***Science* Manuscript Template Instructions**

**General Instructions on using this template and submitting a manuscript to *Science*:** Thank you for preparing a manuscript for submission to *Science*. Using this template, or following the guidelines below, will help us in processing your paper. Our goal is to be able to identify each section of your manuscript so that we can accurately record the title, authors, abstract, etc. and to enrich it by including reference links and an accurate layout.

Please use the actual template, which starts on page 2. When you are ready to submit, please delete the text on this cover page.

You can submit your paper at <https://cts.sciencemag.org>.

Additional information for authors is available at [http://www.sciencemag.org/authors/science-information-authors](http://www.sciencemag.org/site/feature/contribinfo/index.xhtml).

If you are using LaTeX, please convert your paper into a Word .docx file if possible. If this is not possible, please use our LaTeX template and upload a PDF version of your paper. Some conversion approaches are available here: <http://www.tug.org/utilities/texconv/textopc.html>.

So that we can identify the parts of your paper, and even if you do not use our template, please begin each section with the specific key words listed below, some of which are followed by a colon. Please do not use paragraph breaks in the title, author list, or abstract. The author list, corresponding author email(s), and affiliation(s) should be checked carefully because they will be published as listed in the manuscript.

**Title:** No more than 96 characters, lacking jargon and abbreviations where possible.

**Authors:**

**Affiliations:**

**One Sentence Summary:** No more than 40 characters.

**Abstract:** 125 words or less.

**Main Text:**

**References and Notes:** (Followed by a numbered list); only a single reference list should be provided for the main text and supplemental information.

**Acknowledgments:** Split into sections as described below.

**List of Supplementary Materials:** Include a list, noting which references are only cited in the SM.

**Fig. #.** (Begin each figure caption with a label, “**Fig. 1.**”, for example, as a new paragraph.)

**Table #.** (Begin each table caption with a label “**Table 1.**”, for example, as a new paragraph.)

**Supplementary Materials:** Comprising Materials and Methods, figures, and tables; should be in a separate file.

Please use the .docx format (all versions after Word 2007). If you chose not to use this template, please include page numbers in your submitted file. We also encourage use of line numbers.

More specific formatting instructions are provided in the template which follows.

Title: Using Deepfakes to Hack the Unconscious Mind

**Authors:** Sean Hughes[[1]](#footnote-2)\*, Ohad Fried2, Melissa Ferguson3, David Yao4, Ciaran Hughes5, Rian Hughes6, & Ian Hussey1

Conventional wisdom tells us that “seeing is believing”. Yet thanks to recent advances in artificial intelligence this may no longer be the case. A branch of machine learning known as ‘deep learning’ has made it increasingly easy to take a person’s likeness (whether their face, voice, or writing style), feed that data to a computer algorithm, and have it generate a synthetic copy. The results are equal parts impressive and frightening: a digital doppelganger which can convince others that what they’re seeing, reading, or hearing is fact rather than fiction. Although mainly used to mimic real individuals, this technology can also be used to generate images of people who do not exist [ref], synthetic voices that belong to no-one [ref], and synthetic text that sounds human-authored [ref]. Content generated or manipulated in this way is collectively known as ‘synthetic media’.

Synthetic media is rapidly evolving: it’s becoming highly realistic, easier to produce, and thanks to the Internet, can be distributed and shared on a mass scale. One recent report suggests that the number of ‘Deepfakes’ (a subcategory of synthetic media) is doubling online every six months [ref]. What once took a small fortune and a Hollywood special-effects department can now be achieved using only a computer or smartphone.

The technology behind synthetic media can be deployed for both good or ill. Some are using it to generate believable voices and images for those who have lost their own through traumatic injury or cancer [ref], or to allow celebrities such as David Beckham to deliver public health messages about malaria in nine different languages [ref]. Museums are using it to bring the dead back to life (at the Salvador Dalí Museum visitors can interact with a synthetic Dalí to learn about his art [ref]), while combining it with natural language processing could one day lead to smart digital assistants capable of truly natural interactions [ref]. Synthetic media can also be used as a digital veil to swap the faces and voices of witnesses or confidential sources during court proceedings [ref], and is already helping sexual assault victims and marginalized groups share their experiences in film and documentaries [ref]. Voice skins can enable women and children, often subject to internet bullying, to control when and how their gender and age is shared with others, while members of the trans community can use it to more accurately reflect their identity when interacting online [ref].

Nevertheless, this technology is also ripe for abuse. Deepfaking has quickly become a tool of harassment against activists [ref], and is a growing concern for those in the entertainment, business, and political sectors [ref]. The ability to control a person’s voice or appearance opens companies to new levels of identity theft, impersonation, and financial harm [ref]. In one high-profile case, hackers Deepfaked a CEO’s voice and used it to trick an employee into initiating a six-figure wire transfer [ref]. The rich and famous, whose data is widely available online, also represent an easy target. The voices of male celebrities are being synthetically copied and digitally distributed [ref] while their female counterparts are being non-consensually grafted into pornographic scenes so realistic that only they know its fake [ref]. Elsewhere, politicians are being digitally manipulated into endorsing controversial positions [ref], while worry grows that a well-timed video in the days prior to an election could have them ‘confess’ to bribery or sexual assault, political disinformation that distorts democratic discourse and election outcomes [ref].

The dark side of synthetic media goes even further. Deepfakes have sparked a new disinformation frontier where malicious actors are using the technology to pose as journalists, analysts, or consultants [ref]. These fake identities are legitimized through connections to genuine professionals on LinkedIn [ref], and used to manipulate mainstream news outlets into publishing content for political or personal gain [ref]. Elsewhere, intelligence services and think-tanks warn that Deepfakes represent a growing cybersecurity threat, a tool that state-sponsored actors, political groups, and lone individuals could use to trigger social unrest, fuel diplomatic tensions, and undermine public safety [ref]. This technology can be used to distract the public, hijack narratives, and waste opponents’ time and resources by forcing them to fight lies and slander [ref]. Given the speed with which information now proliferates and how quickly individuals, systems, and governments react, these digital lies could be half-way around the world before the truth catches up. And the consequences could be catastrophic.

One day soon it will be impossible to tell with the naked eye or ear if content is genuine or synthetic. Recognizing this inflection point, industry leaders and lawmakers are looking to two forms of protection. Politicians, in Europe and the USA, are advocating for legislation that regulates a technology they believe will further erode the public’s trust in media and push ideologically opposed groups deeper into their own subjective realities [ref]. At the same time, technology giants such as Facebook, Google, and Microsoft are developing algorithms to detect Deepfakes, excise them from their platforms, and prevent their spread [ref]. Although legislative and technological stopgaps are undoubtedly necessary, they are also in a perpetual game of ‘cat-and-mouse’, with certain actors evolving new ways of evading detection, and others rapidly working to catch up. In such a world, no law or algorithm can guarantee that the public will be completely protected from malicious synthetic content [ref].

What is needed then, alongside legislation and technological fixes, is a greater focus on the *human* dimension. It’s imperative that we study the impact of this new technology on our thoughts, feelings, and actions. For instance, can Deepfakes be used to manipulate our (unconscious) attitudes and intentions? How effective are they in doing so, especially when compared to authentic content? Are people aware of this new technology, and perhaps more importantly, can they detect when they are being exposed to it? Finally, does awareness of Deepfaking and the ability to detect when it is present immunize people from its influence?

We carried out seven pre-registered studies (*n* = XXX) to answer these questions. We first created a set of genuine baseline videos in which an unknown target (‘Chris’) disclosed personal information about himself. In one video he emitted positive self-statements while in another he emitted negative statements. One group of participants navigated to YouTube (where the videos were hosted), watched the positive or negative variant, and then completed measures of their self-reported attitudes, automatic attitudes, and behavioral intentions. Results indicated that genuine online content strongly influenced how Chris was perceived (see Fig 1.) stats here. \*

A second group encountered a similar procedure but with one key difference: they watched a Deepfaked video. Deepfakes were created by taking the genuine content outlined above, fitting a parameterized 3D model to the target’s head, and using this model to create computer graphical (CG) renderings of his face and mouth movements. These renderings were then converted to photorealistic synthesized video using a trained Generative Adversarial Network (GAN) [ref]. By inserting the positive synthetic statements into the negative genuine video, and vice-versa, we created a set of Deepfakes wherein a target’s actions were fabricated and manipulated. Selectively exposing people to one of these Deepfakes allowed us to control how the target was perceived, liked by some and despised by others (see Fig X.) stats here.

Similar findings also emerged when a different method was used to create Deepfakes, one that generated content from scratch, rather than extracting it from one video and inserting it into another. This involved taking pre-existing footage from a different actor and using it to generate a 3D head model. This model was then used to perform iterative localized edits on the genuine videos (i.e., to transform positive statements into negative statements and vice-versa). Digitally manipulating the target’s actions in this way allowed us to control attitudes and intentions towards him (see Fig X) stat here.

The above findings also generalized from one synthetic media type (video) to another (audio). We created a training set of the target’s voice and then fed it to a bidirectional text-to-speech (TTS) autoregressive neural network (see [ref]). This resulted in a Deepfake of the target’s voice: a synthetic replica that sounded like the original and which could be manipulated into saying anything. Participants were informed that they would listen to a recording of Chris, and then either exposed to the Deepfaked voice, or a genuine recording of him emitting positive or negative self-statements. We found that by synthetically cloning a person’s voice, and manipulating what he ‘said’, we could control people’s (automatic) attitudes and intentions towards him (see Fig X.) stats here.

Taken together, our findings show that Deepfakes can be used to bias what people think and feel. But how *effective* they are in doing so? Most - including our own - contain video or audio artefacts which represent ‘tell-tale’ signs of manipulation. It’s possible that these artefacts undermine the effectiveness of Deepfakes relative to genuine content. Yet in our studies this was never the case: Deepfakes were just as effective in altering attitudes and intentions as authentic content (see Fig X.) stats here.

It’s also important to know if (a) people are aware that online content can be Deepfaked, and (b) if they can detect when they are being exposed to it. Our findings were not encouraging: roughly half of those who took part in our studies had never previously heard of Deepfaking (XX%), and even after they were told what it entailed, many were unable to detect if the video or audio they had just encountered was genuine or synthetic in nature (i.e., they were unable to make accurate or informed judgements about the authenticity of online content) stats here. That said, people who were previously aware of Deepfaking were XX times more likely to detect when they were exposed to a forgery relative to their unaware counterparts. stats here.

Finally, is it the case that an awareness of Deepfaking, or an ability to detect when it is present, protects the viewer from its influence? Unfortunately, this was never the case in our studies: these individuals were just as likely to be manipulated by Deepfakes as those without awareness or who thought a Deepfaked was genuine (see Fig. X) stats here. Even those who were both aware *and* who detected the Deepfake still fell prey to its influence.

In short, Deepfakes do not need to be undetectable or even perfectly convincing in order to psychologically impact the viewer. They can be used to manipulate attitudes and intentions just as effectively as authentic content. Many are unaware of this new technology, find it difficult to detect when they are being exposed to it, and neither awareness nor detection served to protect individuals from its influence.

Given the dangers posed by malicious synthetic content (Deepfakes) politicians are looking to the law to help regulate its creation and spread while industry leaders invest in technological solutions to help consumers detect and recognize when they are exposed to it. Our findings suggest that this won’t be enough: a single brief exposure to a Deepfake quickly and effectively shifted (unconscious) thought and feeling, even when people were aware of Deepfaking and detected when they were being exposed to it.

What is needed then is a better understanding of the *psychology* of Deepfakes - and in particular - how they exploit our cognitive biases, vulnerabilities, and limitations for maladaptive ends. We need to identify the properties of individuals, situations, and/or content that increase the chances that Deepfakes are believed and spread versus detected and rejected. We need to examine if these lies root themselves quickly and deeply in our minds, and linger on long after efforts to debunk them have ended (as is the case with fake news; [ref]). If so, then approaches currently favored by tech companies, such as tagging Deepfaked videos with a warning, may be less effective than now assumed [ref]. We also need to examine if Deepfakes can be used to manipulate what we remember, either by trigger Mandela effects (i.e., installing false memories that never happened) or by altering genuine memories that did [ref]. If they can influence memory then it is not only the present and future that can be influenced but also the past.

Perhaps the most dangerous aspect of Deepfakes is their capacity to erode our underlying belief in what is real and what can be trusted. Instead of asking if a specific image, video, or audio clip is authentic, this new technology may cause us to question *everything* that we see and hear, thereby accelerating a growing trend towards epistemic breakdown: an inability or reduced motivation to distinguish fact from fiction. This “reality apathy” [ref] is already being exploited by certain actors to dismiss inconvenient or incriminating content as a fabrication (the so-called ‘liars’ dividend’ [ref]). Given that the human mind is built for belief [ref], we need psychological interventions that can inoculate individuals against synthetic media attacks, and together with technology and legislation, to create a ‘shared immune system’ that safeguards our individual and collective belief in truth. Without such safeguards we may be speeding towards a world where our individual and collective ability to agree on what’s true eventually disappears.

**Abstract:** The abstract should be 100-125 words, and organized in this structure: An opening sentence that sets the question that you address and is comprehensible to the general reader, background content specific to this study, results, and a concluding sentence. It should be a single paragraph.

**One Sentence Summary:** A brief teaser statement highlighting main result of the paper, understandable by a scientist not in your field, without jargon or abbreviations. This will appear online adjacent to the title and should not repeat phrases already present there. Please keep to under 125 characters.

**Main Text:** In general, this should include a brief (1-2 paragraph) introduction, followed by a statement of the specific scope of the study, followed by results and then interpretations. Please avoid statements of future work or claims of priority, and avoid repeating the conclusions at the end.

**Subheadings** (“Results”, “Discussion”, or more specific subheadings, but not a leading “Introduction”) may be included in Research Articles or Reviews and should be brief, set off by a line break and formatted in bold face. Reports should not have subheadings.

All Figures and Tables should be cited in order, including those in the Supplementary Material (which should be cited as, for example, “Fig. S1”, and “Table S1”). You may include line or page breaks if you would like to place the figures within the text near where they are referenced. Please do not place figures in text boxes.

References should be cited in parentheses with an italic number (*1*). Multiple reference citations are separated by commas *(2, 3)* or if a series, dashes *(4-6)*. References are cited in order by where they first are called out, through the text, captions, and then the supplementary material.

Equations can be included. We do not recommend using the native Word 2007, 2008, 2010, or 2011 equation editor. This can in some cases produce less reliable MathML, the online markup language we use, which may result in display errors. Instead, use the legacy equation editor in Word (Chose Insert > Insert Object > Word Equation) or use Mathtype (recommended). If you enter equations in simple LaTeX, check that they will convert accurately (Word 2007 and higher can convert simple LaTeX equations).

References and Notes:

1. There is only one reference list including all references in the text, Figure and Table captions, and Supplementary Material. Do not include a second reference list in the Supplementary Material. References only cited in the Supplementary Material are not counted toward length guidelines.
2. Each reference should be on a separate line ending in a period. For a style guide, see [http://www.sciencemag.org/authors/instructions-preparing-initial-manuscript](http://www.sciencemag.org/authors/instructions-preparing-initial-manuscript%20).
3. You should include titles in references and full page ranges. Titles will not be included in the print version of the paper, but will be shown in the online version.
4. Please include the above heading, “References and Notes:”.
5. You can use a numbered list in Word.
6. Each reference should have a separate number.
7. Please do not mix in references with explanatory notes.

**Acknowledgments:** Acknowledgments follow the references and notes but are not numbered. Acknowledgments should be gathered into a paragraph after the final numbered reference. This section should start by acknowledging non-author contributions and then should provide information under the following headings: **Funding:** include complete funding information, including grant numbers; **Author contributions:** a complete list of contributions to the paper [we encourage you to follow the [CRediT](http://docs.casrai.org/CRediT) model]; **Competing interests:** competing interests (including but not limited to patents, financial holdings, professional affiliations, advisory positions, and board memberships) of any of the authors must be listed (all authors must also fill out a separate, internal Conflict of Interest form). Where authors have no competing interests, this should also be declared (e.g., “Authors declare no competing interests.”); and **Data and materials availability:** Include a note explaining any restrictions on materials, such as materials transfer agreements. Note accession numbers to any data relating to the paper and deposited in a public database; include a brief description of the data set or model with the number. If all data are in the paper and supplementary materials include the sentence “All data is available in the main text or the supplementary materials.” All data, code, and materials used in the analysis must be available in some form to any researcher for purposes of reproducing or extending the analysis.

Supplementary Materials:

Materials and Methods

Figures S1-S#

Tables S1-S#

Movies S1-S#

Audio Files S1-S#

External Databases S1-S#

References (*##-##*)

**Fig. 1.** The figure caption should begin with a short descriptive statement of the entire figure followed by additional text. Captions should be immediately after each figure. Figure parts are indicated with capital letters (A). If you prefer, you can place both figures and captions logically throughout the text near where they are cited rather than at the end of the file (but not both). If a paragraph in the main text begins with the name of a figure, write out “Figure” in full (e.g., <para>“Figure 1 shows….”)

**Fig. 2.** You can place graphics in-line above each caption. Please do not use text boxes to arrange figures. All images should be JPEG, PNG, TIFF, or similar standard format. High resolution (preferably editable PDF or Adobe Illustrator format) figures will be requested following review.

**Table 1.** Start this caption with a short description of your table. Format tables using the Word Table commands and structures. Do not create tables using spaces or tab characters.

(Please delete before submission) Supplementary materials should be included in a separate supplementary materials file. A template for this file can be found at: <http://www.sciencemag.org/sites/default/files/Science_Supplementary_Materials_Word_template.docx>.

1. Department of Experimental Clinical and Health Psychology, Ghent University, Ghent, Belgium. 2 Interdisciplinary Center, Herzliya, Israel. 3 Department of Psychology, Yale University, USA. 4 Department of Computer Science, Stanford University, USA. 5 Fermi National Accelerator Laboratory (Fermilab), USA. 6 Rudolf Peierls Centre for Theoretical Physics, Oxford University, UK.

   \***Corresponding author. Email: sean.hughes@ugent.be (S.H.)** [↑](#footnote-ref-2)