COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 46 (2022) e13191
© 2022 Cognitive Science Society LLC.
ISSN: 1551-6709 online
DOI: 10.1111/cogs.13191

Can You Hear Me *Now*? Sensitive Comparisons of Human and Machine Perception

Michael A. Lepori, Chaz Firestone

Department of Psychological & Brain Sciences, Johns Hopkins University

Received 30 April 2020; received in revised form 7 June 2022; accepted 26 June 2022

Abstract

The rise of machine-learning systems that process sensory input has brought with it a rise in comparisons between human and machine perception. But such comparisons face a challenge: Whereas machine perception of some stimulus can often be probed through direct and explicit measures, much of human perceptual knowledge is latent, incomplete, or unavailable for explicit report. Here, we explore how this asymmetry can cause such comparisons to misestimate the overlap in human and machine perception. As a case study, we consider human perception of *adversarial speech* — synthetic audio commands that are recognized as valid messages by automated speech-recognition systems but that human listeners reportedly hear as meaningless noise. In five experiments, we adapt task designs from the human psychophysics literature to show that even when subjects cannot freely transcribe such speech commands (the previous benchmark for human understanding), they can sometimes demonstrate other forms of understanding, including discriminating adversarial speech from closely matched nonspeech (Experiments 1 and 2), finishing common phrases begun in adversarial speech (Experiments 3 and 4), and solving simple math problems posed in adversarial speech (Experiment 5) — even for stimuli previously described as unintelligible to human listeners. We recommend the adoption of such "sensitive tests" when comparing human and machine perception, and we discuss the broader consequences of such approaches for assessing the overlap between systems.

Keywords: Perception; Psychophysics; Adversarial examples; Machine learning; Artificial intelligence

1. Introduction

How can we know when a machine and a human perceive a stimulus the same way? Machine-recognition systems increasingly rival human performance on a wide array of

Correspondence should be sent to Michael A. Lepori or Chaz Firestone, Department of Psychological and Brain Sciences, Johns Hopkins University, 3400 N Charles St, Baltimore, MD 21218, USA. E-mail: mlepori19@gmail.com and chaz@jhu.edu.

tasks and applications, such as classifying images (Krizhevsky, Sutskever, & Hinton, 2012; Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016), transcribing speech (Chan, Jaitly, Le, & Vinyals, 2016; Lamere et al., 2003), and diagnosing mechanical or biological anomalies (Cha, Choi, & Büyüköztürk, 2017; Lakhani & Sundaram, 2017), at least on certain established benchmarks (for critical discussion, see Raji, Bender, Paullada, Denton, & Hanna, 2021). Such advances often call for comparisons between human and machine perception, in which researchers collect responses from human subjects and machine-recognition systems and then ask how similar or different those responses are. In some cases, these comparisons serve to establish standards for determining when a machine-recognition system has reached so-called "human-level" performance (e.g., by recording average human accuracy rates on visual and auditory classification tasks). In other cases, the purpose of such comparisons is subtler, as in work that uncovers aspects of human perception that are shared with various machine-learning systems (Schrimpf et al., 2018). For example, recent work asks whether humans and machines that demonstrate comparable overall accuracy nevertheless show different patterns of errors (Eckstein, Koehler, Welbourne, & Akbas, 2017; Rajalingham et al., 2018), exhibit different biases (Baker, Lu, Erlikhman, & Kellman, 2018; Buolamwini & Gebru, 2018), are susceptible to different illusions and attacks (Elsayed et al., 2018; Jacob, Pramod, Katti, & Arun, 2019; Ward, 2019; Zhou & Firestone, 2019), or more generally display different behavioral and cognitive "signatures" (Taylor, Bastos, Brown, & Allen, 2022).

1.1. Knowing more than you can say

Regardless of their goals, any comparison between human and machine perception must confront several challenges associated with the different constraints these systems face. One such challenge involves finding the right tasks and measures to serve as the basis of these comparisons, since different tasks may be better suited to assessing comprehension in humans versus in machines. Indeed, whereas machine-recognition systems are often evaluated by more direct and explicit measures (since their classification decisions and other outputs are openly available to an experimenter), such measures — especially paradigms involving free descriptions of some stimulus under minimal constraints — are typically understood to be inadequate tests of human perception. One reason for this is that human perception almost always involves incomplete knowledge, thresholded responses, and even unconscious processing that may not be cognitively accessible to the perceiver, such that assessing what someone knows or perceives is rarely as straightforward as asking them to describe, label, or categorize what they see, hear, or feel.

Indeed, decades of research on human perception and cognition have revealed knowledge and abilities that were not initially evident from tasks in which subjects freely report what they know or observe. For example, even when subjects cannot explicitly predict how a swinging ball will travel when released from a pendular trajectory (Caramazza, McCloskey, & Green 1981), they may accurately place a cup to catch the ball — suggesting that they possessed the relevant physical knowledge all along but were able to access it only through more implicit processes (Smith, Battaglia, & Vul, 2018). Similarly, even when subjects incorrectly report the locations of remembered objects, they may perform reliably above chance if asked to take a second or third guess (Wu & Wolfe, 2018; see also Vul & Pashler, 2008). Even when subjects report no awareness of objects that are masked or appear outside the focus of attention, they may nevertheless show priming effects for the unnoticed stimuli, suggesting that those stimuli were processed unconsciously or below the threshold for explicit report (Kouider & Dehaene, 2007; Mack, 2003; for critical discussion, see Phillips, 2018). Even when subjects fail to notice statistical regularities in visual displays, they may still apply those regularities on subsequent trials, implying that they learned and incorporated such regularities implicitly (Chun, 2000; Chun & Jiang, 1998). And in an especially dramatic case, patients with cortical blindness who fail to freely report features of the objects they are looking at (e.g., being unable to answer questions like *"what orientation is the line in front of you?"*) may nevertheless succeed under forced-choice conditions (e.g., being able to correctly answer questions like *"is the line in front of you horizontal or vertical?"*; Weiskrantz, 1986, 1996; also Phillips, 2021).

1.2. Sensitive tests in human–machine comparisons

The above examples involve what we will refer to here as *sensitive tests* of human perception and cognition. Sensitive tests are tasks or measures that go beyond simply asking someone to describe what they see, hear, or know. Such techniques include making subjects act on a piece of information (rather than report it), exploring downstream consequences for other behaviors (as in priming studies), collecting additional responses (such as ranking various options rather than giving a single answer), or using some piece of knowledge to make a discrimination (rather than trying to report that knowledge directly).

How might this apply to comparisons between humans and machines? The fact that human perceptual knowledge can be partial, incomplete, or buried beneath layers of unconscious mental processing creates a challenge for comparisons of human and machine perception. In particular, whenever such comparisons rely mostly or only on explicit descriptions of sensory stimuli, there is a risk that these measures may underestimate what the human subjects really know about the stimuli they perceive, and thereby misestimate the overlap between human and machine perceptual processing (either leading to overestimation or underestimation of such overlap, depending on the comparison at issue).[1]

Here, we suggest that these considerations matter in concrete and measurable ways. In particular, we argue that at least some apparent gaps or disconnects between human perceptual processing and the outputs of various machine-perception systems can be explained in part by insufficiently sensitive tests of human perceptual knowledge. To demonstrate this, we explore an empirical case study of how using more sensitive tests can reveal a perceptual similarity when previous studies seemed to show a deep dissimilarity. Accordingly, we recommend that comparisons of human and machine perception consider such sensitive tests before drawing conclusions about similarities or differences in their perceptual processing.

1.3. A case study: Adversarial misclassification

An especially striking difference in human and machine perception is the one implied by adversarial misclassification (Szegedy et al., 2014). Adversarial examples are inputs designed to cause high-confidence misclassifications in machine-recognition systems, and they may crudely be divided into two types. One type is sometimes called a "fooling" example, in which a stimulus that would otherwise be classified as meaningless or nonsensical (e.g., patterns of image static, or auditory noise) is recognized as a familiar or valid input by a machine (e.g., a dog, or the phrase "OK Google, take a picture"; Carlini et al., 2016; Nguyen, Yosinski, & Clune, 2015). A second type is a "perturbed" example, in which a stimulus that would normally be classified in one way (e.g., as an orange, or a piece of music) can be very slightly altered to make a machine classify it in a completely different way (e.g., as a missile, or the command "Call 911 now") — even when such perturbations seem irrelevant (or are not even noticeable) to human observers (Athalye, Engstrom, Ilyas, & Kwok, 2018; Szegedy et al., 2014).

Such misclassifications are significant for at least two kinds of reasons. First, and more practically, they expose a major vulnerability in the security and reliability of machine-perception systems: If machines can be made to misclassify stimuli in ways that humans cannot notice, then it may be possible to attack such systems in their applied settings (e.g., causing an autonomous vehicle to misread a traffic sign, or making a smartphone navigate to a dangerous website) — a worry that may only intensify as such technologies become more widely adopted (Hutson, 2018). Second, and more theoretically, two systems classifying the same stimulus so differently would seem to undermine any other similarities they might show (Brendel et al., 2020), and even rule out the use of one to model the other.

Crucially, the reason adversarial misclassifications carry such important and interesting consequences is the very strong and intuitive sense that machines perceive these stimuli in ways that humans do not. And indeed, a growing literature has explored this intuition empirically, by asking human subjects to classify such stimuli and noting similarities and differences in their classification decisions (Carlini et al., 2016; Chandrasekaran, Yadav, Chattopadhyay, Prabhu, & Parikh, 2017; Elsayed et al., 2018; Harding, Rajivan, Bertenthal, & Gonzalez, 2018; Yuan et al., 2020; Zhou & Firestone, 2019; see also Baker, Lu, Erlikhman, & Kellman, 2018; Dujmović, Malhotra, & Bowers, 2020; Feather, Durango, Gonzalez, & McDermott, 2019; Golan, Raju, & Kriegeskorte, 2020). But might some of these discrepancies arise in part because of the means of comparison themselves?

1.3.1. Can people understand adversarial speech?

As a case study of this possibility, we consider here the example of adversarial speech. A recent and influential research program shows that it is possible to generate audio signals that are recognized as familiar voice commands by automated speech-recognition systems but that human listeners hear as meaningless noise (Carlini et al., 2016; Fig. 1). In the particular attack we explore here (see the General Discussion for other examples), a normal, human-recognizable voice command can be "mangled" by removing audio features not used by the speech-recognition system, such that it remains perfectly intelligible to that system but becomes incomprehensible to a human listener (see below for more detail).[2] To verify that such stimuli are truly heard as meaningless nonspeech, this work included an empirical comparison between the relevant speech-recognition system and a cohort of human subjects, who were played the audio clips and asked to make judgments about them. In particular, the

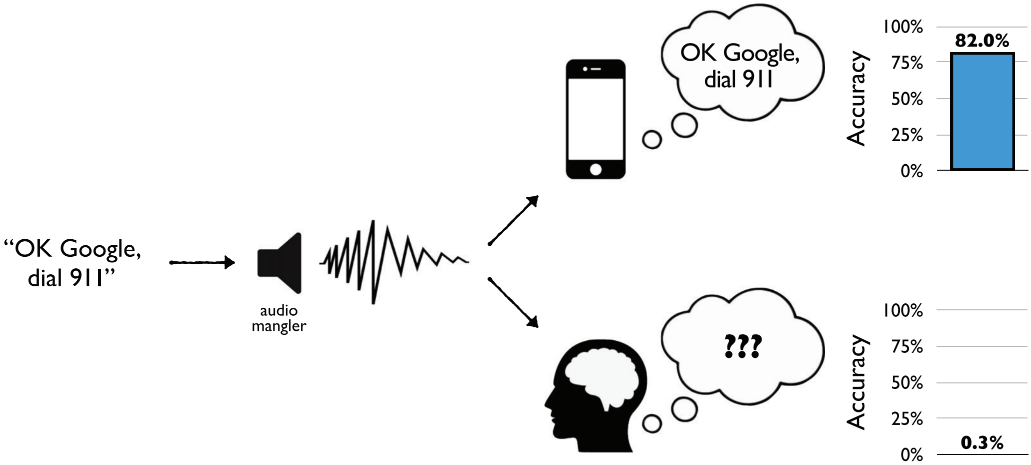

Fig. 1. Adversarial audio commands. Whereas human listeners and automated speech-recognition systems may transcribe normal human speech with comparable accuracy, adversarial speech involves synthetic or modified audio clips that human listeners hear as meaningless noise but that are recognized as valid commands by automated speech-recognition systems. As shown in Carlini et al. (2016), it is possible to take a speech clip and pass it through an "audio mangler" that erases many features that humans rely on in order to understand speech. This procedure results in a signal that is reliably recognized as a valid command by the machine-recognition system targeted by the attack, while human listeners are reportedly unable to understand it (as measured by their ability to freely transcribe its contents).

comparison asked the subjects to freely transcribe the audio files, and found that subjects were unable to transcribe the adversarial speech clips into the underlying messages from which the files were produced. These and similar tests led the authors to conclude that "0%" of subjects heard the hidden messages in the files, that subjects "believed our audio was not speech," and more generally that these commands were "unintelligible to human listeners."

However, accurate transcription is an almost paradigmatically "insensitive" test — an extremely high bar to study comprehension of this sort, including for the practical and theoretical issues raised above. For example, even in an applied setting, it could be valuable to know whether a human listener can tell *that* a hidden voice command was just played, even if the listener does not know the precise nature of the command (so that, e.g., the user could monitor or disable their device if they suspect it is being attacked). Similarly, for at least some purposes, understanding even part of a message might be nearly as good as understanding the whole thing. For example, a smartphone user who heard mostly nonsense in an adversarial speech clip — but managed to pick up the words "9-1-1" — might understand what their phone is about to do, even if they could not make out all or even most of the full message (e.g., "OK Google, please dial 911"). This is perhaps especially likely if they have a sense of the attacker's intentions or probable messages (e.g., the kinds of commands one would give to a phone in the first place). Finally, beyond such applied concerns, discovering that humans could extract these subtler patterns of meaning from adversarial speech could imply greater processing overlap (or least fail to imply a lack of overlap) than previous tests seemed to reveal. Might more sensitive tests show that this is the case?

1.4. The present experiments: Sensitive tests of adversarial speech comprehension

Here, we use perhaps the simplest of such sensitive tests — forced-choice classification — to explore how differences between human and machine perception of such stimuli may be misestimated by insensitive tests of human perceptual knowledge. We generate adversarial speech commands using the same method as in Carlini et al. (2016). But rather than ask subjects to freely transcribe these messages, we probe their understanding by asking them to make discriminations based on the commands. Experiments 1 and 2 ask whether subjects can discriminate adversarial messages that contain English speech from messages that do not contain speech but that are otherwise closely matched. Experiments 3 and 4 ask whether subjects can supply the last word in a familiar phrase begun in adversarial speech (e.g., "*It is no use crying over spilled* ____"), even without any advance knowledge of what words will be played. Finally, Experiment 5 asks whether subjects can answer simple math problems posed in adversarial speech (e.g., " $2 + 5 = ?$ "). To foreshadow the key results, all five experiments show that subjects can demonstrate comprehension of such speech clips when tested under sensitive conditions, even though these clips are not easily transcribed.

Of course, this work covers only one approach to generating adversarial speech (see the General Discussion for a review of other approaches), and adversarial speech is itself only one example of apparent divergence between human and machine perception. Our intention is thus for the particular results we report below to serve as a case study of the more general conclusion that sensitive tests can reveal latent or incomplete perceptual knowledge that less sensitive tests can miss, and in ways that directly inform comparisons between human and machine perception.

2. General methods

Though the five experiments reported here involve different hypotheses, designs, and research questions, all five proceeded in a very similar way and shared many methodological characteristics. What follows is thus a general methods section applying to all of the experiments below, followed by shorter methods sections for each individual experiment covering those key factors that differed.

2.1. Open science practices

The hypotheses, experimental designs, analysis plans, exclusion criteria, and sample sizes were determined in advance — and formally preregistered — for all five experiments. The data, materials, code, and preregistrations for all experiments reported here are available at https://osf.io/kp5j9/.

2.2. *Participants*

Each experiment recruited 100 different subjects (i.e., 500 subjects total across all five experiments) from Amazon Mechanical Turk (https://mturk.com). For discussion of this subject pool's reliability, see Crump, McDonnell, and Gureckis (2013), who use this platform to replicate several classic findings from cognitive and perceptual psychology. All subjects consented to participate in the study and were monetarily compensated for their participation.

All experiments applied exclusion criteria that required successful performance on attention-check trials and tests of the subjects' audio quality. These criteria varied slightly across experiments, and were always preregistered. Across all experiments, an average of 77% of subjects passed all exclusion criteria and were thus included in subsequent analyses. However, we note that no result reported here depended in any way on these exclusions; in other words, all of the results reported below remained statistically significant, in the same direction, even without excluding any subjects at all.

2.3. *Generating audio commands*

All experiments involved the presentation of adversarial speech. Carlini et al. (2016) present two methods for generating these speech stimuli: a black-box method, which is agnostic to the targeted speech recognition system, and a white-box method, which requires prior knowledge of the particular speech system of interest. To generate these speech stimuli, we used the white-box method.[3] We chose the white-box method because it was identified by the original authors as "significantly better than the black-box attack" at fooling human subjects, thus making it a more stringent test of our approach. For this attack, it was claimed that "0%" of subjects understood the speech clips, and that subjects "believed our audio was not speech." The white-box method is designed to fool the CMU Sphinx speech recognition system.[4] In brief, the Sphinx system extracts features from the raw audio input using the mel-frequency cepstrum transformation (mel-frequency cepstrum coefficients; MFCCs), as well as simple transformations on the computed values. These features are used as input to a hidden Markov model (HMM), whose states correspond to phonemes. This HMM outputs a probability distribution over feature vectors that could be generated next. A Gaussian mixture model (GMM) is used to represent this distribution over feature vectors.

The white-box method exploits knowledge of the parameters of the GMM (coefficients, means, and standard deviations for each Gaussian), as well as the dictionary used to transform words into phonemes. Given a string of text, this algorithm decomposes that string into a sequence of phonemes, and then attempts to generate input audio that will cause the Sphinx's HMM to generate that sequence. To do this, the algorithm uses knowledge of the GMM's parameters in order to calculate a series of MFCC feature vectors that maximize the likelihood under the GMM. Finally, the white-box method uses gradient descent in order to generate audio clips that are optimized to produce these MFCC feature vectors. Throughout this process, the audio corresponding to each phoneme is kept as short as possible, which further distorts the final clip.

Carlini et al. (2016) note that Sphinx correctly transcribes the audio clips generated by this method when they are fed directly into the system as audio files, but they are sometimes not transcribed correctly when they are played through a speaker and recorded with a microphone. They describe a method to improve the adversarial audio's over-the-air transcription ability, but we choose to use stimuli that do not correct for these difficulties. We do this to maximize the difficulty of comprehending these hidden audio commands, so that it will be especially clear how sensitive tests can reveal such comprehension.

In a supporting study (run after all of the studies reported below), we confirmed that subjects do not demonstrate understanding of the adversarial speech stimuli when assessed by free transcription. We selected one clip from each of the studies we report below (so that the stimulus sets would be comparable in size to past work), and presented 100 subjects either adversarial or uncorrupted text-to-speech versions of these stimuli. These subjects were tasked with transcribing the audio clips, as in Carlini et al. (2016).

We ran two versions of this task: One in which subjects could play the clips only once, and one in which (a separate group of 100) subjects could play the clips as often as they liked. In the single-play condition, subjects in the adversarial condition transcribed the audio correctly 0.6% of time, and transcribed the uncorrupted stimuli correctly 97% of the time. In the multiplay condition, subjects in the adversarial condition performed at 0% (failing to transcribe even a single example correctly), whereas subjects in the uncorrupted condition transcribed the stimuli correctly 99.1% of the time. This assured us that our adversarial stimuli were as incomprehensible as those in Carlini et al. (2016), who report broadly similar performance.

Finally, we note that the approach we take here (which uses same code as in the original study) builds into the stimulus-generation process all of the internal knowledge of the particular version of Sphinx that was targeted by the original, and so essentially "fools" that internal model of Sphinx as it generates audio.

All audio files used here, as well as the code for generating them, are available in our archive of data and materials.

3. Experiment 1: Which one is speech?

Can more sensitive tests reveal deeper human understanding of adversarial speech? Experiment 1 asked whether forced-choice conditions could allow human subjects to distinguish adversarial speech from closely matched nonspeech (even without requiring that subjects report the *content* of the speech; see Experiments 3–5 for tests of more contentful comprehension). We synthesized several dozen adversarial speech commands that previous work suggested should be "unintelligible to human listeners" and "believed [to be] not speech" (Carlini et al., 2016) and then played these commands to subjects either forwards or backwards (Fig. 2a). We predicted that subjects would hear the forwards-played audio as more speech-like than the backwards-played audio, even though the two kinds of clips were matched on many low-level features (since these two trial types involved the very same audio clips — the only difference was whether they were played forwards or backwards). If so, this would suggest that subjects do hear such audio clips as speech after all, in ways that would suggest a greater overlap in how such audio is processed by human listeners and the relevant speech-recognition systems.

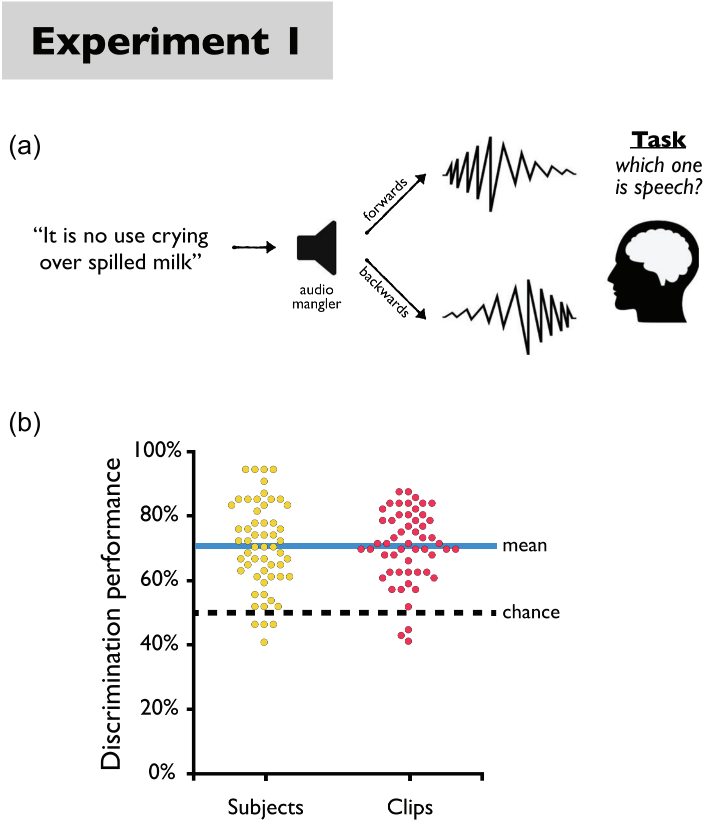

Fig. 2. Design and results of Experiment 1. (a) On each trial, subjects heard two adversarial speech clips, one forwards-played and one backwards-played version of the same adversarial audio command; their task was just to say which contained English speech. (b) Subjects correctly discriminated adversarial speech from these closely matched nonspeech clips. The vast majority of subjects classified the clips correctly more often than they classified them incorrectly (yellow), and the vast majority of speech clips were classified correctly more often than they were classified incorrectly (red).

3.1. Methods

3.1.1. Stimuli

We generated 54 hidden audio commands using the procedure described above. To select the content of the speech commands, we chose common idioms, quotes from history and media, or natural sequences — for example, "laughter is the best medicine," "we have nothing to fear but fear itself," and "1 2 3 4 5 6 7 8 9 10 11 12." (See materials archive for the full list of phrases.) We chose such phrases instead of completely arbitrary collections of words so that we could roughly match the familiarity of the phrases used in Carlini et al. (2016) — which included, for example, the phrases "take a picture" and "text 12345." Thus, in both our study and in past studies, the stimuli included words and phrases that a typical listener would have heard before, even though the subjects had no advance knowledge of which particular words would appear. Notably, this is also the case for the likely words that a malicious attacker might transmit to a smartphone or home assistant (e.g., messages involving key words or phrases, such as “call,” “browse,” “unlock,” etc., as well as small numbers).

In addition to the 54 audio clips containing these messages, we also generated a corresponding set of 54 audio files that simply played those very same clips in reverse. This process ensured that these two sets of stimuli were fairly minimal pairs, matched for many low-level auditory characteristics, including average length, frequency, intensity, and so on (as well as the variance in such characteristics). Thus, these pairs differed primarily in whether they followed the patterns characteristic of human speech (though see Irino and Patterson, 1996, for discussion of important differences between forwards- and backward-played audio).

Finally, we generated one audio file containing a simple tone, to be used as a “catch” trial to ensure that subjects were engaged in the task and paying attention (see below). This file appeared five times in the experiment.

There were thus 109 audio files (54 speech + 54 nonspeech + 1 catch), and 59 trials (54 experimental trials containing one forwards clip and its backwards counterpart, and five catch trials). All clips were generated, stored, and played in .wav format.

3.1.2. Procedure

Subjects were told a brief story to motivate the experiment:

> *A robot has hidden English messages in some of the following audio clips. Can you help us figure out which clips contain the robot’s messages? We know that half of the messages have hidden English and half of them don’t; we want your help figuring out which ones are English and which are not.*

The experiment proceeded in a self-paced manner, with subjects triggering the playing of the clips. On each trial, subjects completed a two-alternative forced-choice task (2AFC). Two embedded audio players appeared onscreen, each loaded with a single clip that was played when the subject hit a “play” button. The two clips were always forwards and backwards versions of the same adversarial audio command (with left-right position on the display always randomized for each trial). After the subject played each clip at least once, they could select whether the left or right clip was the one that contained English speech. Subjects could play each clip additional times if they chose to before responding. After making their selection, the next trial began and proceeded in the same way. The command played on each trial was always randomly chosen (without replacement) from the 59 total trials (54 experimental and 5 catch), each of which was played for every subject.

Note that, even though these clips were of fairly well-known phrases, subjects had no advance knowledge of the particular words they should look for in the clips. As has been noted previously (Carlini et al., 2016), adversarial speech can sometimes be easier to decipher when one knows (or is “primed” by) what the message is supposed to be; but, as in previous work, no such priming was possible here (beyond knowledge that the message might be drawn from the extremely broad class of messages that includes all vaguely familiar phrases, idioms, and sequences in English).

To ensure that subjects were paying attention and that the audio interface was working properly, subjects were instructed how to behave on catch trials: “A few times in the experiment, instead of hearing some sounds from the robot, you will instead hear a simple beep or tone; whenever that happens, make sure to click ‘Right’.” No data from catch trials were included in our analysis, except as criteria for exclusion. Additionally, before beginning the experiment, a “sound test” was performed in which a single audio clip was played (Bach’s *Cello Suite No. 1*), and subjects had to say what kind of audio the clip contained (beeping, clapping, conversation, traffic, music, or ocean). Only if subjects selected “music” could they proceed to the experimental trials; any other option ended the experiment without collecting data.

We excluded subjects based on two criteria. First, any subject who failed to provide a complete dataset was excluded from our analysis. Second, any subject who failed to follow instructions on any one of catch trials (i.e., who selected “Left” when they should have selected “Right”) was excluded entirely. This was done to ensure that all subjects had read the instructions and were focused on the experiment. These exclusion criteria, along with the rest of the design and analysis plan, were preregistered.

Finally, as an additional sanity check, we also repeated the above experimental design using the five original adversarial speech stimuli created by Carlini et al. (2016), so that we could compare our results to the original work (for details, see Appendix A).

Readers can experience this task for themselves at https://perceptionresearch.org/adversarialSpeech/E1.

3.2. *Results and discussion*

Subjects demonstrated an ability to discriminate adversarial speech from closely matched nonspeech. Average accuracy on experimental trials was 70.6%, which significantly differed from chance accuracy (50%), $t(55) = 11.34, p < .001, d = 1.529$ (Fig. 2b, blue).[5] Thus, subjects were able to reliably determine whether an adversarial audio clip contained speech.

In addition to the preregistered analysis we report above, we also conducted two exploratory analyses to determine the generality or pervasiveness of subjects’ performance. Specifically, beyond the “raw” accuracy across all subjects and all trials (i.e., the proportion of trials in which subjects correctly classified a clip as speech or nonspeech), it may also be informative to know (a) the proportion of *subjects* who tended to classify such stimuli correctly (collapsing across all speech clips), as well as (b) the proportion of speech clips that tended to be classified correctly (collapsing across all subjects).

In fact, collapsing across all speech clips, 92.9% of subjects showed classification performance that was numerically above chance (Fig. 2b, yellow). In other words, the vast majority of subjects tended to classify such clips correctly rather than incorrectly, suggesting that the ability to hear adversarial speech as speech was quite widespread. Additionally, collapsing across all subjects, 94.4% of the 54 speech clips were classified correctly more often than they were classified incorrectly (Fig. 2b, red).[6]

Finally, we also discovered a strikingly similar pattern of results when this paradigm was repeated with Carlini et al. (2016)'s own stimuli. When subjects heard the five original adversarial speech clips studied in the past work, they performed at 68.9% discriminating speech from nonspeech, which significantly differed from chance accuracy (50%), $t(82) = 8.20$, $p < .001$, $d = 0.906$ (for additional detail, see Appendix A).

In other words, whereas previous methods involving speech transcription suggested that "0%" of subjects heard the hidden messages in the files, or that subjects "believed our audio was not speech," our approach here suggested that a large majority of subjects heard the speech clips as speech more often that not, and that nearly all clips were heard as speech more often than not. These results thus provided initial evidence that a more sensitive approach (here, using forced-choice classification) could reveal perceptual knowledge and abilities that less sensitive tests were unable to detect.[7]

4. Experiment 2: Speech or nonspeech?

The previous experiment revealed an ability to discriminate adversarial speech from closely matched nonspeech, when subjects are tested under sensitive forced-choice conditions. However, the psychophysically powerful 2AFC design may have given subjects an undue advantage, since success on this task requires only that subjects decide which clip sounds more speechlike, regardless of whether either of the clips actually sounds very much like speech. In other words, it is possible that subjects felt that neither clip sounded like speech much at all, but were still able to succeed on the test as long as they could tell which clip better resembled ordinary speech. Though this is, in many ways, the entire purpose of 2AFC task designs, people in the real world (i.e., where such adversarial attacks might actually be deployed) may not be in the position of subjects in Experiment 1. In that case, a natural question is whether subjects could identify adversarial speech as speech under looser conditions in which they must actively label a single clip as speech, rather than compare two clips.

Experiment 2 investigated this question by asking subjects to label single clips as speech or nonspeech. The design was extremely similar to Experiment 1, except that instead of 54 experimental trials each containing two speech clips (one forwards and one backwards), there were 108 experimental trials each containing one clip (either the forwards or backwards version of the 54 audio commands). Subjects' task was now to classify each clip as speech or nonspeech, rather than to decide which of two clips was more speechlike. Could subjects succeed even under these conditions? (Readers can experience this task for themselves at https://perceptionresearch.org/adversarialSpeech/E2.)

4.1. Results and discussion

Subjects demonstrated an ability to identify adversarial speech clips as speech. Average accuracy on experimental trials was 62.2%, which significantly differed from chance accuracy (50%), $t(82) = 11.34$, $p < .001$, $d = 1.253$.[8] Performance was comparable on forwards trials (63.0%) and backwards trials (61.5%). Thus, subjects were able to reliably determine whether an adversarial audio clip contained speech, by identifying adversarial speech as speech and closely matched nonspeech as nonspeech.

We also performed the same two exploratory analyses as described in Experiment 1. Collapsing across all speech clips, 91.0% of subjects showed classification performance numerically above chance. And collapsing across all subjects, 84.3% of the clips were classified correctly more often than they were answered incorrectly.

Thus, in addition to being able to tell the difference between adversarial speech and adversarial nonspeech, subjects could also identify a given adversarial speech clip as speech.

5. Experiment 3: Fill in the blank

The previous experiments suggested that human subjects can identify and discriminate adversarial speech from closely matched nonspeech. But this result says little or nothing about subjects' ability to understand the content of that speech. Indeed, it is possible that subjects were able to simply distinguish speech and nonspeech using the low-level differences between normal and reversed audio clips. For example, natural audio tends to have sharp attacks and long decays that are not present in reversed audio (Irino & Patterson, 1996). Experiment 3 thus asked whether forced-choice conditions could allow human subjects to display knowledge of the content of adversarial speech, by asking them to identify the next word in an adversarial phrase.

We took a subset of the speech clips from Experiments 1 and 2 and simply removed the last word from the phrases (so that, e.g., "*It is no use crying over spilled milk*" became "*It is no use crying over spilled*"), and then asked subjects to supply the final word under forced-response conditions. If they can do so, this would suggest that subjects can engage in deeper and more contentful processing of adversarial speech, beyond knowing which clips are speech and which are not.

5.1. Methods

Experiment 3 proceeded in the same way as Experiments 1 and 2, except as noted below.

5.1.1. Stimuli

To generate the stimuli, we analyzed performance on the 54 adversarial speech clips from Experiment 2, and selected the 20 phrases with the highest classification accuracy in that experiment. Using the same procedure described earlier, we generated new adversarial speech clips from these 20 phrases, shortened by one word. For example, "laughter is the best medicine" became "laughter is the best." (See materials archive for the full list of phrases.) Importantly, all of these missing "last words" (e.g., "medicine" above) were unique to a given adversarial speech clip, and none of these words were spoken in any of the other adversarial speech clips. These clips served as the experimental stimuli.

We also generated three files containing nonmangled speech using an online text-to-speech system. These commands contained an uncorrupted human voice reciting portions of the alphabet (e.g., "A, B, C, D, E..."). These were used as catch trials to ensure that subjects were engaged in the task and paying attention.

5.1.2. Procedure

Subjects were told a modified version of the story from Experiments 1 and 2:

> *A robot has hidden English messages in some audio transmissions that we've recovered. However, these audio clips have been "corrupted", in two ways. First, most of the transmissions sound very strange and garbled; the robot's "voice" is very different than a human voice. Second, they've all been cut short by at least one word; for example, a message that was supposed to be "O say can you see by the dawn's early light" might actually come across as "O say can you see by the dawn's early".*

After the subjects played the clip on a given trial (e.g., "It is no use crying over spilled"), the two buttons that appeared either contained the correct next word in the current phrase (e.g., "milk"), or a word that would complete a different experimental phrase (e.g., "medicine"). The incorrect option was randomly selected from the pool of last words for other phrases in the experiment (without replacement), such that each last word appeared twice in the experiment, once as the correct option and once as the incorrect option. The pairs of options were randomly generated for each subject, as was the order in which the clips were shown. Note that, even though these clips were of fairly famous phrases, subjects had no advance knowledge of the particular kinds of words they should look for in the clips, just as in Experiments 1 and 2. The only property uniting these phrases was that they were likely to be vaguely familiar (rather than, say, being likely to be about food, sports, or some other theme).

As in previous experiments, we excluded any subject who failed to provide a complete dataset, as well as any subject who answered any of the catch trials incorrectly.

Readers can experience this task for themselves at https://perceptionresearch.org/adversarialSpeech/E3.

5.2. Results and discussion

Subjects demonstrated an ability to identify the next word of a phrase contained in an adversarial speech command. Average accuracy on experimental trials was 67.6%, which significantly differed from chance (50%), $t(69) = 7.98$, $p < .001$, $d = 0.960$ (Fig. 3b, blue). Thus, subjects could reliably understand at least some of the content of the adversarial speech commands, and could apply this comprehension to identify the next word of the hidden phrase. Collapsing across all speech clips, 79.3% of subjects showed classification performance numerically above chance (Fig. 3b, yellow). Thus, most subjects were able to correctly identify the next word in the hidden phrase. Collapsing across all subjects, 19 of the 20 clips were completed correctly more often than they were completed incorrectly (and one was completed correctly exactly half of the time; Fig. 3b, red).

These results suggest that subjects can not only hear adversarial speech as speech, but can also decipher at least some of the content of such speech. Whereas the task in Experiments

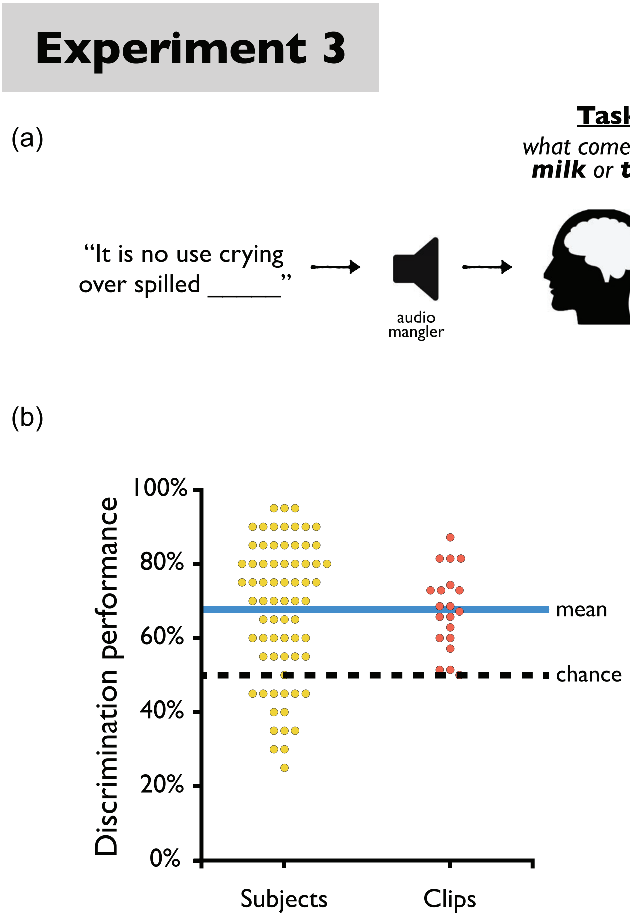

Fig. 3. Design and results of Experiment 3. (a) On each trial, subjects heard an adversarial speech clip that was missing its last word; their task was to identify the word that should come next. (b) Subjects correctly identified the next word in adversarial speech clips, and this was again broadly true across both subjects (yellow) and speech clips (red).

1 and 2 could have been solved by attending to phonetic or phonological features (such as prosody, intonation, or picking up one or two phonemes), Experiment 3 required some understanding at the level of words. This ability thus goes beyond what was previously shown in transcription tasks, further showing how sensitive tests can reveal perceptual knowledge that less sensitive tests cannot.

6. Experiment 4: Fill in the blank, after a single play

Experiment 3 suggested that subjects can not only hear adversarial speech as speech, but can also comprehend at least some of the content of that speech. However, it is perhaps possible that the forced-choice options themselves assisted subjects in parsing the speech, and in a way that could potentially undermine our interpretation of that experiment. When one is given a clue about what to hear in obfuscated speech, it may be easier to hear that speech in line with one's expectations (suggesting that speech processing is subject to top-down influence; see especially Remez, Rubin, Pisoni, & Carrell, 1981, as well as discussion in Firestone and Scholl, 2016a, 2016b; Vinson et al., 2016). In that case, one could imagine the following sequence of events: First, subjects heard an adversarial speech clip (e.g., "laughter is the best ___ ") and initially found it completely incomprehensible; second, they noticed the possible answers (e.g., "milk" vs. "medicine"); third, it occurred to them that not many well-known phrases end in "medicine," such that one of the possible phrases might be "laughter is the best medicine" (or "a taste of your own medicine," but perhaps few others); fourth, and finally, they replayed the adversarial speech clip while paying special attention to whether it might be the phrase "laughter is the best," and indeed were able to hear it that way.

Though this interpretation would still attribute to subjects some ability to comprehend adversarial speech, it is perhaps less impressive than subjects exhibiting this ability even without any such hints. So, to find evidence of this comprehension without any chance to hear the clips with hints, Experiment 4 repeated Experiment 3 with two small changes. First, the "last word" options on a given trial were revealed only after the adversarial speech clip had completely finished playing, such that every subject first heard the clip without any keywords that might tell them what to attend to. Second, though we still gave subjects the ability to play the clips multiple times, we recorded the number of plays a given subject made on a given trial, and preregistered a follow-up analysis including only those trials in which the subject chose not to replay the clip (i.e., those trials on which the first, clue-less play was the only play). If subjects can still identify the missing word in a clip even without any hint in advance about which particular words might be in the clip, and even on only a single clue-less play of the clip, this would be especially compelling evidence that subjects can and do comprehend aspects of the content of adversarial speech.

Readers can experience this task for themselves at https://perceptionresearch.org/adversarialSpeech/E4.

6.1. Results and discussion

Subjects again performed successfully. First, considering all trials (including multiplay trials), subjects anticipated the next word in the phrases with an accuracy rate of 67.8%, which significantly differed from chance (50%), $t(76) = 9.66, p < .001, d = 1.107$; this result replicates Experiment 3. But second, and more tellingly, even on trials in which subjects played the clip only a single time (such that they knew the last word options only after hearing the adversarial speech clip), subjects retained the ability to anticipate the next word in the adversarial speech phrases, with an accuracy rate on such trials of 69.8%, which significantly differed from chance (50%), $t(69) = 7.04, p < .001, d = 0.848$. Thus, subjects can comprehend the content of adversarial speech, even without advance knowledge of which words to hear in such clips.

7. Experiment 5: Math problems

Even though the results of Experiments 3 and 4 suggest that human listeners can hear meaning in adversarial speech commands, the task could still have been completed with a very minimal understanding. For example, a subject who was played “It is no use crying over spilled” could fail to comprehend almost the entire message, but still correctly guess “milk” over relevant foils if they just heard one telling keyword (e.g., “crying”). Could subjects instead complete a task that required parsing an entire message spoken in adversarial speech?

Experiment 5 asked whether subjects could correctly answer simple arithmetic problems that are contained in adversarial speech commands. We generated 20 audio commands containing these arithmetic problems (e.g., “six minus two”), and then asked subjects to select the correct answer from two options. On one hand, this approach constrains the space of possible messages from the extremely broad space explored earlier (i.e., the space of words that appear in familiar phrases) to a more limited space involving permutations of the integers 0 through 9 and the operations of addition and subtraction. On the other hand, this task remains especially challenging, because mishearing even one word of the problem would make it difficult or impossible to answer the question correctly. So, if subjects can succeed at this task, this would suggest that they can essentially understand every word of an adversarial speech command when tested under more constrained conditions.

7.1. Methods

7.1.1. Stimuli

We generated 20 arithmetic problems using the procedure described above. The problems contained two digits (from 0 to 9) and one instance of either the addition or subtraction operation. The problems were generated so that each digit (0–9) was the correct solution to two problems: one addition problem and one subtraction problem.

Similarly to Experiments 3 and 4, three files contained arithmetic problems spoken by an uncorrupted human voice, and served as catch trials to ensure that subjects were engaged in the task and paying attention.

7.1.2. Procedure

Subjects were told a modified version of the story from previous experiments:

> *A robot has hidden simple math problems in some audio transmissions that we’ve recovered. However, these audio clips have been “corrupted”. Most of the transmissions sound very strange and garbled; the robot’s “voice” is very different than a human voice.*
>
> [...]
>
> *We want you to help us by solving the math problems. On each trial, you will listen to a short audio clip. These clips contain simple addition or subtraction problems, containing two numbers from zero to nine. After you play the clip, you will be presented with*

two possible answers to the problem. Your job is just to select the correct answer. For example, the robot voice might say "5 plus 3". If that happens, you should select "8".

This experiment proceeded in a very similar way to Experiments 3 and 4. After the subject played the clip, one button showed the correct answer, and one showed an incorrect answer. The incorrect answers were randomly generated for each subject, as was the order in which the clips were shown.

Experiment 5 also contained catch trials of the same form as Experiments 3 and 4, and used the same exclusion criteria.

Readers can experience this task for themselves at https://perceptionresearch.org/adversarialSpeech/E5.

7.2. Results and discussion

Subjects demonstrated an ability to correctly answer arithmetic problems posed in adversarial speech. Average accuracy on experimental trials was 64.6%, which significantly differed from chance (50%), $t(79) = 8.27$, $p < .001$, $d = 0.930$ (Fig. 4b, blue). Thus, subjects could reliably comprehend most or all of the content of these audio commands, since such knowledge was necessary to answer the problem correctly.

Moreover, collapsing across all speech clips, 81.9% of subjects showed classification performance numerically above chance (Fig. 4b, yellow). Collapsing across all subjects, 18 out of 20 arithmetic problems were answered correctly more often than they were answered incorrectly (and one was answered correctly exactly half of the time) (Fig. 4b, red).

Whereas Experiments 3 and 4 could have been solved by understanding just one or two salient words, these results show that subjects displayed the ability to correctly decipher most (if not all) of the adversarial speech commands. In previous experiments, a subject who was played "It is no use crying over spilled" and noticed only the word "spilled" could still use that information to select "milk" as opposed to "time." However, here in Experiment 5, every word must be understood in order to solve these arithmetic problems, since even a single misheard word would make it nearly impossible to answer correctly. Thus, subjects can demonstrate an ability to understand whole phrases of adversarial speech, when tested in more sensitive ways.

One potential concern about this result is that it may not, after all, demonstrate comprehension of full speech commands, because subjects who heard only two words (e.g., "7 plus ***") or even just the operation itself (e.g., "*** minus ***") could still use this knowledge to perform above chance, even without comprehending the whole arithmetic problem. For example, consider a trial in which "9 minus 5" was played, but you heard only "*** minus ***"; if you then noticed that the two options for that trial were "9" and "4," you might make the educated guess that "4" is more likely than "9" to be the correct answer, because there is only a single problem that could have the word "minus" in it while still having the answer "9" (i.e., the problem "9 minus 0"), whereas many more problems containing "minus" could have the answer "4" (including, e.g., "9 minus 5," "8 minus 4," "7 minus 3," etc.). Since we have shown only that subjects perform above chance (and not, e.g., that they perform perfectly), it

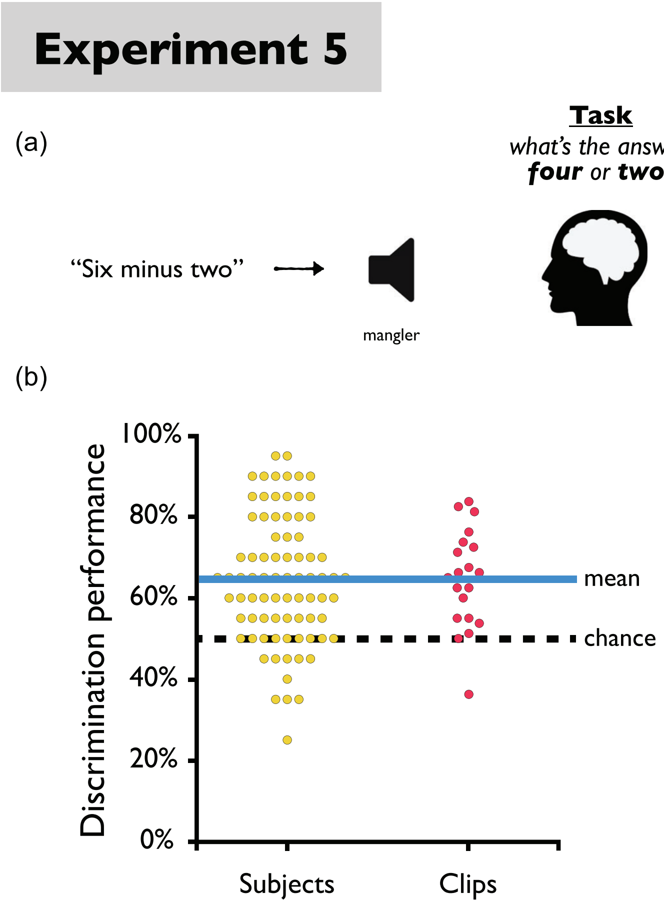

Fig. 4. Design and results of Experiment 5. (a) On each trial, subjects heard an adversarial speech clip expressing a simple arithmetic problem; their task was to supply the answer. (b) Subjects correctly answered the adversarial arithmetic problems, and this was again broadly true across both subjects (yellow) and speech clips (red).

is possible that this above-chance performance merely reflects strategic responding based on partial knowledge, rather than complete understanding of adversarial speech commands.

However, this concern can be overcome by examining only those trials in which the correct answer was the less probable one given the logic above. For example, suppose "9 minus 0" was played, and the options were again "9" and "4." If subjects on such trials correctly answer "9" at rates above chance, then that above-chance performance could not be explained by strategic responding after hearing "*** minus ***" (or even "9 minus ***"), since the optimal strategy in such cases of partial hearing would be not to answer "9." More generally, consider all trials in which either (a) the problem included "plus" and the correct answer was the lesser of the two options, or (b) the problem included "minus" and the correct answer was the greater of the two options. These trials are ones in which hearing only "plus" or "minus" would lead you away from the correct answer — and so above-chance performance on such trials could not be explained by such strategic responding.

In fact, when we carried out this analysis, it revealed that even on those trials in which strategic responding based on partial knowledge would produce the incorrect answer, subjects still performed far above chance: 67.8%, $t(79) = 8.53$, $p < .001$, $d = 0.960$. This analysis provides especially strong evidence that successful performance on math problems posed in adversarial speech goes beyond mere strategic responding, and implies a fuller and more complete understanding of the messages contained in such audio signals.

8. General discussion

What does it take to demonstrate that a human does not perceive a stimulus in a way that a machine does? Whereas previous work had identified a class of stimuli that machines comprehend but humans reportedly do not, here we showed that human subjects could display various forms of understanding when tested in more sensitive ways. By taking adversarial speech as a case study, we showed that even when humans could not easily transcribe an adversarial speech stimulus (the previous benchmark for human understanding), they nevertheless could discriminate adversarial speech from closely matched nonspeech (Experiments 1 and 2), finish common phrases started in adversarial speech (Experiments 3 and 4), and solve simple math problems posed in adversarial speech (Experiment 5) — even though such stimuli have been previously described as "unintelligible to human listeners."

Of course, this work covers only one approach to generating adversarial speech (see below for discussion of other approaches), and adversarial speech is itself only one example of apparent divergence between human and machine perception. Nevertheless, these results show how sensitive tests can reveal perceptual and cognitive abilities that were previously obfuscated by relatively "insensitive" tests, and in ways that directly inform comparisons between human and machine perception.

8.1. Increasing sensitivity

In Experiments 1 and 2, subjects reliably discriminated adversarial speech clips from the same clips played backwards. Whereas previous work suggested that subjects "believed our audio was not speech," our tests showed that subjects not only can identify such clips as speech, but can do so even when compared to audio signals that are matched on numerous low-level properties (since they were just the very same clips played in reverse). At the very least, these experiments suggest that subjects can attend to the phonetic or phonological cues that minimally distinguish speech from nonspeech, even when those cues are obscured by the adversarial-speech-generation process.

Experiments 3 and 4 showed that subjects can not only hear adversarial speech as speech, but can also comprehend the content of that speech. Whereas a subject could succeed in Experiments 1 and 2 simply by picking up patterns of prosody or segmentation, Experiments 3 and 4 asked subjects to fill in the last word of a phrase, and so required at least some comprehension of the adversarial speech clip. Moreover, this result could not be explained by straightforward forms of "priming." For example, Carlini et al. (2016) rightly note that, if one is told in advance what to hear in an adversarial speech clip, it is surprisingly easy to hear that message. But in our Experiments 3 and 4, the subjects were given no advance knowledge about which words they would encounter. Indeed, even though subjects knew that the phrases would likely be familiar, simply being told that one will hear a familiar phrase says little about which words will appear in that phrase. Instead, it tells you only that, if you can make out one or two salient words (e.g., "picture" and "thousand"), you might be likely to correctly guess what will come next (e.g., "words"). But even this situation first requires understanding some aspects of the speech clip, and so demonstrates a kind of comprehension that previous studies failed to reveal.

Finally, Experiment 5 created conditions that demonstrate complete (or nearly complete) comprehension of adversarial speech, by playing subjects spoken arithmetic problems and asking them to select the answer. For these problems (e.g., "six minus two"), every word of the problem was crucial to completing the task, since even a single misheard word would undermine one's ability to answer it. In a sense, this task is nearly equivalent to the free transcription task from previous work (since it required near-perfect comprehension to succeed), except that this experiment involved a more constrained space of word possibilities (since the subjects knew to expect the numbers 0 through 9, and the words "plus" and "minus"). Nevertheless, there was still considerable uncertainty for any message: Even under these constraints, there were 110 possible problems that could have appeared on any given trial, such that subjects did not have a straightforward or trivial way to know what they were "supposed to hear" in the message.

8.2. *Psychophysically inspired*

Our approach here was motivated by a central insight from the human psychophysics literature: simply asking people to describe what they see, hear, feel, or know can severely underestimate the nature and extent of such knowledge. Of the many reasons for this, one is that subjects may adopt conservative or idiosyncratic criteria when generating such explicit reports. We perceptually experience more than we can hope to include in any report of such experiences, and so when faced with very unconstrained tasks (such as freely describing what we hear in an audio clip), we must make choices about which aspects of our experience to describe, and in how much detail. These choices will be influenced by a host of factors, including which aspects of our experience stand out as remarkable, which we think the experimenter wants to hear about, and even just how engaged and motivated we are by the task. By contrast, more constrained tasks of the sort we explore here have the ability to zero in on those aspects of the subject's knowledge we are interested in, and in ways that require the subject to make very few (if any) choices about the relevance of various kinds of information. Indeed, similar considerations may even apply to tests of machine perceptual knowledge itself. For relevant work on this issue, see Zoran et al., 2015; Ritter et al., 2017.

In fact, strictly speaking, what we have explored here is not literally a task that is somehow more sensitive, but rather a task that makes adequate a previously inadequate measure.

Whereas previous work adopted "% correct" as the standard for comprehension of adversarial speech, this measure turns out to have been inadequate to reveal such comprehension when applied to a free transcription task. Perhaps some other measure (e.g., more advanced text mining) could, when applied to free transcription, capture the higher levels of comprehension that we observe here. But for our two-alternative forced-choice and forced-response tasks, the "% correct" measure was indeed adequate, because the task constrained subjects' responses to the variables and dimensions of interest, thereby revealing knowledge that was hidden by previous approaches.

The present work also joins other projects that have begun to explore similar themes. For example, Vadillo and Santana (2020) use adversarial speech stimuli consisting of regular audio overlaid with specific noise patterns that cause a speech recognition system to misclassify the audio clips. These authors also advocate for careful human subjects testing before describing an adversarial audio attack as "imperceptible," though their studies primarily focus on the more basic capacity to distinguish clips containing adversarial attacks from normal clips. By contrast, our present studies reveal that, in at least some cases, subjects can not only identify that adversarial clips contain speech, but can also comprehend the content of some adversarial audio attacks.[9] Other work has investigated human versus machine perception of "deepfake" stimuli, including both auditory and visual examples (Groh, Epstein, Firestone, & Picard, 2022; Müller, Markert, & Böttinger, 2021).

8.3. *Consequences for human–machine comparisons*

Increased understanding of the sort revealed here matters for at least two kinds of reasons:

First, and more practically, a major source of interest in adversarial attacks and other machine misclassifications derives from the security concerns they raise about the vulnerability of various machine-learning applications. For example, adversarial images could be used to fool autonomous vehicles into misreading street signs, and adversarial speech could be used to attack smartphones and home assistants without a human supervisor knowing. But the present results suggest that, at least in some circumstances, humans may be more aware of such attacks than was previously evident. Indeed, at least for the attack explored here, our Experiments 1 and 2 suggest that humans may well know *that* they are being attacked, even without knowing the precise way in which they are being attacked. And our Experiments 3 and 5 suggest that they may even have more sophisticated knowledge of the content of such attacks. Indeed, the constraint in Experiment 5, involving numbers ("0," "1," "2," ..., "9") and a few keywords ("plus," "minus"), is closely analogous to certain kinds of attacks that malicious actors might actually deliver to a phone (e.g., "dial 911").

But second, and more theoretically, human understanding of stimuli that fool machines is of interest to work in cognitive science (including both psychology and artificial intelligence) that explores similarities and differences between human perception and the analogous processes in various machine-learning systems. Adversarial attacks in particular are frequently invoked as evidence of "an astonishing difference in the information processing of humans and machines" (Brendel et al., 2020; see also Hendrycks and Gimpel, 2016; Sabour, Cao, Faghri, & Fleet, 2015; Serre, 2019; Buckner, 2019; Ilyas et al., 2019). But while it seems clear that humans and machines do not perceive such stimuli identically, it is still of interest to know just how similar or different their perception of such stimuli is. As we show here, the answer to this question can be surprisingly subtle: Humans who cannot freely report the content of such stimuli may nevertheless decipher them under certain conditions, in ways directly relevant to claims about overlapping or nonoverlapping processing across such systems.

Of course, none of the present experiments imply that all adversarial attacks will be comprehensible to humans in this way (though there are indeed other audio adversarial attacks that subjectively sound speechlike; Abdullah et al., 2019). For example, some audio adversarial attacks involve ultrasonic frequencies beyond the range of human hearing (Zhang et al., 2017); these attacks will certainly be incomprehensible to people. At the same time, that very fact can make them "weaker" or less threatening as attacks. For example, the developers of this attack note that an automated speech-recognition system could "defend" against it by restricting its processing to the frequency bands of human hearing (e.g., by modifying a microphone to "suppress any acoustic signals whose frequencies are in the ultrasound range"). And more generally, if the indecipherability of such attacks to humans owes fully to their occurring outside the range of human auditory perception, then this too is less an example of deep underlying differences in speech processing and more so an example of superficially different "performance constraints" facing these two systems (Firestone, 2020).

Relatedly, it has also been shown that adversarial speech can be embedded in clips of normal human speech (Carlini & Wagner, 2018a; Qin, Carlini, Cottrell, Goodfellow, & Raffel, 2019). These attacks seem particularly difficult to decipher, but at the same time, the original authors note here too that they are audible if you "listen closely" (Carlini & Wagner, 2018b), or more generally that they "can still be differentiated from the clean audio," in such a way that could still make them detectable as attacks. All these cases, then, show just how much nuance is required to make valid comparisons across human and machine perception.

8.4. Broader lessons

Though the present experiments explore sensitive comparisons in a case study involving adversarial speech, this approach could apply much more broadly to nearly any comparison of human and machine perception. Indeed, even recent work that does not use the language of "sensitive tests" may still be considered within this framework. For example, it had previously been claimed that it is possible to generate bizarre visual images that machines recognize as familiar objects but which are "totally unrecognizable to human eyes" (Nguyen et al., 2015). However, follow-up work using forced-responding showed that human observers can actually anticipate machine classifications when given relevant alternatives to choose from (Zhou & Firestone, 2019; though see Dujmović et al., 2020). For example, a series of cross-hatched red lines on a white background is recognized as a "baseball" by AlexNet; even though a human may not have initially been inclined to give the image that label, they are able to *select* "baseball" over certain alternatives under forced-response conditions not unlike those we explore here.

More generally, the approach we advocate here could apply to much broader questions about the overlap of human and machine processing. Even beyond adversarial misclassification, recent work has shown that natural language inference systems often rely on surface-level heuristics to make judgments about whether one sentence logically entails another (McCoy, Pavlick, & Linzen, 2019). This work also includes a human–machine comparison, where it is concluded that "human errors are unlikely to be driven by the heuristics targeted" in that work. This work could also benefit from a more sensitive test, such as one that eliminates the ability to reread the sentences, or perhaps introduces a time constraint. Under these strict constraints, it is possible that human judgments would be more informed by surface-level heuristics, revealing that some aspects of human cognition are reflected in the mistakes of the natural language inference systems.

Another human–machine difference that could benefit from more sensitive tests is the finding that Deep Convolutional Neural Nets tend to classify images based on texture rather than shape (Baker et al., 2018), whereas human subjects tend to classify based on shape rather than texture (Landau, Smith, & Jones, 1988). But the human subjects in Baker et al. (2018) were almost completely unconstrained in their testing conditions, being able to view the images and consider their labels for as long as they like. Perhaps forcing subjects to classify quickly, and after only a brief presentation, could bring the human and machine classification judgments into better alignment (see also Geirhos et al., 2018). For an exploration of this possibility, see Hermann (2022). Whereas the present work has used sensitive tests to reveal machine-like capabilities on a difficult task, these two examples demonstrate ways in which sensitive testing can be used to reveal machine-like deficiencies on fairly straightforward tasks.

Finally, even though the example we explore here shows how sensitive tests can reveal similarities where there previously seemed to be dissimilarities, the opposite pattern of results is possible as well. First, for some future class of stimuli, sensitive tests could well reveal that humans cannot comprehend, perceive, or process them the way a machine does. In that case, researchers could become especially confident in a given human–machine difference that survives even a sensitive test. Second, sensitive tests could also isolate very specific differences in how a human and a machine classify a stimulus. For example, if a human views the "baseball" image from Nguyen et al. (2015) and consistently prefers a specific alternative label (e.g., "chainlink fence"), this would suggest all the more strongly that the two systems represent this image differently.

8.5. Concluding remarks

People know and experience more than they freely report. Though this is a familiar and often-studied problem in cognitive psychology and perception research, it is also relevant to research comparing human perception and cognition to their analogous processes in machines. Here, we have shown how lessons from human perception research can directly inform and advance such comparisons, including in ways that reveal latent or implicit knowledge that was not evident from initial (and perhaps insensitive) comparisons. We thus advocate the adoption of more sensitive tests of human and machine perception, so that we can better explore when humans and machines do — or do not — perceive the world the same way.

Acknowledgments

For helpful discussion and/or comments on previous drafts, we thank Tom McCoy, Ian Phillips, and members of the JHU Perception & Mind Lab. For assistance with beeswarm plots, we thank Stefan Uddenberg. For resources relating to the production of adversarial speech commands, we thank Nicholas Carlini. This work was supported by a JHU ASPIRE Grant (M.L.) and the JHU Science of Learning Institute (C.F.).

Open Research Badges

This article has earned Open Data, Open Materials badges, and was pre-registered. Data are available at https://osf.io/kp5j9/, materials are available at https://osf.io/kp5j9/ and pre-registrations are available at https://osf.io/kp5j9/.

Data availability statement

The data, materials, code, and preregistrations supporting all of the above experiments are available at https://osf.io/kp5j9/.

Note

[1] Whether similar considerations also apply to tests of machine perceptual knowledge is an interesting open question. For relevant work on this issue, see Zoran, Isola, Krishnan, and Freeman (2015); Ritter, Barrett, Santoro, and Botvinick (2017). For a more general discussion of similar themes, see Firestone (2020).

[2] According to some restrictive definitions of "adversarial examples," stimuli only get to count as adversarial if they very closely resemble the original stimuli from which they were created. By those standards, the present case involves a misclassification, but perhaps not an adversarial one. By contrast, here we assume the broader and more popular definition given by Goodfellow et al. (Goodfellow et al., 2017; see also Szegedy et al., 2014): "Adversarial examples are inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake." By this definition, the case we explore here easily fits the bill.

[3] We thank Nicholas Carlini for generously sharing the code to create these stimuli.

[4] Note that CMU Sphinx (Lamere et al., 2003) has a non-neural architecture (based on a hidden Markov model) and so is in many ways unlike the systems that are much discussed today as sharing deeper aspects of human perception and cognition. But this fact only serves to strengthen any positive results in our experiments: If we can demonstrate greater-than-expected overlap between human perception and a machine-perception system that is not typically thought to have architectural similarities to the human mind and brain, then it should seem all the more impressive if humans are able to understand such messages. (Moreover, it is not actually clear that HMMs and other non-neural architectures are necessarily "worse" as models of human perception and cognition; indeed,

they were regularly used for just that purpose in a previous generation of computational cognitive science; Kaplan, 2008; Miller, 1952.)

5 A direct replication of Experiment 1, with extra reminders for how to behave on catch trials, excluded fewer subjects and produced a similar pattern of results: 63.8%, $t(72) = 8.43$, $p < .001$, $d = 0.994$.

6 Note that this does not mean that each of these subjects performed significantly above chance, or that each of these clips were identified as speech significantly above chance — only that 92.9% of subjects got the right answer more often than they got the wrong answer (whereas chance performance would predict that only 50% of subjects would do so), and that 94.4% of speech clips were classified correctly more often than they were classified incorrectly (whereas chance performance would predict that only 50% of clips would be correctly identified in this way). As stated in the main text, mean performance on any given trial was 70.6%.

7 See Appendix A for a replication of this experiment using stimuli from Carlini et al. (2016).

8 The nearly identical t-values in Experiment 1 and Experiment 2 are purely coincidental. (Means and SDs differed across the experiments, with Experiment 2 yielding worse overall performance but less variance and more included subjects.)

9 We note that Vadillo and Santana (2020) also demonstrate that subjects found some adversarial stimuli to be less natural than nonadversarial stimuli. This difference is more in line with the present work, as it demonstrates subjects' perception of specific differences between clips that may signal that the stimuli contain an adversarial attack.

References

Abdullah, H., Garcia, W., Peeters, C., Traynor, P., Butler, K. R. B., & Wilson, J. (2019). Practical hidden voice attacks against speech and speaker recognition systems. *NDSS'19*, 1369–1378.

Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2018). Synthesizing robust adversarial examples. In *35th International Conference on Machine Learning* (pp. 284–293).

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLOS Computational Biology*, *14*(12), e1006613.

Brendel, W., Rauber, J., Kurakin, A., Papernot, N., Veliqi, B., Mohanty, S. P., Laurent, F., Salathé, M., Bethge, M., Yu, Y., Zhang, H., Xu, S., Zhang, H., Xie, P., Xing, E. P., Brunner, T., Diehl, F., Rony, J., Hafemann, L. G., Cheng, S., Dong, Y., Ning, X., Li, W., Wang, Y. (2020). Adversarial vision challenge. In *The NeurIPS'18 Competition* (pp. 129–153). Springer.

Buckner, C. (2019). The comparative psychology of artificial intelligences. *philsci-archive*.

Buolamwini, J., & Gebru, T. (2018). Gender shades. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume *81*, (pp. 77–91).

Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M., Shields, C., Wagner, D., & Zhou, W. (2016). Hidden voice commands. In *25th USENIX Security Symposium* (pp. 513–530).

Carlini, N., & Wagner, D. (2018a). Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)* (pp. 1–7). IEEE.

Carlini, R., & Wagner, D. (2018b). Audio adversarial examples [https://nicholas.carlini.com/code/audio_adversarial_examples/].

Cha, Y. J., Choi, W., & Büyüköztürk, O. (2017). Deep learning-based crack damage detection using convolutional neural networks. *Computer-Aided Civil and Infrastructure Engineering*, *32*(5), 361–378.

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4960–4964). IEEE.

Caramazza, A., McCloskey, M., & Green, B. (1981). Naive beliefs in "sophisticated" subjects: misconceptions about trajectories of objects. *Cognition*, *9*(2), 117–123. https://doi.org/10.1016/0010-0277(81)90007-x

Chandrasekaran, A., Yadav, D., Chattopadhyay, P., Prabhu, V., & Parikh, D. (2017). It takes two to tango: Towards theory of AI's mind. *arXiv preprint arXiv:1704.00717*.

Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences*, *4*(5), 170–178.

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, *36*(1), 28–71.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One*, *8*(3), e57410.

Dujmović, M., Malhotra, G., & Bowers, J. S. (2020). What do adversarial images tell us about human vision? *eLife*, *9*, e55978.

Eckstein, M. P., Koehler, K., Welbourne, L. E., & Akbas, E. (2017). Humans, but not deep neural networks, often miss giant targets in scenes. *Current Biology*, *27*(18), 2827–2832.e3.

Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018). Adversarial examples that fool both computer vision and time-limited humans. In *Advances in Neural Information Processing Systems* (pp. 3910–3920).

Feather, J., Durango, A., Gonzalez, R., & McDermott, J. (2019). Metamers of neural networks reveal divergence from human perceptual systems. In *Advances in Neural Information Processing Systems* (pp. 10078–10089).

Firestone, C. (2020). Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, *117*(43), 26562–26571.

Firestone, C., & Scholl, B. J. (2016a). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences*, 39, *e229*.

Firestone, C., & Scholl, B. J. (2016b). Seeing and thinking: Foundational issues and empirical horizons. *Behavioral and Brain Sciences*, *e264*.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). Imagenet-trained CNNS are biased towards texture; Increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations, ICLR 2018*.

Golan, T., Raju, P. C., & Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, *117*(47), 29330–29337.

Goodfellow, I., Papernot, N., Huang, S., Duan, R., Abbeel, P., & Clark, J. (2017). Attacking machine learning with adversarial examples. *OpenAI Blog*.

Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, *119*(1), e2110013119.

Harding, S. M., Rajivan, P., Bertenthal, B. I., & Gonzalez, C. (2018). Human decisions on targeted and non-targeted adversarial samples. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*.

Hendrycks, D., & Gimpel, K. (2016). Visible progress on adversarial images and a new saliency map. *CoRR*, *abs/1608.00530*.

Hermann, K. (2022). *Understanding feature use divergences between human and machine vision*. PhD thesis, Stanford University.

Hutson, M. (2018). Hackers easily fool artificial intelligences. *Science*, *361*(6399), 215–215.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems* (pp. 125–136).

Irino, T., & Patterson, R. D. (1996). Temporal asymmetry in the auditory system. *Journal of the Acoustical Society of America*, *99*(4), 2316–2331.

Jacob, G., Pramod, R., Katti, H., & Arun, S. (2019). Do deep neural networks see the way we do? *bioRxiv*, 860759.

Kaplan, D. (2008). An overview of Markov chain methods for the study of stage-sequential developmental processes. *Developmental Psychology*, *44*(2), 457–467.

Kouider, S., & Dehaene, S. (2007). Levels of processing during non-conscious perception: A critical review of visual masking. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 857–875.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097–1105).

Lakhani, P., & Sundaram, B. (2017). Deep learning at chest radiography. *Radiology*, *284*, 574–582.

Lamere, P., Kwok, P., Gouvea, E., Raj, B., Singh, R., Walker, W., Warmuth, M., & Wolf, P. (2003). The CMU SPHINX-4 speech recognition system. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong*, volume *1* (pp. 2–5).

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Psychology*, *3*(3), 299–321.

Mack, A. (2003). Inattentional blindness: Looking without seeing. *Current Directions in Psychological Science*, *12*(5), 180–184.

McCoy, T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3428–3448).

Miller, G. A. (1952). Finite Markov processes in psychology. *Psychometrika*, *17*(2), 149–167.

Müller, N. M., Markert, K., & Böttinger, K. (2021). Human perception of audio deepfakes. *arXiv preprint arXiv:2107.09667*.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Computer Vision and Pattern Recognition* (pp. 427–436).

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163.

Phillips, I. (2018). Unconscious perception reconsidered. *Analytic Philosophy*, *59*(4), 471–514.

Phillips, I. (2021). Blindsight is qualitatively degraded conscious vision. *Psychological Review*, *128*(3), 558–584.

Qin, Y., Carlini, N., Cottrell, G., Goodfellow, I., & Raffel, C. (2019). Imperceptible, robust, and targeted adversarial examples for automatic speech recognition, 5231–5240.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, *38*, 7255–7269.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, *212*(4497), 947–949.

Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study, 2940–2949.

Sabour, S., Cao, Y., Faghri, F., & Fleet, D. J. (2015). Adversarial manipulation of deep representations. *arXiv*.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2018). Brain-Score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*.

Serre, T. (2019). Deep Learning: The good, the bad, and the ugly. *Annual Review of Vision Science*, *5*(1), 399–426.

Smith, K. A., Battaglia, P. W., & Vul, E. (2018). Different physical intuitions exist between tasks, not domains. *Computational Brain & Behavior*, *1*(2), 101–118.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2818–2826).

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*.

Taylor, A. H., Bastos, A. P. M., Brown, R. L., & Allen, C. (2022). The signature-testing approach to mapping biological and artificial intelligences. *Trends in Cognitive Sciences*, *26*(9), 738–750. https://doi.org/10.1016/j.tics.2022.06.002

Vadillo, J., & Santana, R. (2020). On the human evaluation of audio adversarial examples. *arXiv preprint arXiv:2001.08444*.

Vinson, D. W., Abney, D. H., Amso, D., Chemero, A., Cutting, J. E., Dale, R., & Spivey, M. (2016). Perception, as you make it. *Behavioral and Brain Sciences*, 39, *e260*.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647.

Ward, E. J. (2019). Exploring perceptual illusions in deep neural networks. *Journal of Vision*, *19*(10), 34b.

Weiskrantz, L. (1986). *Blindsight: A case study and implications*. Oxford University Press.

Weiskrantz, L. (1996). Blindsight revisited. *Current Opinion in Neurobiology*, *6*(2), 215–220.

Wu, C. C., & Wolfe, J. M. (2018). A new multiple object awareness paradigm shows that imperfect knowledge of object location is still knowledge. *Current Biology*, *28*(21), 3430–3434.e3.

Yuan, L., Xiao, W., Kreiman, G., Tay, F. E., Feng, J., & Livingstone, M. S. (2020). Adversarial images for the primate brain. *arXiv preprint arXiv:2011.05623*.

Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., & Xu, W. (2017). Dolphinattack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (pp. 103–117). New York: ACM Press.

Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. *Nature Communications*, *10*(1), 1–9.

Zoran, D., Isola, P., Krishnan, D., & Freeman, W. T. (2015). Learning ordinal relationships for mid-level vision. In *The IEEE International Conference on Computer Vision (ICCV)* (pp. 388–396).

APPENDIX A: Replicating Experiment 1 with original adversarial speech clips

All of our experiments generated new stimuli from scratch using the white-box method described by Carlini et al. (2016). However, our methods apply equally well to other stimuli generated using this method, including those presented in the original work. In a preregistered supplementary study, we adopted the same procedure from Experiment 1, except we use the five stimuli provided at https://www.hiddenvoicecommands.com/white-box. We converted those files to .wav format, trimmed them to have a shorter period of silence following the command, amplified them, and then played them either forwards or backwards. Hundred subjects (before exclusions) were recruited from Prolific (see Peer, Brandimarte, Samat, & Acquisti, 2017). We included one catch trial of the same form as Experiment 1. Eighty-three subjects were included in the analysis after exclusions. The mean accuracy on experimental trials was 68.9%, which significantly differed from chance accuracy (50%), $t(82) = 8.20$, $p < .001$, $d = 0.906$. These results demonstrate that subjects can also reliably determine whether the audio attacks presented in Carlini et al. (2016) contain speech, consistent with our findings in Experiment 1.