***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 Human 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”, but thanks to recent advances in artificial intelligence this age old saying may no longer be the case. A new branch of machine learning, known as ‘deep neural networks’, has emerged and 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 be used to convince others that what they’re seeing, reading, or hearing is fact rather than fiction. Although mainly used to mimic real individuals, this technique can also be used to generate photographs of people who do not exist [ref], synthetic voices that belong to no-one [ref], and synthetic text that sounds human-authored [ref].

Synthetic media is becoming highly realistic, easier to produce, and according to a recent report, the online presence of one subcategory known as ‘Deepfakes’ is doubling 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. Like any new technology it can be used 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 Deepfaked Dalí to learn about his art [ref]), while combining it with natural-language learning 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].

However, this technology is also ripe for abuse. Deepfakes have quickly become tools of harassment against activists [ref], and are 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 recent 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 inserted into pornographic scenes so realistic that only they know its fake [ref]. are being

The dark side of synthetic media goes even further. Deepfakes are giving rise to a new disinformation frontier where malicious actors are using the technology to create identities for themselves as journalists, analysts, and political consultants. These identities are akin to a digital prosthesis, worn by a real human and used for political or personal gain [ref]. These actors have sought to legitimize themselves by building connections with professional targets on LinkedIn [ref] and have already tricked dozens of main-stream news outlets into publishing their opinion pieces [ref]. Deepfakes have also been used to rewrite history (one such example involves President Nixon claiming that the Apollo 11 moon landing never happened [ref]), while intelligence services and think-tanks warn that they 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]. A convincing video of a world leader declaring the launch of a nuclear weapon or severe economic sanctions on a trading partner could incite violence, trigger stock-market volatility, and damage foreign relations. 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 gets its boots on. And the consequences could be catastrophic.

Soon it may be impossible to tell with the naked eye or ear whether a video or audio clip is authentic. Recognizing this inflection point, industry leaders and lawmakers are looking to two forms of protection. Politicians, both 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 even 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 always be protected from contact with malicious synthetic content [ref].

What is needed then, alongside legislation and technological fixes, is a greater focus on the human dimension, namely, the *psychology* of Deepfakes. We need to understand how synthetic media comes to shape what people think, how they feel, and what they ultimately do. To illustrate, take a branch of psychology known as person perception. So far Deepfakes have been used to target prominent individuals such as Barack Obama, Donald Trump, and Mark Zuckerberg [ref]. However, as the availability of this technology grows, and the online sharing of personal data continues, it will become increasingly easy to scrape anyone’s data, use it to generate a Deepfake, and have them apparently act in ways that either enhance or destroy their reputation. This raises the question: how easily can a Deepfake shift our attitudes and intentions towards another person, especially one we’ve just met?

To answer this, we created a set of videos wherein an actor (‘Chris’) disclosed personal information on his YouTube channel. In one video he emitted positive self-statements (e.g., “*When I was in college I helped my friend study for his final exam*”) while in another he emitted negative statements (e.g., “*I won’t give up my seat on a bus if I see a heavily pregnant woman standing. It’s not my problem if she needs it more than me*”). A first group of participants were asked to navigate to YouTube (where the videos were hosted), watch either the positive or negative variant, and then provided measures of their self-reported attitudes, automatic attitudes, and behavioral intentions towards Chris. Results indicated that genuine videos strongly influenced people’s first impressions of Chris (see Fig X.): stats here.

A second group encountered the same procedure but with one key difference: they watched a Deepfaked video of Chris. This video was generated by feeding authentic footage of the actor to a deep neural network and having it output a synthetic copy. These videos digitally manipulated Chris into emitting the same statements as in the genuine footage (for examples of the videos see link). We found that a single exposure to one of these Deepfakes strongly influenced how the target was perceived, and that by digitally controlling his actions, we could influence how much he was either liked or despised (see Fig X.): stats here. These findings replicated when a different algorithmic process was used to generate the Deepfakes and also generalized from one synthetic media type (videos) to another (audio). That is, in a separate set of studies, we fed audio recordings of the actor to a deep neural network and had it create a completely synthetic clone of his voice. One group listened to the clone saying the same statements as in the videos while another listened to genuine recordings of the actor. Synthetically cloning the target’s voice and manipulating what he said allowed us to control (automatic) attitudes and intentions towards the target (see Fig X.): stats here.

Demonstrating that Deepfakes influence person perception is an important first step. Nevertheless, many questions remain. For instance, how effective are Deepfakes in changing attitudes and intentions? Most - including our own - contain minor video or audio artefacts which, to a discerning eye, represent ‘tell-tale’ signs of manipulation. It’s possible that these artefacts undermine the believability and thus validity of synthetic media as an information source. If so, then we would expect Deepfakes to influence people to a lesser extent than authentic content. Yet comparing the meta-analytic effects of Deepfakes to those produced by the authentic videos revealed that the former were just as effective in shifting attitudes and intentions as the latter (see Fig X.) stats here. It seems that Deepfakes don’t have to be perfect in order to hack the human mind and do so in ways that are comparable to authentic media sources.

We were also interested in ‘Deepfake detection’ in humans. When synthetic media is used for malicious purposes the recipient is exposed to a forgery masquerading as something genuine. If people can detect when they are being exposed to such content they can prevent it from influencing their decision making. Unfortunately, when we explained the concept of a Deepfake to participants and informed them that they had just encountered one during the study, only a small fraction (X%) of those who were exposed to the Deepfake recognized it for what it was. The vast majority (X%) believed that what they had encountered was an authentic recording. Perhaps more worryingly, when we told participants in the genuine content group that the videos or audio were Deepfakes, many agreed (X%). This combination of poor false negative and false positive rates means that individuals are highly likely to make poorly informed decisions regarding the authenticity of the content (Sensitivity = XX, Specificty = XX, Youden's J = XX). It seems that in a world where Deepfakes exist, the mere suggestion that content is fake may be enough to make some question the validity of what they see or hear, even when that content is genuine [ref].

Finally, we wanted to know if ‘Deepfake detection’ serves to protect the individual from its influence. If one recognizes that the information they are being exposed to has been intentionally manipulated in order to influence them, then the recipient should reject that content or at least attempt to minimize its control. Worrying, however, we found that this was not the case: those who detected that the video or audio clip was a Deepfake were just as likely to be manipulated into liking or disliking Chris as those who failed to do so (see Fig. X; stats here).

Taken together, it seems that a single (brief) exposure to a Deepfake can alter our perceptions of others; that it impacts them in a similar way to genuine content; that most people are unaware they are being exposed to a synthetic forgery; and that Deepfakes do not need to be undetectable or even perfectly convincing in order to hack the human mind.

So far society has focused on legislation to regulate, and technological solutions to detect and filter out Deepfakes. This is not enough. We need to start studying the psychology of Deepfakes - and in particular - the capacity of this new technology to exploit our cognitive biases, vulnerabilities, and limitations for maladaptive ends. Future work should identify those properties of the individual, situation, and/or content that increase the chances of Deepfakes being believed and spread versus detected and rejected. Others could examine if these lies root themselves quickly and deeply in our minds, and linger on as insinuation or by association long after efforts to debunk them have ended (as is the case with more traditional forms of 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 is now assumed [ref]. Still others could examine if Deepfakes can be used to manipulate what we remember, either by trigger Mandela effects (i.e., false memories that never happened) or by altering memories of events that did [ref]. If they can then it is not only the present and future that is up for change but also the past.

Perhaps the most dangerous aspect of Deepfakes is their capacity to erode our belief in what is real and what can be trusted *in general*. Instead of questioning a single image, video, audio, or text this new technology may push us towards questioning *everything* we see and hear, thereby accelerating an already growing trend towards epistemic breakdown: an inability or reduced motivation to distinguish fact from fiction. This “reality apathy” [ref] is already being exploited by some 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 to start developing interventions that inoculate individuals against synthetic media attacks, and together with technology and legislation, create a ‘shared immune system’ that safeguards our individual and collective belief in truth. Without such safeguards we may be moving towards a world where seeing is no longer believing, and where our individual and collective ability to agree on what’s true slowly 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)