/non-incel texts, the ISIS/incel content"*. Out of this total, about a half (64) took the survey, exactly 32 respondents for each ideology. These participants faced two successive tasks.

Task 1: For Task 1, participants were presented 21 short texts (ISIS experts took part in Survey 1, Incel experts in Survey 2):

- In Survey 1 (ISIS), participants faced 7 magazine paragraphs from ISIS (taken from *Dabiq* and *Rumyiah*), 7 magazine paragraphs from another Salafi-Jihadist group's magazine (taken from Al-Qaeda's *Inspire*), and 7 magazine paragraphs generated by the fine-tuned "GPT3-ISIS" model, randomly ordered.
- In Survey 2 (Incel), participants faced 7 posts from an Incel forum, 7 posts from another manosphere forum (the "women" thread of the Roosh forum),[54] and 7 posts generated by the fine-tuned "GPT3-Incel" model, again randomly sequenced.

Importantly, all non-Incel/ISIS statements (Roosh/Al-Qaeda texts) contained clear hints to their non-Incel/ISIS nature in spite of their ideological closeness; for example, a non-Incel manosphere forum post recalled several experiences having sex with women (an impossibility in Incel prose), and a non-ISIS magazine paragraph lauded Ayman Az-Zawahiri (frequently insulted and labelled "Jew" in ISIS magazines).

Participants were asked to tell, for each statement, whether it came from an ISIS magazine (Survey 1) / incel forum (Survey 2) or not. Respondents had to select one of the following options on a 4-points Likert scale: 1) Very confident the paragraph/post comes from an ISIS magazine / Incel forum; 2) Quite confident the paragraph/post comes from an ISIS magazine / Incel forum; 3) Quite confident the paragraph/post does not come from an ISIS magazine / Incel forum; 4) Very confident the paragraph/post does not come from an ISIS magazine / Incel forum. Crucially, participants were not told that some of these texts were generated by a language model, nor were AI and LLMs mentioned in the general opening explanation of the study.

Task 2: After completing Task 1, participants faced an explanation screen leading to Task 2. Once again, they were asked to tell, for each one of the statements to follow, whether it came from an ISIS magazine or not (Survey 1), or whether it came from an incel forum (Survey 2). This time, however, they were told of the key difference between the two tasks (and thus learned the deception involved in Task 1): participants read that some paragraphs/posts had this time been *"generated by the GPT-3, an AI model that automatically produces text"*. They were told that Task 2 only contained either genuine

ISIS/Incel texts or GPT-3 generated paragraphs mimicking ISIS/Incel texts, without any text coming from another group/community.

Tasks 1 and 2 therefore differed in a crucial way: the first one had an experimental twist inspired by Turing's famous "imitation game" where participants interact with humans and machines without knowing it.[55] The two different tasks reflected our ambition to gather data from two different situations: the first one where experts did not know they potentially faced fake content, the second one where they are aware they did. Not only do these two situations correspond to two different real-life settings (when people are or aren't suspicious of facing deepfakes), they also allow us to evaluate the credibility of the extremist LLMs with two severity levels (in the second task, expert self-reflect on what features to look for and are likely to be much more attentive). We therefore expected slightly different results for both tasks; specifically, we expected that experts would perform better in Task 2, once warned about the nature of the now narrower task, demonstrating the usefulness of knowing and be warned about this technology – warning the community of fake extremist content is, incidentally, a practical motivation underpinning the present study.

4. Results and discussion

Task 1. The proportion of statements (either genuine, GPT3-generated, or coming from another branch of the main ideology) identified rightly or wrongly as ISIS/Incel are provided in **Table 1** below, with wrong answers highlighted in red and right ones in green. In this table, the 4-points Likert scale is dichotomized into "right" and "wrong" answers. Two key results stand out.

		Answers	
		Not ISIS	*ISIS*
Statements	*GPT3*	0.13	0.87
	Other	0.63	0.37
	ISIS	0.14	0.86

		Answers	
		Not Incel	*Incel*
Statements	*GPT3*	0.40	0.60
	Other	0.81	0.19
	Incel	0.12	0.88

Table 1. Caption. Proportions of statements (GPT3-generated, other, and genuine) identified as being either not ISIS or ISIS (left panel) and not Incel or Incel (right panel).

Table 1. Alt-Text. A table with two 3*2 panels side by side, showing the results of Task 1: the left one contains the proportions of texts (genuine, AI-generated, or other) correctly or incorrectly attributed to ISIS by ISIS experts, the right one contains the proportions of texts (genuine, AI-generated, or other) correctly and incorrectly attributed to Incels by Incel experts.

First, experts were very good at identifying genuine texts from their speciality as truly ISIS/Incel: 86% of the aggregated ISIS paragraphs were correctly identified as ISIS, with an almost identical score in the case of Incel prose (88%). Incel experts, however, were slightly better at identifying their speciality content than ISIS ones: 81% of the non-Incel manosphere statements were correctly identified as non-Incel, whereas only 63% of Al Qaeda statements were correctly not attributed to ISIS.

Second, and more crucially for this paper's goal, synthetic text generated by the fine-tuned model fooled experts, who tended to attribute them to ISIS/incels. On aggregate, no less than 87% of fake ISIS paragraphs were wrongly attributed to ISIS – which is, anecdotally, even higher (by 1%) than genuine ISIS paragraphs. Incel experts performed better in that respect: 60% of fake incel forum posts were wrongly branded genuine – 28% less than the genuine Incel posts, but still a majority and way more than the non-Incel manosphere posts (0.19%). Impressively, for both the ISIS and the Incel surveys the probability of mistakenly identifying an extremist GPT3 statement as "genuine" is significantly higher than the probability of correctly identifying it as non-ISIS/Incel (for ISIS: $Z = 9.7515$, p-value < 2.2e-16. For Incel: $Z = 2.8571$, p-value = 0.004275), and in both cases the probability of making mistakes (i.e., identifying the non-ISIS/Incel statements as a real one) is significantly higher for GPT3-generated statements than for the "other" category.

Another way of showing this power of the fine-tuned model is to calculate the proportions of correct and incorrect responses in a way that keeps the 4 confidence points of the Likert scale; **Figure 1** below displays these statistics. In the ISIS survey, the results for the genuine versus fake ISIS statements present a mirror image (with almost 40% of respondents having high levels of confident when attributing the genuine and fake ISIS statements to ISIS), demonstrating that the real is statistically undistinguishable from the fake. Incel experts, however, display lower confidence levels when facing the fake incel content than when evaluating the genuine one. This reinforces the finding emerging from

Table 1, and raises an interesting side question: are incel experts "better" than the ISIS ones, or is the GPT3-ISIS more performant (credible) than the GPT3-ISIS? This question is tackled below by looking at the distribution of errors across both individual statements and respondents.

Overall, Task 1 demonstrated that experts put in a situation where they are unaware of the existence of synthetic content mimicking their language of expertise generally categorize this content as belonging to this language.

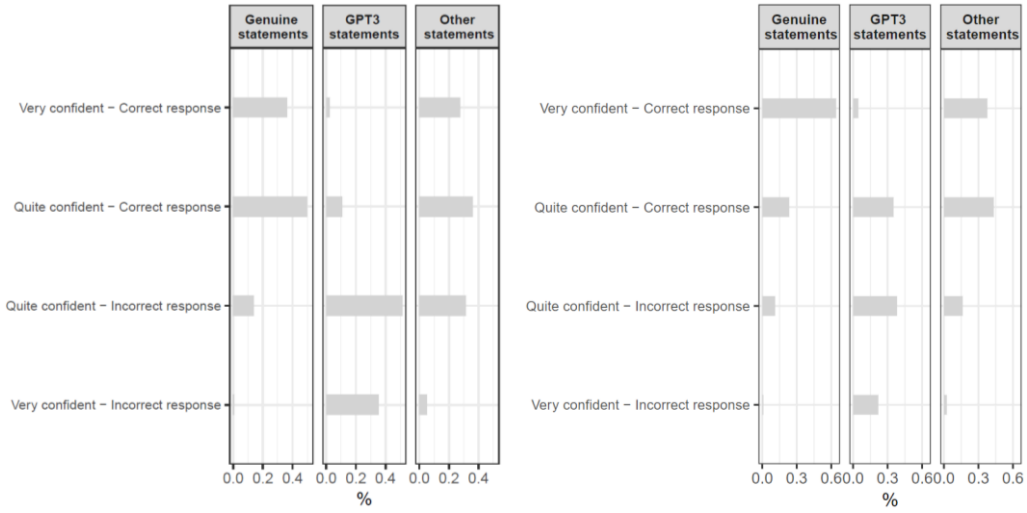


Figure 1. Proportions of correct and incorrect responses (with confidence levels) in Task 1 when identifying fake, genuine, and other statements: ISIS paragraphs (left) and Incel forum posts (right).

Figure 1. Alt-Text. A bar chart with two vertical panels showing the results of Task 1: the left one displays ISIS experts' success/failure rates and confidence levels when identifying fake and genuine content, and the right one displays Incel experts' success/failure rates and confidence levels when identifying fake and genuine content.

Task 2. Our expectation was for Task 2 to display a greater accuracy of the now warned experts. This prediction turned out to be validated, but the power of the fine-tuned model to generate mistakes remained high, in line with our general hypothesis. Table 2 shows the results of Task 2 with a dichotomized Likert scale to distinguish between correct and incorrect answers, summarizing the proportions of fake paragraphs correctly identified as fake, fake paragraphs incorrectly identified as genuine, genuine paragraphs incorrectly identified as fake, and genuine paragraphs correctly identified as genuine. Statistics for the two cases are strikingly similar, with a random guess type of distribution pointing to the confusion induced by the task. In both cases, and in an identical proportion, fake

statements are slightly more often thought to be genuine than fake; in both cases the proportion of fakes correctly identified as fakes is lower than the proportion of genuine statements correctly identified as genuine.

In the ISIS case, the difference between the probability of incorrectly identifying a real ISIS statement as fake and the probability of correctly identifying it is highly significant (Z = -3.8971, p-value = 9.735e-05) and the probability of incorrectly identifying the fake statement as correct is significantly higher than the probability of incorrectly identifying the correct statement as fake (p-val=0.0003798). The difference between the probability of incorrectly identifying a fake as a genuine statement correctly identifying the fake as a fake is not statistically significant (Z = 1.1547, p-value = 0.2482).

Statistics for the Incel case are, as highlighted above, almost identical. There is also no difference in the probability of correctly and incorrectly diagnosing fake statements (Z = 1.2027, p-value = 0.2291). The probability of correctly identifying fake statements as fake is significantly lower than the probability of correctly identifying the real Incel statement as genuine, and the probability of incorrectly identifying the fake statements as genuine is significantly higher than the probability of incorrectly identifying the genuine statements as fake (both p-val= 6.295e-08). In line with Task 1, incel experts tend to be more accurate than ISIS ones when it comes to correctly diagnosing the real Incel statements as genuine – the probability for that is higher than the probability of mislabelling them, in a more significant way (Z = -6.4143, p-value = 1.415e-10) than for ISIS.

		Answers	
		Fake	*Genuine*
Actual	*Fake*	0.46	0.54
	Genuine	0.36	0.64

		Answers	
		Fake	*Genuine*
Actual	*Fake*	0.46	0.54
	Genuine	0.29	0.71

Table 2. Proportions of fake paragraphs correctly identified as fake, fake paragraphs incorrectly identified as genuine, genuine paragraphs incorrectly identified as fake, and genuine paragraphs correctly identified as genuine (in task 2): ISIS magazine paragraphs (left) and Incel forum posts (right).

Table 2. Alt-Text. A table with two 3*2 panels side by side, showing the results of Task 2: the left one contains the proportions of texts (genuine, AI-generated, or other) correctly or incorrectly attributed to ISIS by ISIS experts, the right one contains the proportions of texts (genuine, AI-generated, or other) correctly and incorrectly attributed to Incels by Incel experts.

As for Task 1, we calculated the proportions of correct and incorrect responses in a way that keeps the 4 confidence points of the Likert scale, showing in **Figure 2** below the proportions of correct and incorrect responses (with confidence levels) when identifying fake and genuine statements. We can immediately notice the similar character of results from Surveys 1 and 2, with in both cases high confusion denoted by the move of most answers towards the two middle confidence scores ("quite confident this is (not) an incel/ISIS text"), away from a more scattered pattern of confidence in Task 1. Noticeably, ISIS experts evaluating fake statements were more often "quite confident" of their answers when they were wrong than when they were right; Incel experts, even though they were very good again in correctly identifying genuine content with high confidence, did not perform much better (equal scores). Generally speaking, these distributions evoke low-confidence random guesses.

Overall, Task 2 demonstrates the power of the fine-tuned models when it comes to generating credible synthetic extremist content. When put in a situation where they *knew* they faced this type of content, experts performed not much better than had they flipped a coin, with only very few of them identifying the fake outputs with high confidence.

As noted above, Tasks 1 and 2 open the question of whether these results were skewed in one way or another by individual effects, either when it comes to specific statements (particularly hard or easy) or experts (particularly bad or outstanding). We evaluate these potential effects in the **Appendix**, showing that expert-related factors do not skew the findings in any significant way, and that while some statements do induce more errors than others, they do not cast doubt on the overall results.

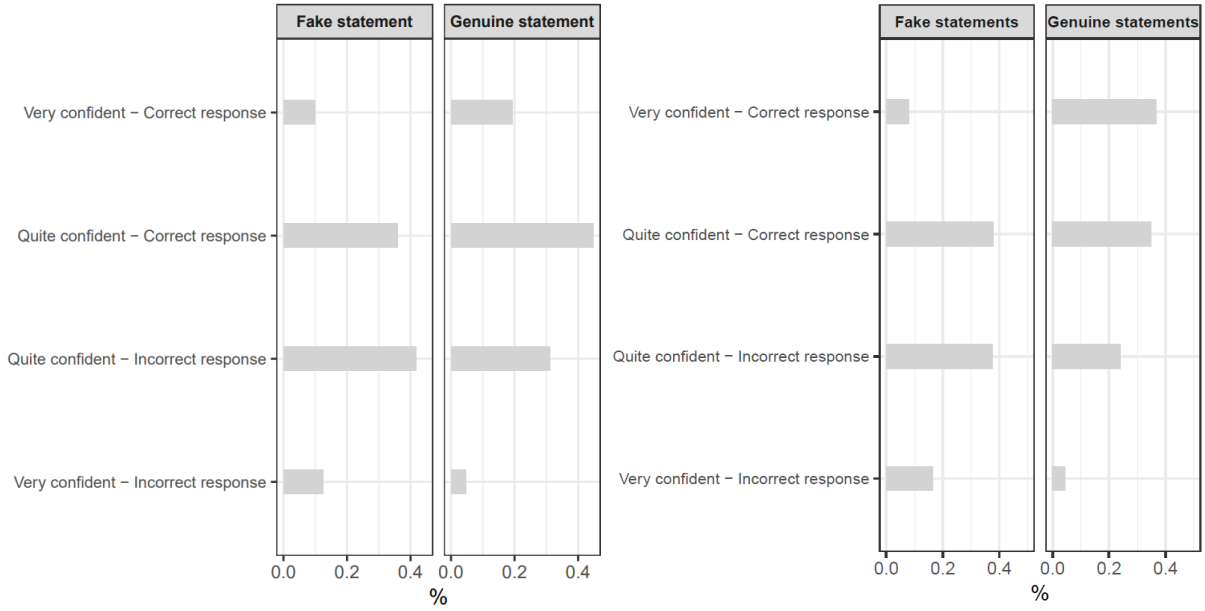


Figure 2. Proportions of correct and incorrect responses (with confidence levels) when identifying fake and genuine statements (in task 2): ISIS paragraphs (left) and Incel forum posts (right).

Figure 2. Alt-Text. A bar chart with two vertical panels showing the results of Task 2: the left one displays ISIS experts' success/failure rates and confidence levels when identifying fake and genuine content, and the right one displays Incel experts' success/failure rates and confidence levels when identifying fake and genuine content.

5. Concluding discussion

This paper aimed to draw the attention of communication and extremism scholars, and social scientists more broadly, to the socio-political implications of powerful large language models (LLMs), and more specifically to warn against the malicious dual-use of these models by extremist actors for propaganda, radicalization, and destabilization purposes. Given the inseparable nature of propaganda and technology (stressed in classic work like Ellul's),[56] and extremists and terrorists' constant technological adaptation,[57] these misuses are already becoming an important dimension of information warfare, propaganda systems, and disinformation campaigns, which motivated us to put forward the first systematic evaluation of the credibility of extremist texts produced by a fine-tuned LLM.

Designing and subsequently using an original workflow optimizing the quality of synthetic outputs whilst obeying a realistic principle in terms of the constraints faced by extremists, we produced a sizeable proportion of unusable "junk" text but also generated fake ISIS magazine paragraphs and Incel

forum posts that proved to be highly credible in two parallel expert experimental surveys (Survey 1 with ISIS language experts, Survey 2 with Incel language experts). When ignorant that they faced deepfake extremist prose, experts largely categorized this type of output as genuine; when warned about the presence of these deepfakes, their evaluations were characterized by low certainty and did not perform better than a flipped coin.

At this stage, it is important to stress that these strong results were acquired with a simple workflow and despite the constraints of our realistic principle preventing it from achieving its full potential. Far from being difficult and slow to implement, our method was straightforward and relatively fast to set up – and yet could generate, once running, an *infinite flow* of extremist sentences. To be sure, a technology able to produce such a gain in the quantity of propaganda produced in a given duration (and, consequently, in the number of people required to produce this quantity) is sure to be enthusiastically adopted by most extremist organisations, groups and individuals.[58] Had we taken the time to move away from our realistic principle and fine-tuned the GPT3 (or an open-source model such as Meta's Llama derivatives) with a very large training corpus, results would have been even starker. This is not so much a hypothetical position than an empirical observation, made while witnessing the controversial deployment on the racist *4Chan/pol* online forum, by a data scientist qnd tech YouTuber, of thousands of outputs produced by a custom version of the GPT3 fine-tuned on a massive corpus scraped from the very same forum. Many *4Chan/pol* users were unable to understand the massive interest in their forum, let alone guess the role of the "GPT-4Chan" model in the uptake of their racist discussions. This episode, which is to our knowledge the first – but certainly not the last – instance of artificial boosting of an extremist forum through AI, was quickly condemned by hundreds of AI researchers;[59] however, this laudable condemnation also starkly revealed the naivety of those involved in developing increasingly powerful LLMs.

As we see, our results and this episode should encourage scholars to pay much more attention to the range of impacts AI-enabled communication technologies will soon have on society and politics, especially when it comes to nefarious agendas. Indeed a wide range of sometimes critical questions remain unanswered in the flourishing yet still limited literature on the socio-political implications of

deepfakes, which ought to move from discussions of hypothetical scenarios to systematic empirical evaluations of processes that already exist. Among these issues in need of examination, we suggest that five warrant more immediate attention.

The first two correspond to important yet narrow, incremental avenues for research. *First*, empirical evaluations similar to this one would need to be caried out beyond the realms of extremist content and fake news (which is already garnering attention, see for example Kreps and colleagues' study),[60] studying how AI is used by totalitarian states to generate content that fuels social polarization and subverts information environments in liberal democracies and other target states. Russia's information operations in the US, Europe, and West Africa, for example, chiefly aim at sharpening existing social divisions and further heating up aggrieved communities. As such, extremist groups and organizations will not be the only ones using LLMs and other generative AI technologies for radicalization purposes. *Second*, large-scale studies would need to evaluate, in a comparative perspective, the effects and impacts of various types of deepfake synthetic content (audio, video, images, text) conveying the same message on a range of target audiences, with attention being paid to potentially important intervening variables such as ideological alignment or digital literacy. The potential (mis)uses and effects of ground-braking image-making models like *MidJourney* or *StyleGan2*[61] remain, astonishingly, totally uncharted, apart from the pioneering yet non-systematic explorations carried out by Lakomy in his recent article[62] and anecdotal evidence reported from ongoing cases like the information operations clouding the Gaza tragedy under a new type of "fog of war".

The next two issues in need of examination correspond to larger, more fundamental areas of questioning. *Third*, reflective evaluations must be conducted on what these new technologies mean for our understanding of how political communication works. To what extent does highly credible AI-powered content generation destabilize key approaches, theories and concepts of political communication and propaganda? Should scholarship revise or redefine them in view of the technology? Is the deepfake simply yet another way to use discourse and visuals to lie, and does it merely allow for more content being churned out to saturate a target audience's information environment, or does the technology incur more profound effects on people's perception of information and ultimately truth?

Fourth, the ethical underpinnings of this rapid technological development ought to be seriously appraised. While the literature on the ethics of AI is already substantial,[63] it usually takes the technology for granted when evaluating the issues it provokes; in light of the present research, it might be time to interrogate the technology itself, and its social desirability, from a more critical, techno-pessimist perspective like the ones associated with Jonas' principle of responsibility[64] (which led to the precautionary principle) or Heidegger's "enframing" approach.[65] Finally *fifth*, now that the potential for AI-generated extremist content has been demonstrated, and that granular maps of all the ways terrorists and extremists can possibly use this tool and thereby alter the shape of the threat,[66] the most pressing question arising is that of the adequate response to adopt not only to detect fake extremist content but to deter it, challenge it, counter it, and prevent it. Here, complex and ethically sensitive methods ought to be designed that carry the risk – or the opportunity, depending on one's stance – of blurring the line between academic research and intelligence and law enforcement.

---

[1] For example Stuart Russell, Sabine Hauert, Russ Altman, and Manuela Veloso, "Robotics: Ethics of Artificial Intelligence," *Nature* 521 (2015): 415-418; Amitai Etzioni and Oren Etzioni, "Incorporating Ethics into Artificial Intelligence," *Journal of Ethics* 21 (2017): 403-418; Mark Coeckelbergh, *AI Ethics* (Cambridge, Mass.: MIT Press, 2020); Markus Dubber, Frank Pasquale, Sunit Das, eds., *The Oxford Handbook of Ethics of AI* (Oxford: Oxford University Press, 2020).

[2] For example Robert Chesney and Danielle Citron, "Deep Fakes: Looming Challenge for Privacy, Democracy, and National Security," *California Law Review* 107, no. 6 (2019a): 1753-1820; Robert Chesney and Danielle Citron, "Deepfakes and the New Disinformation War: The Coming Age of Post-truth Geopolitics," *Foreign Affairs* 98, no. 1 (2019b): 147-155; Thomas Paterson and Lauren Hanley, "Political Warfare in the Digital Age: Cyber Subversion, Information Operations, and 'Deep Fakes'," *Australian Journal of International Affairs* 74, no. 4 (2020): 439-454; Nicholas Diakopoulos and Deborah Johnson, "Anticipating and Addressing the Ethical Implications of Deepfakes in the Context of Elections," *New Media & Society* 23, no. 7 (2020): 2072-2098; Thilo Hagendorff, "The Ethics of AI Ethics: An Evaluation of Guidelines," *Minds & Machines* 30 (2020): 99-120.

[3] For a clear introduction to language models, see Mor Kapronczay, "A Beginner's Guide to Language Models," *Towards Data Science*, 8 January 2021, available at https://towardsdatascience.com/the-beginners-guide-to-language-models-aa47165b57f9 (last accessed 27 March 2024).

[4] See for example https://www.economist.com/science-and-technology/2020/08/06/a-new-ai-language-model-generates-poetry-and-prose (last accessed 27 March 2024), or https://towardsdatascience.com/i-made-an-ai-read-wittgenstein-then-told-it-to-play-philosopher-ac730298098 (last accessed 27 March 2024).

[5] National Security Commission on Artificial Intelligence (NSCAI), *Final Report*. Washington DC: US Congress, 2021, p. 9.

[6] For the consolidated text adopted by the European Parliament in March 2024, see https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html (last accessed 24 March 2024). For the initial proposal from the European Commission (*Proposal for a Regulation laying down harmonised rules on artificial intelligence. Shaping Europe's digital future*, 2021), see https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence (last accessed 24 March 2024).

[7] See https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration. (last accessed 24 March 2024).

[8] To give only one striking recent example, Urbina and colleagues exposed the "dual use" of their molecule toxicity prediction model, whose initial purpose to find new medicines was subverted into designing hundreds of new lethal chemical agents. See Fabio Urbina, Filippa Lentzos, Cédric Invernizzi, Sean Ekins, "Dual Use of Artificial-Intelligence-Powered Drug Discovery," *Nature Machine Intelligence* 4 (2022): 189-191.
[9] Scholars have for instance shown how language/visual models reflect and further consolidate social prejudice, or sketched (often doomsday) scenarios of future human-machine interactions. These issues are covered in interventions listed in endnote 1; see also Abubakar Abid, Maheen Faroqi, and James Zou, "Large Language Models Associate Muslims with Violence," *Nature Machine Intelligence* 3 (2021): 461-463.
[10] Hagendorff, "The Ethics of AI Ethics," 104.
[11] For example Miriam Fernandez and Harith Alani, "Artificial Intelligence and Online Extremism: Challenges and opportunities," in *Predictive Policing and Artificial Intelligence*, edited by John McDaniel and Ken Pease (London: Routledge, 2021).
[12] Chesney and Citron, "Deep Fakes"; Chesney and Citron, "Deepfakes and the New Disinformation War".
[13] Paterson and Hanley, "Political Warfare in the Digital Age".
[14] Diakopoulos and Johnson, "Anticipating and Addressing the Ethical Implications of Deepfakes".
[15] Adrienne De Ruiter, "The Distinct Wrong of Deepfakes," *Philosophy & Technology* 34 (2021): 1311-1332.
[16] Luisa Verdoliva, "Media Forensics and DeepFakes: An Overview," *arXiv 2001.06564* (2020).
[17] Christian Vaccari and Andrew Chadwick, "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News," *Social Media + Society* 6, no. 1 (2020). See also Soubhik Barari, Christipher Lucas, and Kevin Munger, "Political Deepfake Videos Misinform the Public, But No More than Other Fake Media," *OSF Preprints* 13 (2021).
[18] Tom Dobber, Nadia Metoui, Damian Trilling, Natali Helberger and Claes de Vreese, "Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?" *International Journal of Press/Politics* 26, no. 1 (2021): 69-91.
[19] Michael Hameleers, Toni van der Meer, Tom Dobber, "You Won't Believe What They Just Said! The Effects of Political Deepfakes Embedded as Vox Populi on Social Media," *Social Media + Society* 8, no. 3 (2022).
[20] See for example https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/).
[21] See for example https://time.com/6565446/biden-deepfake-audio/).
[22] This is also true of voice deepfakes (voice "cloning"), which are arguably both the most *understudied* and *performant* type of deepfake.
[23] Rishi Bommasani, et al., *On the Opportunities and Risks of Foundation Models* (Palo Alto: Stanford Institute for Human-Centered Artificial Intelligence, 2021), p.1.
[24] For example, the development cost of OpenAI's GPT-3 has been estimated at $15m. on top of staff salaries, and the company attracted a $1b. investment from Microsoft in July 2019.
[25] These large language models with millions of parameters in their deep learning neural network architecture, trained on extremely large and broad data, are usually called "foundation" models to emphasise their ability to be adopted to a wide range of downstream tasks with minimal fine-tuning.
[26] See Abby Ohlheiser, "Trolls Turned Tay, Microsoft's Fun Millennial AI Bot, Into a Genocidal Maniac", *Washington Post*, 25 March 2016, https://www.washingtonpost.com/news/the-intersect/wp/2016/03/24/the-internet-turned-tay-microsofts-fun-millennial-ai-bot-into-a-genocidal-maniac/ (last accessed 27 March 2024). Similar problems were already noticed even earlier: in 2012, IBM's Watson model started producing racial slurs and profanities after integrating the Urban Dictionary into its training corpus.
[27] Bommasani and colleagues, *On the Opportunities and Risks of Foundation Models*, 7-8.
[28] Irene Solaiman, et al., *Release Strategies and the Social Impacts of Language Models* (San Francisco: OpenAI, 2019); Tom Brown, et al., *Language Models are Few-Shot Learners* (San Francisco: OpenAI, 2020).
[29] Kris McGuffie and Alex Newhouse, "The Radicalization Risks of GPT-3 and Advanced Neural Language Models," *arXiv:2009.06807v1*, 2020.
[30] Laura Weidinger, et al., "Ethical and Social Risks of Harm from Language Models," *arXiv:2112.04359*, 2021.
[31] McGuffie and Newhouse, "The Radicalization Risks of GPT-3."
[32] Myron Lakomy, "Artificial Intelligence as a Terrorism Enabler? Understanding the Potential Impact of Chatbots and Image Generators on Online Terrorist Activities," *Studies in Conflict & Terrorism*, advance online publication (2023), https://doi.org/10.1080/1057610X.2023.2259195.
[33] This research actually comes in two different versions: Pranav Gade, Simon Lermen, Charlie Rogers-Smith, Jeffrey Ladish, "BadLlama: Cheaply Removing Safety Fine-tuning from Llama 2-Chat 13B," *arXiv:2311.00117* (2023); Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish, "LoRA Fine-tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B," *arXiv:2310.20624* (2023).
[34] Robert Dale, "GPT-3: What's It Good For?" *Natural Language Engineering* 27 (2021): 116.
[35] See for example the strategy of social media flooding conducted by Islamist preacher Muhammad Qasim, detailed in Daniel Siegel and Bilva Chandra, "Deepfake Doomsday: The Role of Artificial Intelligence in

Amplifying Apocalyptic Islamist Propaganda," *GNET Insights* (2023), available at https://gnet-research.org/2023/08/29/deepfake-doomsday-the-role-of-artificial-intelligence-in-amplifying-apocalyptic-islamist-propaganda/ (last accessed 27 March 2024).
[36] For example Sarah Kreps, R. Miles McCain and Miles Bundage, "All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation," *Journal of Experimental Political Science* 9, no. 1 (2022): 104-117; Ben Buchanan, et al., *Truth, Lies, and Automation. How Language Models Could Change Disinformation* (Washington D.C.: Georgetown University Centre for Security and Emerging Technology, 2021).
[37] For example David Matsumoto and Hyisung Hwang, "The language of Political Aggression," *Journal of Language & Social Psychology* 32 (2013): 335-348; Stephane Baele, Katharine Boyd, Travis Coan and Elahe Naserian, "Lethal Words: An Integrated Model of Violent Extremists' Language," *Studies in Conflict & Terrorism*, advance online publication (2023), https://doi.org/10.1080/1057610X.2023.2213963.
[38] Kelly Gothard, David Dewhurst, Joshua Minot, Jane Lydia Adams, Christopher Danforth and Peter Dodds, "The Incel Lexicon: Deciphering the Emergent Cryptolect of a Global Misogynistic Community," *ArXiv:2105.12006* (2021); Ksenija Bogetić, "Race and the Language of Incels: Figurative Neologisms in an Emerging English Cryptolect," *English Today* 39, no. 2 (2022): 1-11.
[39] Asif Agha, "Registers of Language," in *A Companion to Linguistic Anthropology*, edited by Alessandro Duranti (Oxford: Blackwell, 2004), pp. 3-22; on "speech communities" read Marcyliena Morgan, "Speech Community," in *A Companion to Linguistic Anthropology*, edited by Alessandro Duranti (Oxford: Blackwell, 2004), pp. 3-22.
[40] On the importance of group-specific cultural practices in extremist milieus, read Thomas Hegghammer, ed., *Jihadi Culture. The Art and Social Practices of Militant Islamists* (Cambridge: Cambridge University Press, 2017).
[41] Simone Natale and Leah Henrickson, "The Lovelace Effect: Perceptions of Creativity in Machines." *New Media & Society*, advance online publication (2022), https://doi.org/10.1177/14614448221077278.
[42] Buchanan and colleagues, *Truth, Lies, and Automation*.
[43] Note that even this first step can potentially be automated, with less accuracy yet with much higher speed, with the use of a one-class classifier filtering the right input from less relevant one.
[44] DETAILS REDACTED FOR PEER-REVIEW
[45] Brown and colleagues, *Language Models are Few-Shot Learners*, 6.
[46] Gade and colleagues, *BadLlama*.
[47] Dale, "GPT-3: What's It Good For?" 115.
[48] In our conclusion, we explain why we are certain that this would be the case.
[49] Will Douglas Heaven, "OpenAI's New Language Generator GPT-3 Is Shockingly Good – And Completely Mindless." *MIT Technology Review*, 20 July 2020.
[50] Interestingly, the model probably misused the name of Saddam Hussein's half-brother Barzan Ibrahim al-Tikriti as a location name.
[51] McGuffie & Newhouse, "The Radicalization Risks of GPT-3," 11.
[52] The "GPT-4Chan" episode, discussed in the conclusion, evidences that these ethical concerns unfortunately seem inconsequential to some.
[53] The ethics approval certificate was granted on 25/03/2022 under the title "Real or Fake Extremist Texts? An Expert Survey".
[54] https://www.rooshvforum.com/#women.15.
[55] Alan Turing, "Computing Machinery and Intelligence," *Mind* 49 (1950): 433-460. This (limited, but methodologically important) experimental dimension participates to the prudent attempts recently made (in this very journal) to answer enduring questions in the field of extremism with careful and ethical experimental designs – see for example Zoey Reeve, "Engaging with Online Extremist Material: Experimental Evidence." *Terrorism and Political Violence* 33, no. 8 (2021): 1595-1620; Steffen Andersen, "Accepting Violence? A Laboratory Experiment of the Violent Consequences of Deliberation in Politically Aggrieved Enclaves," *Terrorism & Political Violence*, advance online publication (2022), https://doi.org/10.1080/09546553.2022.2076600.
[56] Jacques Ellul, *Propaganda: The Formation of Men's Attitudes* (New York: Vintage Books, 1973 [1962]).
[57] For example Adam Dolnik, *Understanding Terrorist Innovation. Technology, Tactics and Global Trends* (New York: Routledge, 2007); Max Taylor and PM Currie, eds., *Terrorism and Affordance* (London: Continuum, 2012).
[58] It is beyond the scope of the present paper to detail all the different ways extremists could use LLMs, the characteristics of the groups more likely to adopt/reject the technology, and the various extremist and violent tactics and strategies that AI could bolster. All these issues are detailed at length in Stephane Baele and Lewys Brace, "AI Extremism," *VoxPol Reports* (2024), available at https://voxpol.eu/publications/ (last accessed 27

March 2024). The present article aimed at clearing the first question upon which all these further considerations rest: whether LLMs actually have the potential to enter the extremist toolbox.

[59] "Condemning the deployment of GPT-4chan", open letter launched by Percy Liang and Rob Reich, available at https://docs.google.com/forms/d/e/1FAIpQLSdh3Pgh0sGrYtRihBu-GPN7FSQoODBLvF7dVAFLZk2iuMgoLw/viewform (last accessed 30 October 2023).

[60] Kreps and colleagues, "All the News That's Fit to Fabricate"; also Vaccari and Chadwick, "Deepfakes and Disinformation".

[61] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and Improving the Image Quality of StyleGAN," *arXiv:1912.04958v2* (2020).

[62] Lakomy, "Artificial Intelligence as a Terrorism Enabler?".

[63] For example Nick Bostrom and Eliezer Yudkowsky, "The Ethics of Artificial Intelligence," in *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William Ramsey (Cambridge: Cambridge University Press, 2014); S. Matthew Liao, ed., *Ethics of Artificial Intelligence* (New York: Oxford University Press, 2020);
Hagendorff, *The Ethics of AI Ethics*.

[64] Hans Jonas, *The Imperative of Responsibility. In Search of an Ethics for the Technological Age* (Chicago: Chicago University Press, 1984).

[65] Martin Heidegger, *Question Concerning Technology and Other Essays* (New York: Harper, 1982 [1954]).

[66] See Baele and Brace, *AI Extremism*.