nature research

Peer Review Information

Journal: Nature Human Behaviour
Manuscript Title: Deep neural network models of sound localization reveal how perception is adapted to real-world environments
Corresponding author name(s): Andrew Francl and Josh H. McDermott

Reviewer Comments & Decisions:

Decision Letter, initial version:
29th September 2020

Dear Dr McDermott,

Thank you once again for your manuscript, entitled "Deep neural network models of sound localization reveal how perception is adapted to real-world environments", and for your patience during the peer review process.

Your Article has now been evaluated by 3 referees. You will see from their comments copied below that, although they find your work of [considerable] potential interest, they have raised quite substantial concerns. In light of these comments, we cannot accept the manuscript for publication, but would be interested in considering a revised version if you are willing and able to fully address reviewer and editorial concerns.

We hope you will find the referees' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

All reviewers agree that your work presents solid engineering work. Reviewers 1 and 3 however point out that the objective of the current work is not entirely clear and also the relevance of your work within the context of human behaviour needs to be better specified. We agree with these comments and would like you to carefully adddress them in your revised manuscript.

Reviewer 2 points out the lack of empirical evidence to validate predictions of your model. Furthermore, Reviewer 3 notes that your model has been validated in extremely limited task settings.

COVID-19 has dramatically limited opportunities for laboratory-based research and we will not insist on the provision of novel empirical evidence to validate predictions. However, we expect that you will fully address Reviewer 3's requests for further validation leveraging existing datasets.

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. I have also attached a template manuscript file that exemplifies our policies and formatting requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

If you wish to submit a suitably revised manuscript we would hope to receive it within 6 months. We understand that the COVID-19 pandemic is causing significant disruptions which may prevent you from carrying out the additional work required for resubmission of your manuscript within this timeframe. If you are unable to submit your revised manuscript within 6 months, please let us know. We will be happy to extend the submission date to enable you to complete your work on the revision.

With your revision, please:

• Include a "Response to the editors and reviewers" document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be used by the editors to evaluate your revision and sent back to the reviewers along with the revised manuscript.

• Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

Thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further.

Sincerely,

Samantha Antusch
Editor
Nature Human Behaviour

Reviewer expertise:

Reviewer #1: human sound localization/spatial hearing and formal models of sound localization/spatial hearing

Reviewer #2: machine learning/deep neural networks (with interest in sensory perception/audition), engineering/machine models of hearing

Reviewer #3: machine learning/deep neural networks (with interest in sensory perception/audition), engineering/machine models of hearing

REVIEWER COMMENTS:

Reviewer #1:
Remarks to the Author:
This manuscript presents a newly developed deep neural network (DNN) trained to mimic some properties of the human spatial hearing. The evaluation is extensive and covers many aspects of spatial hearing, even beyond the classical sound-localization performance. The writing is clear and detailed, the figures are well-laid out and descriptive.

While I appreciate the work per se, I wonder what the gain in the general knowledge from this work is. If this manuscript were submitted to one of the IEEE journals, the goal would be clear: to present a new "black box", that functionally mimics some properties of the human auditory system. The focus would be engineering and the proposed DNN would find many applications. In Nature Human Behavior, I rather expect to gain some knowledge – in the case of the submitted manuscript, it could be something new about the properties of the spatial hearing, or about the correspondence of some model stages to the human auditory neural system.

Given that, I was seeking for new insights from the proposed work. The proposed "black box" uses a rather simple linear cochlear pre-processing, which is not state of the art but it is commonly used in the field of audio engineering. For example, it misses the cochlear nonlinear properties or the contribution of efferents – that's valid, but nothing new. The general DNN design (convolution→ ReLU → batch normalization → next layer) is widely used in many areas of machine learning. The choice of the best DNN configurations was based on testing over 1500 architectures and selecting the 10 best-performing ones. This does not seem to be a hypothesis-driven approach, it rather reminds me of the brute-force approaches usually used to find a solution in a stochastic way such as Monte Carlo. The combination of the cochlear processing and the DNNs is clearly a solid and great engineering work, but from the scientific point of view, I don't see much new insight here.

One typical advantage of having a predictor of human behavior is that it can be applied to predict data in conditions not testable in humans. In the manuscript, the authors write that this corresponds to training the DNN to unnatural conditions. Three "unnatural" conditions were tested:
#1: Anechoic condition. Trained to anechoic sounds, the DNN predictions failed in the precedence-effect task. This is actually trivial: no reflections, no precedence effect.
#2: Noiseless condition. Here, the predictions failed in the bandwidth task. This is again not surprising given that the system does not implement internal noise in the cochlea, meaning that despite cochlear bandpass filtering, each channel transmits information of all frequencies down to the numerical accuracy. That's why state-of-the-art cochlear simulators implement internal noise corresponding to

the human absolute hearing thresholds, see the already quite vintage Breebaart et al. (2001) and compare to the more recent comparison in Saremi et al. (2016).
#3: Narrowband-sounds. Here, predictions failed in conditions simulating listening with others' ears. This is not surprising at all because these conditions do require wideband stimuli (up to 16 kHz). The corresponding references for that knowledge go back in time for decades, e.g., Middlebrooks (1992), so again, this is not a new insight, it is rather an important aspect in the evaluation of the proposed box.
Taken together, no surprises here, no gain in knowledge – just a well-trained DNN, being well-evaluated.

Hence, I suggest to submit this work to an engineering journal – I'm sure that it will find application by others.

Also, note that without having the DNN (code and training data) published, the research cannot be reproduced at all. Reproducibility is one of the main criteria for scientific quality, and in case of software it is easy to provide. Thus, even if resubmitted to any journal, I strongly recommend to publish the DNN (code and data for the training as well as the trained network). An article describing DNNs without the published working code is not much of use.

References:

Saremi, A., Beutelmann, R., Dietz, M., Ashida, G., Kretzberg, J., and Verhulst, S. (2016). "A comparative study of seven human cochlear filter models," The Journal of the Acoustical Society of America, 140, 1618–1634.

Breebaart, J., van de Par, S., and Kohlrausch, A. (2001). "Binaural processing model based on contralateral inhibition I Model structure," J Acoust Soc Am, 110, 1074–1088.

Middlebrooks, J. C. (1992). "Narrow-band sound localization related to external ear acoustics," J Acoust Soc Am, 92, 2607–2624.

Reviewer #2:
Remarks to the Author:
This manuscript presents a deep learning model of sound localization, with human-like inputs (including impulse-response functions from human ears) and trained on simulated natural sound environments. When tested on various human psychophysical benchmarks of sound localization, it displays qualitative effects similar to human listeners. This qualitative match is reduced when the model is trained on unnatural environments (e.g. no reverberation), or without the human-ear IRF.
As far as I can tell, this is the first model of sound localization trained end-to-end, directly from (simulated) sound sources. The dataset generation procedure and model architecture search are colossal computational feats, and the fact that some of the results are made available to the community (model checkpoint and training code on github) will be invaluable. I would encourage the authors to also share the training data on a public repository.
The approach and the model presented here are informative in many respects. By training variants of the model with or without certain properties (of the architecture and/or of the dataset), we can learn

about the origin and functional role of many idiosyncratic properties of the human auditory system (e.g. precedence effect is an adaptation response to echoic reverberation). We can run "thought experiments" about how audition may have evolved in alternate worlds. On the practical side, the model can help determine how to improve sound stimuli for optimal localization.
One important contribution of such a model would be the ability to make novel predictions about human auditory localization performance under various conditions, and to verify these predictions experimentally. The authors have initiated this strategy, mapping the model-estimated quality of localization behavior for many different musical instruments. Unfortunately, the latter part of the strategy, verifying these predictions experimentally, could not be performed because of COVID19. Instead, the authors report anecdotal evidence that the predictions might hold, and defer the actual experiment to a later study. I have no clear opinion on this decision, which reflects an exceptional situation in which standard guidelines and criteria cannot be blindly applied—I merely wish to draw the Editor's attention to this issue.
In short, I am extremely positive about this submission, and believe it will have a strong impact on and beyond our community. I have only minor suggestions for improvement.
-figure 7: I am curious to see how human-model similarity compares with between-human similarity. I wonder if there is a way to estimate this, e.g. using standard deviation or standard error measures from each experiment.
-line 208: "more simpler" => "simpler"
-line 488: "randomly selection locations" => "randomly selected locations"

Reviewer #3:
Remarks to the Author:
This manuscript proposed a neural network model for binaural sound localization and compared its performance against human psychoacoustic results at the behavioral level. It is very well written, and covers a large number of literatures and experiments.

The major concern the reviewer has is about the objective of this study. Did the authors intend to make a good machine, or do
they intend to analyze the mechanism of the human hearing? The purpose of the research is somewhat vague to readers.

Also although the proposed neural network has shown able to model human localisation behaviors in many psychoacoustic experiments, it is less clear how much is task-dependant. Would be interesting to see if the behaviors can still be replicated if a different sound localization task is used, e.g. vertical localization, or one that involves front-back confusions.

The reviewer would like the authors to address the following comments before making a recommendation:

1/ Lines 59-64: The authors used the "duplex" theory as an example to support the importance of behavior models. But in fact it has been proved to be inaccurate for decades, long before any established behavior models were proposed.

2/ Line 98: The cochlea filter output was downsampled to 4 kHz - this effectively removes all the

components above 2kHz. I understand phase locking is only present for low frequencies for most species and high frequency sounds cannot be localized by phase differences. However, I think the upper frequency limit of phasing locking varies across species and remains unclear in humans [Verschooten et al. (2019) "The upper frequency limit for the use of phase locking to code temporal fine structure in humans: A compilation of viewpoints", Hearing Research]. The cited study by Palmer and Russel measured guinea pigs whose upper limit we know is below 2kHz. I think phase locking is still present for low frequency tones up to 4kHz? Furthermore, high frequencies are still useful for localization and provide intensity cues and spectral cues due to the shape and pinnae and head. By removing frequency components above 2kHz, the neural networks used in this study did not get to learn any cues in the high frequencies, which are available and used by human listeners. Please elaborate.

Is the downsampling related to reducing the dimensionality of the input vector for the neural networks?

3/ Line 112: what is the length of each stimulus sample here? Is it 2-sec or 10 ms which is typically used by many machine sound localization systems? Were overlapping windows used to frame the signals? What were the frame shifting rate and frame length?

4/ Line 122: please specify how many data samples were used per spatial location to train the neural networks. And what materials were used as natural sound sources?

5/ Lines 124-125: please clarify if the reverberation was simulated in a binaural setting or a monaural setting. If the latter, as indicated by line 125 that the direction of reflections was created by convolving the monaural reflection with HRTFs, then this is perhaps very artificial in that source reflection is identical in all directions. This is not the case if binaural room impulse responses are measured.

6/ Line 128: please clarify how many randomly chosen locations were used for background noise and why. Could the choice have an impact on the results? What about diffuse noise?

7/ Line 131: How were the neural network outputs mapped to the location? Were the azimuth/elevation pair labels used in a softmax fashion or a regression fashion? What are the range of azimuth and elevation? How many labels are there?

8/ Line 192: the reviewer is surprised to see that the authors only compared their system to microphone-array localization systems. There are a large number of studies in binaural sound localization systems that exploit head/torso related transfer functions in both anechoic and reverberant conditions. To name a few:
* May et al. (2011) "A probabilistic model for robust localization based on a binaural auditory front-end," IEEE TASLP
* Woodruff and Wang (2012) "Binaural localization of multiple sources in reverberant and noisy environments," IEEE TASLP
* Ma et al. (2017) "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments", IEEE TASLP
* Vecchiotti et al. (2019) End-to-end binaural sound localisation from the raw waveform, IEEE ICASSP
* There is also a chapter on "Binaural Sound Localization" by R Stern in the book Computational

Auditory Scene Analysis edited By DeLiang Wang and Guy Brown

9/ Please clarify how human listeners' responses were recorded - were the humans able to see (or informed of) the 11 loudspeakers? How long were the stimuli?

10/ Figure 4B: what is the Y-axis unit? Although the proposed model shows a V-shape result pattern across azimuth, the localization errors at the lateral positions seem very large (30-40 degrees) compared to human data.

11/ While it is interesting to see the proposed model is able to replicate many psychoacoustic experiments, all the experiments are focused on single-source localization and the authors did not include any multi-source localization experiments. In realistic listening conditions there are often multiple sources present and indeed this is the setting where the authors trained the neural networks. There are some studies in human sound localization in such settings, e.g.
* N. Kopco, V. Best, and S. Carlile, "Speech localization in a multitalker mixture," J. Acoust. Soc. Amer., vol. 127, no. 3, pp. 1450--1457, 2010.

12/ Lines 401-403: The reviewer finds it difficult to accept this explanation. The neural networks are very easy to be trained to work well in a match condition. When compared to a generative model, neural networks are prone to overfitting and often fail to generalize to unseen conditions. Thus it is important to include "noise" during training of neural networks. The increased dissimilarity is probably due to the mismatch between training and testing conditions, rather than "human-like spatial hearing emerged from task optimization only for naturalistic training conditions". The reviewer suspects if the human listeners did the tests in an anechoic room, and anechoic test data was used for the neural network trained in the anechoic condition, the similarity between humans / machines is properly still there.

Also, many studies (e.g. May et al. (2011), Woodruff and Wang (2012), Ma et al. (2017)) have shown that by including white noise during training in the anechoic condition will significantly increase the robustness of a machine localization system to reververantions.

13/ It is also a shame that the authors did not include any experiments in elevation localization, despite that the neural network was trained to do so. It is understandable that COVID-19 creates a huge challenge to run more listening tests, but perhaps the authors could verify the machine performance against the previous human listening results?

14/ Although the authors have mentioned the front-back confusion and the head movements during human sound localization, this is not examined further. There are classical studies examining this aspect, e.g.:
H. Wallach (1940) "The role of head movements and vestibular and visual cues in sound localization," Journal of Experimental Psychology, vol. 27, no. 4

Author Rebuttal to Initial comments

Francl & McDermott – Response to Reviews

Editor Comments:

All reviewers agree that your work presents solid engineering work. Reviewers 1 and 3 however point out that the objective of the current work is not entirely clear and also the relevance of your work within the context of human behaviour needs to be better specified. We agree with these comments and would like you to carefully adddress them in your revised manuscript.

> ***These were very useful comments and we took them to heart. We have extensively revised the manuscript to highlight the understanding that we gained from our approach and to make it clear that the goal is to use techniques from engineering to understand human behavior. We added a new panel to Figure 1 to explicitly illustrate the approach and its relation to human behavior. We also moved the comparisons to existing engineering methods to a supplementary figure, to avoid the impression that the primary objective was to build a system that beat state-of-the-art methods in engineering.***

Reviewer 2 points out the lack of empirical evidence to validate predictions of your model. Furthermore, Reviewer 3 notes that your model has been validated in extremely limited task settings. COVID-19 has dramatically limited opportunities for laboratory-based research and we will not insist on the provision of novel empirical evidence to validate predictions. However, we expect that you will fully address Reviewer 3's requests for further validation leveraging existing datasets.

> ***We have further validated the model predictions with multiple additional experiments, explaining new aspects of elevation perception (Figure 4M-O) as well as localization of multiple sources at once (Figure 6), as detailed below. The correspondence between model and human results for these new experiments is striking. Together the results make a strong case for the validity of the model predictions.***

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. I have also attached a template manuscript file that exemplifies our policies and formatting requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

> ***We have complied with all the formatting requirements.***

Please Note that all line numbers provided below correspond to the PDF for Review that has the figures embedded for ease of reading.

REVIEWER COMMENTS:

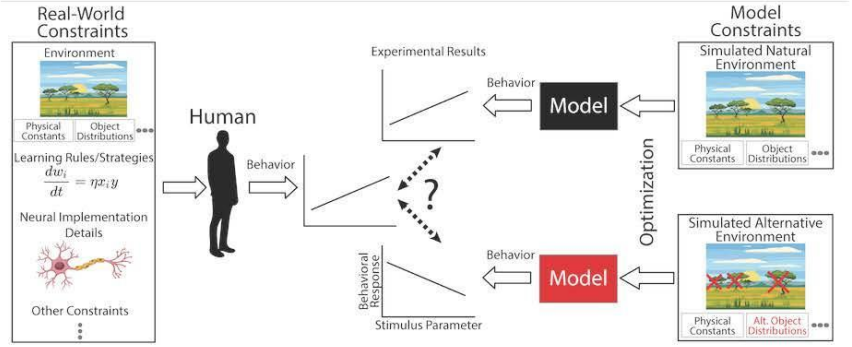
Reviewer #1:

This manuscript presents a newly developed deep neural network (DNN) trained to mimic some properties of the human spatial hearing. The evaluation is extensive and covers many aspects of spatial hearing, even beyond the classical sound-localization performance. The writing is clear and detailed, the figures are well-laid out and descriptive.

While I appreciate the work per se, I wonder what the gain in the general knowledge from this work is. If this manuscript were submitted to one of the IEEE journals, the goal would be clear: to present a new "black box", that functionally mimics some properties of the human auditory system. The focus would be engineering and the proposed DNN would find many applications. In Nature Human Behavior, I rather expect to gain some knowledge – in the case of the submitted manuscript, it could be something new about the properties of the spatial hearing, or about the correspondence of some model stages to the human auditory neural system.

Given that, I was seeking for new insights from the proposed work. The proposed "black box" uses a rather simple linear cochlear pre-processing, which is not state of the art but it is commonly used in the field of audio engineering. For example, it misses the cochlear nonlinear properties or the contribution of efferents – that's valid, but nothing new. The general DNN design (convolution→ ReLU → batch normalization → next layer) is widely used in many areas of machine learning. The choice of the best DNN configurations was based on testing over 1500 architectures and selecting the 10 best-performing ones. This does not seem to be a hypothesis-driven approach, it rather reminds me of the brute-force approaches usually used to find a solution in a stochastic way such as Monte Carlo. The combination of the cochlear processing and the DNNs is clearly a solid and great engineering work, but from the scientific point of view, I don't see much new insight here.

> ***Thank you for the constructive feedback. Your comments forced us to sharpen and clarify our message and argument, and we hope the revision makes the relevance to human behavior clearer.***
>
> ***To highlight the scientific approach up front, we added a graphical description of the overall method:***



"Figure 1. Overview of approach. A. Illustration of general method. A variety of constraints (left) shape human behavior. Models optimized under particular environmental constraints (right) can illustrate the effect of these constraints on behavior. Environment simulators can be used to instantiate naturalistic environments as well as alternative environments in which particular properties of the world are altered, to examine the constraints that shape human behavior." (lines 165-169)

We have also made extensive edits to the Introduction to help clarify the scientific goal of the work:

"Here we extend ideas from ideal observer theory to investigate the environmental constraints under which human behavior emerges, using contemporary machine learning to optimize models for behaviorally relevant tasks in simulated environments. Human behaviors that emerge from machine learning under a set of naturalistic environmental constraints can be understood as a consequence of optimization for those constraints (Fig. 1A)." (lines 43-48)

"…we aim to use the neural network as a way to find an optimized solution to a difficult real-world task that is not easily specified analytically, for the purpose of comparing its behavioral characteristics to those of humans." (lines 69-71)

"The approach we employ is broadly applicable to other sensory modalities, providing a way to test the adaptedness of aspects of human perception to the environment and to understand the conditions in which human-like perception arises." (lines 88-90)

You are correct to note that our engineering methods are state-of-the-art, but not novel per se. This was by design. The novelty is in

the application of these engineering methods to understanding behavior, and we sought to use comparatively well-understood technical ingredients to this end. We have clarified this in the revised paper:

At the start of the Results section:
"The output of the two cochlea formed the input to a standard convolutional neural network (Figure 1C)." (lines 105)

And in the Methods section:
"The components of the CNNs were standard; they were chosen because they have been shown to be effective in a wide range of sensory classification tasks." (lines 856-858)

We have also clarified the motivation for the various choices that were made in building the model, including the choice of cochlear pre-processing:

"The cochlear model was chosen to approximate the time and frequency information in the human cochlea subject to practical constraints on the memory footprint of the model and the dataset. Cochleagrams were generated using a filter bank like that in previous work from our lab[53]. However, the cochleagrams we used provided fine timing information to the network by passing rectified subbands of the signal instead of the envelopes of the subbands. This came at the cost of substantially increasing the dimensionality of the input relative to an envelope-based cochleagram. The dimensionality was nonetheless considerably lower than what would have resulted from a spiking model of the auditory nerve, which would have been prohibitive given our hardware." (lines 798-805)

We have also moved the comparison to other two-microphone localization algorithms to a supplementary figure, as featuring this in a main figure early in the paper had the potential to give the wrong impression as to the paper's focus.

One typical advantage of having a predictor of human behavior is that it can be applied to predict data in conditions not testable in humans. In the manuscript, the authors write that this corresponds to training the DNN to unnatural conditions.

We think it is important to distinguish two ways in which model predictions might be used. The first is to serve as an efficient means of estimating what a human would hear in some set of conditions that a person might encounter, by TESTING the model in those conditions. Model predictions are quick and cheap once the model is

trained, and are much more efficient than running experiments on human listeners, particularly in cases where large numbers of stimuli might need to be tested, or in conditions that are difficult to replicate in the lab. This is the way in which the model was used to evaluate the accuracy of musical instrument localization (Figure 8). These provide non-trivial predictions that could be practically useful, for instance in designing auditory displays, and that would not be possible without the model.

The second application of model predictions is to TRAIN the model in particular conditions, to yield a system that is optimized for those conditions. The model can then be TESTED in standard experimental conditions to gain insight into whether the characteristics of human hearing measured in those experimental conditions reflect optimization for particular environmental conditions. This model application can help us understand how human behavior is shaped by aspects of the natural environment. Such understanding is a widespread goal of cognitive science and biology more generally, and our approach offers a new way of achieving it. We suspect from the reviewer's comments that this use of the modeling approach was not clearly communicated in the text, in part by our use the term "experiment" in both cases.

We have clarified these two uses in the revised manuscript, which we think were not sufficiently explained/highlighted in the original text. In particular, in the Results section we now use the term "experiment" exclusively to refer to laboratory experiments simulated on a trained model.

In the Introduction:

"When tested on stimuli from classic laboratory experiments, the resulting model replicated a large and diverse array of human behavioral characteristics. We then trained models in unnatural conditions to simulate evolution and development in alternative worlds. These alternative models deviated notably from human-like hearing." (lines 84-86)

In the Results:

"To assess whether the trained networks replicated the characteristics of human sound localization, we simulated a large set of behavioral experiments from the literature, intended to span many of the best-known and largest effects in spatial hearing." (lines 194-196)

"To assess the extent to which the properties of biological spatial hearing are adapted to the constraints of localization in natural environments, we took advantage of the ability to optimize models in virtual worlds altered in various ways, intended to simulate the optimization that would occur over evolution and/or development in alternative environments (Fig. 1A)." (lines 429-432)

In the Discussion:

"The general method involves two nested levels of computational experiments: optimization of a model under particular conditions, followed by a suite of psychophysical experiments to characterize the resulting behavioral phenotype. " (lines 568-570)

Three "unnatural" conditions were tested:
#1: Anechoic condition. Trained to anechoic sounds, the DNN predictions failed in the precedence-effect task. This is actually trivial: no reflections, no precedence effect.

We respectfully disagree, and think this could have been otherwise. In fact, other researchers have proposed explanations for the precedence effect that do not involve reflections. We have added mention of this and a citation to one example paper proposing such an alternative explanation, by a respected hearing researcher (Pat Zurek). We note also that this finding was specifically mentioned by Reviewer 2 as one that provided insight.

In the Results:

"The most interpretable example of environment-driven localization strategies is the precedence effect. This effect is often proposed to render localization robust to reflections, but others have argued that its primary function might instead be to eliminate interaural phase ambiguities, independent of reflections[98]." (lines 473-475)

In the Discussion:

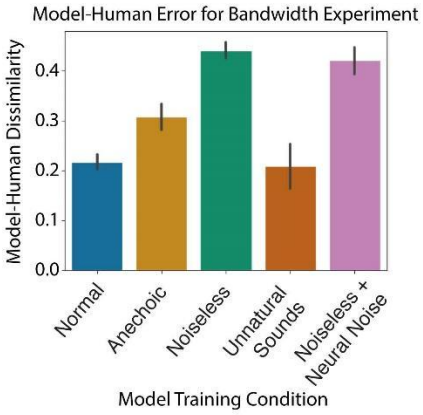
"This approach provides an additional tool with which to examine the constraints that yield biological solutions[103,104], and thus to understand evolution[105]. It also provides a way to link experimental results with function. In some cases these links had been hypothesized but not definitively established. For example, the precedence effect was often proposed to be an adaptation to reverberation[19,91], though other functional explanations were also put

forth[98]. Our results suggest it is indeed an adaptation to reverberation (Fig. 7D)." (lines 570-575)

#2: Noiseless condition. Here, the predictions failed in the bandwidth task. This is again not surprising given that the system does not implement internal noise in the cochlea, meaning that despite cochlear bandpass filtering, each channel transmits information of all frequencies down to the numerical accuracy. That's why state-of-the-art cochlear simulators implement internal noise corresponding to the human absolute hearing thresholds, see the already quite vintage Breebaart et al. (2001) and compare to the more recent comparison in Saremi et al. (2016).

To address this concern we ran a version of the model with internal noise in the cochlear stage akin to that used in the Breebaart et al. paper cited by the reviewer. Specifically, we added independent Gaussian noise to each frequency channel at a level 60.6 dB below the average power across frequency channels (producing noise at 9.4 dB SPL assuming sources at 70 dB SPL – these numbers were taken from Breebaart et al.). The results show that the internal noise has little effect.

Specifically, the results of the bandwidth experiment remain aberrant. This is evident in the human-model dissimilarity for this experiment (here compared to that for the four main training conditions featured in the paper):



It thus appears that the change in the learned strategy is not a result of the presence/absence of neural noise. We have revised the relevant section of the Results to describe this additional result:

"We confirmed that this result was not somehow specific to the absence of internal neural noise in our cochlear model, by training an additional model in which noise was added to each frequency channel (see Methods). We found that the results of training in noiseless environments remained very similar." (lines 463-465)

and have described this control experiment in the Methods section:

"Models with internal noise
To test for the possibility that the noiseless training environments might have had effects that were specific to the lack of internal noise in the cochlea model used as input to our networks, we trained an alternative model with internal noise added to the output of the cochlear stage. This alternative model was identical to the main model used throughout the paper except that independent Gaussian noise was added to each frequency channel prior to the rectification stage of the cochlear model. The noise was sampled from a standard normal distribution and then scaled so that its power was on average 60.6 dB below the average power in the subbands of the input signal (intended to produce noise at 9.4 dB SPL assuming sources at 70 dB SPL[153]). In practice we pre-generated 50,000 noise arrays, sampled one at random on each trial, and added it to the output of the cochlear filters at the desired SNR. " (lines 1542-1551)

#3: Narrowband-sounds. Here, predictions failed in conditions simulating listening with others' ears. This is not surprising at all because these conditions do require wideband stimuli (up to 16 kHz). The corresponding references for that knowledge go back in time for decades, e.g., Middlebrooks (1992), so again, this is not a new insight, it is rather an important aspect in the evaluation of the proposed box.

We agree that the result in this case is not astonishing, but it was also not completely obvious a priori that it would work out in this way. Many of the notches in the HRTFs are narrower than the half-octave bandwidth of the training stimuli in this condition, so it seemed plausible to us that the model might learn to use spectral cues. But independent of whether one regards this particular result as surprising, we feel that it is critically important to be able to validate predictions of this nature rather than rely exclusively on intuition – this is an essential aspect of science – and this is what

our approach enables that is not otherwise possible. In this case, intuitions proved correct, but in others they might not.

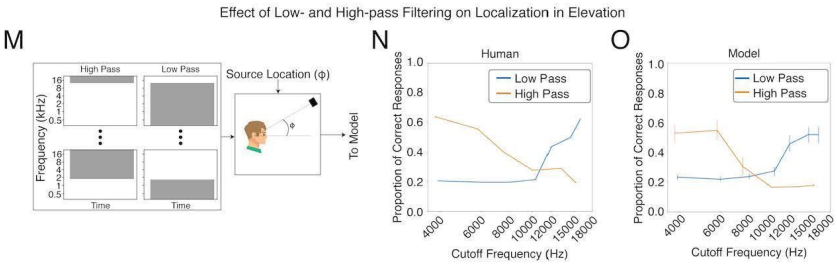
We have revised the discussion to make this point more clearly:

"This approach provides an additional tool with which to examine the constraints that yield biological solutions[103,104], and thus to understand evolution[105]. It also provides a way to link experimental results with function. In some cases these links had been hypothesized but not definitively established. … We similarly provide evidence that sensitivity to spectral cues to elevation emerges only with the demands of localizing somewhat broadband sounds[106]." (lines 570-576)

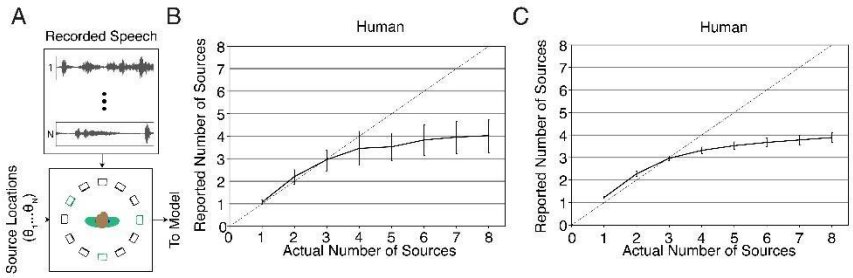
Taken together, no surprises here, no gain in knowledge – just a well-trained DNN, being well-evaluated.

We believe there may have been a misunderstanding about the purpose of the model predictions, and thus the gain in knowledge. We hope the text revisions along with our responses have clarified the contribution.

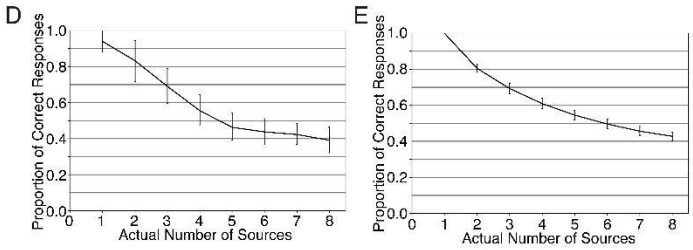
We also note that we have added results for two new experiments, both of which feature what we consider compelling matches to the human data. One experiment measures the effect of highpass and lowpass filtering on the localization in elevation, showing very similar effects in the model and in human listeners:



The other experiment measures localization of multiple concurrent sources. Here again the model reproduces the results quite well. The number of estimated sources shows a very similar dependence on the actual number of sources:



And the dependence of localization accuracy on number of sources is also reproduced well:



Both of these results did not previously have an explanation, and we think it highly nonobvious that multi-source localization should be limited to approximately 4 sources. And yet our model reproduces that effect with a striking degree of accuracy, indicating that this effect reflects limitations of the cues available to listeners.

Here is the revised text explaining this result and its significance:

"Humans are able to localize multiple concurrent sources, but only to a point[95-97]. The reasons for the limits on multi-source localization are unclear[96]. These limitations could reflect human-specific cognitive constraints. For instance, reporting a localized source might require attending to it, which could be limited by central factors not specific to localization. Alternatively, localization could be fundamentally limited by corruption of binaural cues by concurrent sources or other ambiguities intrinsic to the localization problem.

To assess whether the model would exhibit limitations like those observed in humans, we replicated an experiment[97] in which humans

judged both the number and location of a set of speech signals played from a subset of an array of speakers (Fig. 6A). To enable the model to report multiple sources we fine-tuned the final fully-connected layer (freezing all weights in earlier layers) to indicate the probability of a source at each of the location bins, and set a probability criterion above which we considered the model to report a sound at the corresponding location (see Methods). We then tested the model on the experimental stimuli.

Humans accurately report the number of sources up to three, after which they undershoot, only reporting about four sources in total regardless of the actual number (Fig. 6B). The model reproduces this effect, also being limited to approximately four sources (Fig. 6C). Human localization accuracy also systematically drops with the number of sources (Fig. 6D); the model again quantitatively reproduces this effect (Fig. 6E). The model-human similarity suggests that these limits on sound localization are intrinsic to the constraints of the localization problem, rather than reflecting human-specific limitations." (lines 387-407)

We hope these additions help further illustrate the contribution of the work, by helping to explain the origins of some otherwise unexplained behavioral effects.

Hence, I suggest to submit this work to an engineering journal – I'm sure that it will find application by others.

Also, note that without having the DNN (code and training data) published, the research cannot be reproduced at all. Reproducibility is one of the main criteria for scientific quality, and in case of software it is easy to provide. Thus, even if resubmitted to any journal, I strongly recommend to publish the DNN (code and data for the training as well as the trained network). An article describing DNNs without the published working code is not much of use.

We completely agree and note that all code and trained models have been available on GitHub since our initial submission. We provided a link to the GitHub page in the Methods, in the Data and Code Availability Statement section as per the journal formatting guidelines:

"Data and Code Availability Statement
Code and data used to train and analyze the model in this paper, as well as the weights of the trained networks in the model are available at: www.github.com/afrancl/BinauralLocalizationCNN" (lines 641-643)

In addition, we have now made all training data available via the GitHub link.

References:

Saremi, A., Beutelmann, R., Dietz, M., Ashida, G., Kretzberg, J., and Verhulst, S. (2016). "A comparative study of seven human cochlear filter models," The Journal of the Acoustical Society of America, 140, 1618–1634.

Breebaart, J., van de Par, S., and Kohlrausch, A. (2001). "Binaural processing model based on contralateral inhibition I Model structure," J Acoust Soc Am, 110, 1074–1088.

Middlebrooks, J. C. (1992). "Narrow-band sound localization related to external ear acoustics," J Acoust Soc Am, 92, 2607–2624.

Reviewer #2:
Remarks to the Author:
This manuscript presents a deep learning model of sound localization, with human-like inputs (including impulse-response functions from human ears) and trained on simulated natural sound environments. When tested on various human psychophysical benchmarks of sound localization, it displays qualitative effects similar to human listeners. This qualitative match is reduced when the model is trained on unnatural environments (e.g. no reverberation), or without the human-ear IRF.
As far as I can tell, this is the first model of sound localization trained end-to-end, directly from (simulated) sound sources. The dataset generation procedure and model architecture search are colossal computational feats, and the fact that some of the results are made available to the community (model checkpoint and training code on github) will be invaluable. I would encourage the authors to also share the training data on a public repository.

We have uploaded the training data to a cloud server, and now provide the link via the GitHub site linked to in the paper:

"Data and Code Availability Statement
Code and data used to train and analyze the model in this paper, as well as the weights of the trained networks in the model are available at: www.github.com/afrancl/BinauralLocalizationCNN" (lines 641-643)

The approach and the model presented here are informative in many respects. By training variants of the model with or without certain properties (of the architecture and/or of the dataset), we can learn about the origin and functional role of many idiosyncratic properties of the human auditory system (e.g. precedence effect is an adaptation response to echoic reverberation). We can run "thought experiments" about how audition may have evolved in alternate worlds. On the practical side, the model can help determine how to improve sound stimuli for optimal localization.
One important contribution of such a model would be the ability to make novel predictions about human auditory localization performance under various conditions, and to verify these predictions experimentally. The authors have initiated this strategy, mapping the model-estimated quality of localization behavior for many different musical instruments. Unfortunately, the latter part of the strategy, verifying these predictions experimentally, could not be performed because of COVID19. Instead, the authors report anecdotal evidence that the predictions might hold, and defer the actual experiment to a later study. I have no clear opinion on this decision, which reflects an exceptional situation in which standard guidelines and criteria cannot be blindly applied—I merely wish to draw the Editor's attention to this issue.
In short, I am extremely positive about this submission, and believe it will have a strong impact on and beyond our community. I have only minor suggestions for improvement.

Thank you.

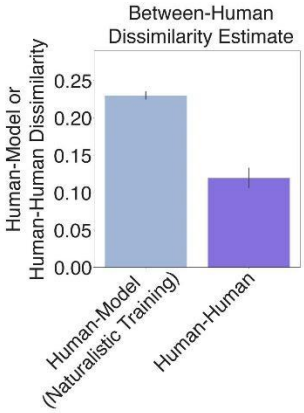
-figure 7: I am curious to see how human-model similarity compares with between-human similarity. I wonder if there is a way to estimate this, e.g. using standard deviation or standard error measures from each experiment.

We would love to be able to do this, as some of the human-model discrepancies are undoubtedly due to the imperfect reliability of the human experimental results. Unfortunately, some of the experiments are from decades ago, and error bars were not provided in the original publications. We were thus unable to estimate the full aggregate dissimilarity that one would expect between two groups of human subjects.

Instead, we computed such an estimate for the subset of experiments for which error bars were provided. We assumed that the human experiment results for each experimental condition were Gaussian distributed with a standard deviation equal to the standard error of the mean (which we estimated from the scanned error bars), and that each condition was independent of the others. We then randomly sampled data points from these Gaussians (simulating a replication of the experiment), and computed the root-mean-square

error between these sampled points and the actual human data for an experiment. This dissimilarity measure was then averaged across the 5 experiments for which we had error bars to work with. We repeated this process 10,000 times, and used the mean and 95% confidence intervals from the resulting dissimilarity distribution to estimate the human-human dissimilarity.

The results show that the human-human dissimilarity is substantial, though not quite as high as the human-model dissimilarity for this subset of experiments:



We now reference this analysis in the Results section:

"the absolute dissimilarity is not meaningful (in that it is limited by the reliability of the human results, which is not perfect; see Supplementary Figure 4)…" (lines 452-454)

and describe it in detail in the Methods section:

"Between-human dissimilarity
The dissimilarity that would result between different samples of human participants puts a lower bound on model-human dissimilarity, and would thus be useful to compare to the dissimilarity plotted in Figure 7B. This between-human dissimilarity could be estimated using data from the original individual human participants. Unfortunately, the individual participant data was unavailable for nearly all of the experiments that we modeled, many of which were conducted several decades ago. Instead, we used the error bars in the published results figures to simulate different

samples of human participants given the variability observed in the original experiments. Error bars were provided for only some of the original experiments (the exceptions being the experiments in Figures 2 and 4N), so we were only able to estimate the between-human dissimilarity for this subset. We then compared the estimated between-human dissimilarity to the model-human dissimilarity for the same subset of experiments (Supplementary Fig. 4).

We assumed that human data for each experimental condition were independently normally distributed with a mean and variance given by the mean and error bars for that condition. Depending on the experiment, the error bars in the original graphs plotted the standard deviation, the standard error of the mean (SEM), or the 95% confidence interval of the data. In each case we estimated the variance from the mean of the upper and lower error bar (for SD: the square of the error bar; for SEM: $variance = (\sqrt{N} \times SEM)^2$***; for 95% CI:*** $variance = (\sqrt{N} \times (error\ bar\ width)/1.96)^2$ ***, where N is the number of participants). To obtain behavioral data for one simulated human participant, we sampled from the Gaussian distribution for each condition. We sampled data for the number of participants run in the original experiment, and obtained mean results for this set of simulated participants. We then calculated the root-mean-squared error (described in Analysis of Results of Alternative Training Conditions) between the simulated human data and actual human data. We repeated this process 10,000 times for each experiment, yielding a distribution of dissimilarities for each experiment. We then calculated the mean dissimilarity across experiments and samples. Supplementary Figure 4 plots this estimated between-human dissimilarity (with confidence intervals obtained from the distribution of between-human dissimilarity) alongside the model-human dissimilarity for the same subset of experiments." (lines 514-540)***

and provide the figure as a supplementary figure.

-line 208: "more simpler" => "simpler"

Corrected.

-line 488: "randomly selection locations" => "randomly selected locations"

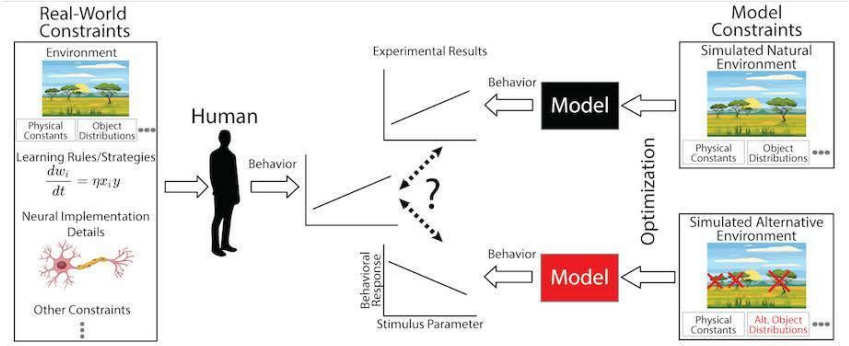
Corrected.

Reviewer #3:
Remarks to the Author:
This manuscript proposed a neural network model for binaural sound localization and compared its performance against human psychoacoustic results at the behavioral level. It is very well written, and covers a large number of literatures and experiments.

Thank you.

The major concern the reviewer has is about the objective of this study. Did the authors intend to make a good machine, or do they intend to analyze the mechanism of the human hearing? The purpose of the research is somewhat vague to readers.

We have extensively revised the introduction to clarify the objective. The purpose of the study was to use a new type of model to get insight into human hearing. However, the modeling approach we employed was to build a machine system that functions under the same constraints as human listeners. We have added a figure depicting the scientific logic of the approach:



"Figure 1. Overview of approach. A. Illustration of general method. A variety of constraints (left) shape human behavior. Models optimized under particular environmental constraints (right) can illustrate the effect of these constraints on behavior. Environment simulators can be used to instantiate naturalistic environments as well as alternative environments in which particular properties of the world are altered, to examine the constraints that shape human behavior." (lines 165-169)

We also believe the scientific objective is clearer in the revised introduction:

"Here we extend ideas from ideal observer theory to investigate the environmental constraints under which human behavior emerges, using contemporary machine learning to optimize models for behaviorally relevant tasks in simulated environments. Human behaviors that emerge from machine learning under a set of naturalistic environmental constraints can be understood as a consequence of optimization for those constraints (Fig. 1A)." (lines 43-48)

"…we aim to use the neural network as a way to find an optimized solution to a difficult real-world task that is not easily specified analytically, for the purpose of comparing its behavioral characteristics to those of humans." (lines 69-71)

"The approach we employ is broadly applicable to other sensory modalities, providing a way to test the adaptedness of aspects of human perception to the environment and to understand the conditions in which human-like perception arises." (lines 88-90)

Also although the proposed neural network has shown able to model human localisation behaviors in many psychoacoustic experiments, it is less clear how much is task-dependant. Would be interesting to see if the behaviors can still be replicated if a different sound localization task is used, e.g. vertical localization, or one that involves front-back confusions.

We note that the original submission included the classic experiment by Hofman et al., which required localization in both vertical and horizontal dimensions, as well as an experiment by Kulkarni and Colburn that reflects elevation perception. However, in the revised paper we added another experiment testing the effects of highpass and lowpass filtering on vertical localization, and again observe a fairly compelling match to human results (Figure 4M-O in the revised manuscript; see below).

We also added an experiment in which multiple sources are localized at once (Figure 6 in the revised manuscript; see below). Both the number of heard sources (which undershoots the true number in human listeners) and the dependence of localization accuracy on the number of sources are closely reproduced by the model (shown below). Neither of these effects had an explanation prior to our

model, and we consider this a fairly compelling modeling success in a different localization task.

The reviewer would like the authors to address the following comments before making a recommendation:

1/ Lines 59-64: The authors used the "duplex" theory as an example to support the importance of behavior models. But in fact it has been proved to be inaccurate for decades, long before any established behavior models were proposed.

The point we were trying to make is that duplex theory is an example of where intuitions can prove incorrect. But we see how this was confusing, and have removed the reference to duplex theory here as part of the revision of the introduction.

2/ Line 98: The cochlea filter output was downsampled to 4 kHz - this effectively removes all the components above 2kHz. I understand phase locking is only present for low frequencies for most species and high frequency sounds cannot be localized by phase differences. However, I think the upper frequency limit of phasing locking varies across species and remains unclear in humans [Verschooten et al. (2019) "The upper frequency limit for the use of phase locking to code temporal fine structure in humans: A compilation of viewpoints", Hearing Research]. The cited study by Palmer and Russel measured guinea pigs whose upper limit we know is below 2kHz. I think phase locking is still present for low frequency tones up to 4kHz?

This was a typo in the first paragraph of the Results section. The Methods section gave the correct description, which is that the cochlear filter output was lowpass-filtered with a cutoff of 4 kHz to match the presumptive upper limit of phase locking (which is typically estimated to be in the neighborhood of 4 kHz), and then downsampled to 8 kHz. We have corrected this in the revised manuscript:

"The cochlea was simulated with a bank of bandpass filters modeled on the frequency selectivity of the human ear[52,53], whose output was rectified and low-pass filtered to simulate the presumed upper limit of phase locking in the auditory nerve[54]." (lines 98-100)

We have also clarified that the output was downsampled to 8 kHz following low-pass filtering with a 4 kHz cutoff:

"The results were then half-wave rectified to simulate the auditory nerve firing rates and were lowpass filtered with a cutoff frequency

of 4 kHz to simulate the upper limit of phase-locking in the auditory nerve[54]... The results of the lowpass filtering were then downsampled to 8 kHz to reduce the dimensionality of the neural network input..." (lines 815-819)

Furthermore, high frequencies are still useful for localization and provide intensity cues and spectral cues due to the shape and pinnae and head. By removing frequency components above 2kHz, the neural networks used in this study did not get to learn any cues in the high frequencies, which are available and used by human listeners. Please elaborate.

We note that these operations apply to the output of each filter (once rectified). So the cochlear representation retains information about high audio frequencies – those are represented by the activation of the high-frequency filters. The nature of the representation is influenced by the simulated phase-locking limit, but because this limit is applied after rectification, the result retains information about the original audio frequencies (just with lower temporal fidelity, in order to more accurately model what is believed to happen in the ear).

We have clarified this issue in the text:

"Because the lowpass filtering and downsampling was applied to rectified filter outputs, the representation retained information at all audible frequencies, just with limits on fidelity that were approximately matched to those believed to be present in the ear." (lines 820-822)

Is the downsampling related to reducing the dimensionality of the input vector for the neural networks?

Yes. Once the input is lowpass-filtered to simulate phase-locking, it can be downsampled without loss of information, and this reduces the memory footprint. We have clarified this in the Methods:

"The results of the lowpass filtering were then downsampled to 8 kHz to reduce the dimensionality of the neural network input (without information loss because the Nyquist limit matched the lowpass filter cutoff frequency). " (lines 818-820)

3/ Line 112: what is the length of each stimulus sample here? Is it 2-sec or 10 ms which is typically used by many machine sound localization systems? Were overlapping windows used to frame the signals? What were the frame shifting rate and frame length?

We have clarified that each training example was one second in length:

"The architecture search consisted of training each of a large set of possible architectures for 15000 training steps with 16 1s stimulus examples per step (240k total examples; see Supplementary Figure 1 for distribution of localization performance across architectures)." (lines 114-116)

The examples were the output of the cochlear model (filter bank → power compression → rectification → lowpass filtering + downsampling), so there were no frames. This has been clarified in the methods section:

"We note also that the input was not divided into "frames" as are common in audio engineering applications, as these do not have an obvious analogue in biological auditory systems." (lines 822-824)

4/ Line 122: please specify how many data samples were used per spatial location to train the neural networks.

This has been clarified:

"The training data was based on a set of ~500,000 stereo audio signals with associated 3D locations relative to the head (on average 988 examples for each of the 504 location bins; see Methods)." (lines 124-125)

And what materials were used as natural sound sources?

We have added a table listing all of the natural sound sources (Supplementary Table 2):

Air hockey	Chainsaw Cutting 2	Doorbell 4	Humming 1	Revving Engine 2	Tapdancing 1
Airplane	Chainsaw Revving	Door knocking	Humming 2	Ringing Phone 1	Tapdancing 2
Alarm 1	Chair Rolling	Drawer opening	Ice Cream Truck	Ringing Phone 2	Tapping Finger
Alarm 2	Cheering	Drilling screw	Insect chirping	Ringing Phone 3	Tapping Object
Alarm 3	Person clapping	Drilling into wood 1	Jackhammer 1	Ringing Phone 4	Tearing
Alarm clock	Chewing 1	Drilling into wood 2	Jackhammer 2	Road traffic	Telephone Ringing
Animal noises 1	Chewing 2	Drinking	Jackpot sound effect	Rocket Launch	Tennis Rally
Animal noises 2	Chicken Clucking	Driving sounds	Jumping rope 1	Rocking Chair	Thunder
Animal noises 3	Chimes 1	Drum Roll	Jumping rope 2	Rooster 1	Ticking Clock
Baby Crying	Chimes 2	Drums Beat	Kettle whistling	Rooster 2	Toothbrushing
Basketball Dribbling 1	Chimes 3	Duck quack 1	Person Laughing 1	Rooster 3	Train 1
Basketball Dribbling 2	Chimes 4	Duck quack 2	Person Laughing 2	Rotary Telephone Dialer	Train 2
Bear	Chopping Wood	Eating	Person Laughing 3	Rubbing Hands	Train 3
Bee 1	Chopping Food	Duck quack 3	Lawn mower 1	Running 1	Trainbell 1
Bee 2	Church Bells	Electric Hand Drill Starting	Lawn mower 2	Running 2	Trainbell 2
Beeping 1	Cicadas 1	Electric Shaver	Lawn mower 3	Running Up Stairs	Trainbell 3
Beeping 2	Cicadas 2	Elevator door	Lion 1	Running water faucet 1	Train Leaving Station
Beeping 3	Clanking	Engine 1	Lion 2	Running water faucet 2	Train Warning Bell
Bells Chiming 1	Clapping 1	Engine 2	Lion 3	Running water faucet 3	Train whistle 1
Bells Chiming 2	Clapping 2	Engine 3	Machine Running	Running water faucet 4	Train whistle 2
Bells Chiming 3	Clapping 3	Eruption	Marching	Sanding	Train whistle 3
Bells Chiming 4	Clashing Metal	Explosion 1	Metal Clinking 1	Hand saw 1	Trampoline
Bells Chiming 5	Clattering 1	Explosion 2	Metal Clinking 2	Hand saw 2	Treadmill
Bells Chiming 6	Clattering 2	Film Reel	Metal Clinking 3	School bell	Truck
Bike bell 1	Clinking Glasses	Finger Tapping	Monkey Scream	Scraping	Truck Backing Up 1
Bike bell 2	Clock ticking 1	House Fire	Morse code 1	Scratching	Truck Backing Up 2
Bird 1	Clock ticking 2	Fire Fighters	Morse code 2	Screwing Off Lid	Truck Backing Up 3
Bird 2	Clock Tower	Fire Alarm	Motor 1	Scrubbing	Truck horn
Bird 3	Coin Dropping 1	Fire Crackers	Motor 2	Seagull 1	Turkey
Bird 4	Coin Dropping 2	Fireworks	Motor 3	Seagull 2	Typewriter
Bird 5	Coloring	Flushing	Motor 4	Seal	Typing 1
Bird 6	Construction 1	Fountain	Motor 5	Sharpening knives	Typing 2
Blender	Construction 2	Cooking Bacon	Motorboat 1	Sheep	Vacuum
Boat	Cow Mooing 1	Gargling	Motorboat 2	Shopping Cart	Vegetable Peeler
Boat Horn	Cow Mooing 2	Gavel 1	Motorcycle Revving	Shower 1	Velcro
Boiling Water	Cow Mooing 3	Gavel 2	Music Box	Shower 2	Walking in Leaves 1
Bowling Pins Falling	Cracking	Geese 1	News Paper Rustling	Shuffling Cards	Walking in Leaves 2
Breaking Glass 1	Creaky Door	Geese 2	Opening Letter	Sink	Walking in Leaves 3
Breaking Glass 2	Crushing Can	Geese 3	Owl	Siren 1	Walking on Gravel
Brushing Hair	Crinkling paper 1	Glass Shattering	Pepper Grinder	Siren 2	Walking on Hard Surface
Brushing Teeth 1	Crinkling paper 2	Goats 1	Pig Oinking 1	Siren 3	Walking with Heels
Brushing Teeth 2	Crow	Goats 2	Pig Oinking 2	Siren 4	Water dripping
Busy Signal 1	Laughing	Goats 3	Pig snorting	Siren 5	Water Flowing
Busy Signal 2	Crumpling paper	Grandfather Clock 1	Ping-Pong 1	Siren 6	Water Splashing
Saw Cutting	Cuckoo clock	Grandfather Clock 2	Ping-Pong 2	Siren 7	Waves at Beach
Camera shutter 1	Cutting with scissors 1	Grating Food	Ping-Pong 3	Skateboarding 1	Weedwhacker
Camera shutter 2	Cutting with scissors 2	Growling 1	Plane crash	Skateboarding 2	Whales
Car crash	Dancing	Growling 2	Pool balls Colliding	Skateboarding 3	Whip 1
Car Accelerating	Dentist Drill	Gunfire	Popcorn	Slicing	Whip 2
Car Alarm	Dial Tone	Guns shooting 1	Pouring Liquid	Smashing Things	Whip 3
Car Driving 1	Dishes Clanking	Guns shooting 2	Pouring water 1	Smoke alarm 1	Whistle 1
Car Driving 2	DJ Record Scratching	Guns shooting 3	Pouring water 2	Smoke alarm 2	Whistle 2
Car Driving 3	Dog Lapping Water	Guns shooting 4	Pouring water 3	Songbird	Whistle 3
Car Driving 4	Dog panting 1	Guns shooting 5	Pouring water out of bottle	Splashing Water	Whistle 4
Car Driving 5	Dog panting 2	Guns shooting 6	Power tools	Sports Arena Buzzer	Windchimes
Car engine Starting 1	Dog panting 3	Hammering 1	Printing 1	Aerosol Can Shaking	Winding up device
Car engine Starting 2	Dog barking 1	Hammering 2	Printing 2	Spraying Aerosol can	Writing 1
Car Horn	Dog barking 2	Hawk	Printing 3	Stomach Growling	Writing 2
Car window rolling down	Dog barking 3	Heart Beat 1	Puppy whining	Stove	Writing on Chalkboard 1
Car Skidding	Dog barking 4	Heart Beat 2	Radio Tuning	Stream 1	Writing on Chalkboard 2
Car Sputtering	Dog barking 5	Heart Beat 3	Rain	Stream 2	
Cash Register	Dog barking 6	Horse neigh 1	Ratchet	Stream 3	
Castanets	Doorbell 1	Horse neigh 2	Rattling	Suitcase rolling	
Cell Phone Vibrating	Doorbell 2	Horse neigh 3	Reception Desk Bell	Swimming	
Chainsaw Cutting 1	Doorbell 3	Horse neigh 4	Revving Engine 1	Swords Clashing	

5/ Lines 124-125: please clarify if the reverberation was simulated in a binaural setting or a monaural setting. If the latter, as indicated by line 125 that the direction of reflections was created by convolving the monaural reflection with HRTFs, then this is perhaps very artificial in that source reflection is identical in all directions. This is not the case if binaural room impulse responses are measured.

The reverberation was simulated in a binaural setting, and we have clarified this in the revised text:

In the Results section:
"Each reflection was then filtered by the (binaural) head-related impulse response for the direction of the reflection[51]." (lines 128-129)

And in the Methods:
"This simulator used the image-source method, which approaches an exact solution to the wave equation if the walls are assumed to be

rigid[57], as well as an extension to that method that allowed for more accurate calculation of the arrival time of a wave[140]. This enabled the simulator to correctly render the relative timing between the signals received by the two simulated ears, including reflections (enabling both the direct sound and all reflections to be rendered with the correct spatial cues)." (lines 693-698)

6/ Line 128: please clarify how many randomly chosen locations were used for background noise and why. Could the choice have an impact on the results? What about diffuse noise?

We have clarified in the results section that we used between 3-8 noise locations, uniformly distributed (over both the number of sources and their positions). We have added a brief note to the Results section to make this clear independent of the Methods section.

"Background noise was synthesized from the statistics of a natural sound texture[58], and was rendered at between 3 and 8 randomly chosen locations using the same room simulator, in order to produce noise that was diffuse but non-uniform, intended to replicate common real-world sources of noise." (lines 131-134)

We made this choice on grounds of ecological validity, as noise sources are almost always directional to some extent. By adding noises rendered at different locations we obtained background noise that was not as precisely localized as the target sound sources, which seemed a reasonable approximation of common real-world conditions. This motivation has been clarified in the revised text:

"Backgrounds were created by spatially rendering between 3 and 8 exemplars of the same texture at randomly chosen locations using the virtual acoustic simulator described above. We made this choice on grounds of ecological validity, based on the intuition that noise sources are typically not completely spatially uniform[95] despite being more diffuse than sounds made by single organisms or objects. By adding noises rendered at different locations we obtained background noise that was not as precisely localized as the target sound sources, which seemed a reasonable approximation of common real-world conditions." (lines 773-779)

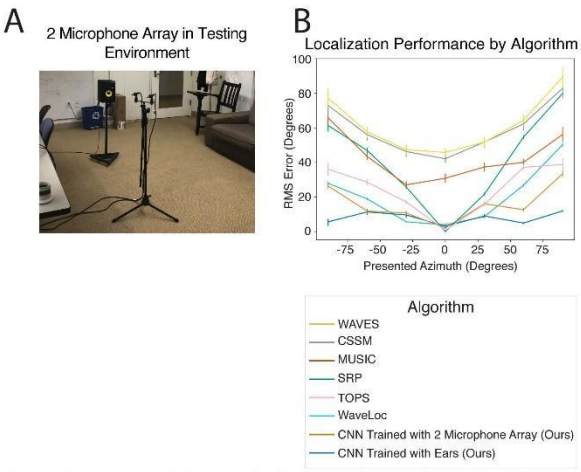
7/ Line 131: How were the neural network outputs mapped to the location? Were the azimuth/elevation pair labels used in a softmax fashion or a regression fashion? What are the range of azimuth and elevation? How many labels are there?

We used a softmax function over the location labels. To clarify this, we have added a summary of the output representation to the Results section (it is specified in detail in the Methods section):

"This network instantiated a cascade of simple operations – filtering, pooling, and normalization – culminating in a softmax output layer with 504 units corresponding to different spatial locations (spaced 5° in azimuth and 10° in elevation)." (lines 105-108)

8/ Line 192: the reviewer is surprised to see that the authors only compared their system to microphone-array localization systems. There are a large number of studies in binaural sound localization systems that exploit head/torso related transfer functions in both anechoic and reverberant conditions. To name a few:
* May et al. (2011) "A probabilistic model for robust localization based on a binaural auditory front-end," IEEE TASLP
* Woodruff and Wang (2012) "Binaural localization of multiple sources in reverberant and noisy environments," IEEE TASLP
* Ma et al. (2017) "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments", IEEE TASLP
* Vecchiotti et al. (2019) End-to-end binaural sound localisation from the raw waveform, IEEE ICASSP
* There is also a chapter on "Binaural Sound Localization" by R Stern in the book Computational Auditory Scene Analysis edited By DeLiang Wang and Guy Brown

Thank you for the suggestions. We are limited by the public availability of the methods in question. Only one of the papers listed by the reviewer has corresponding available code, but we now include that algorithm (Vecchiotti et al. 2019; 'WaveLoc') in our comparison. We note that the comparisons to other systems was not the main point of the paper, and have moved it to a supplementary figure in response to the comments of the editor and of Reviewer 1. Here is the revised figure:



Supplemental Figure 2. *Comparison of our model to alternative two-microphone localization systems. A. Photo of two-microphone array. Microphone spacing was the same as in the mannequin, but the recordings lacked the acoustic effects of the pinnae, head, and torso. B. Localization accuracy of standard two-microphone localization algorithms, our neural network localization model trained with ear/head/torso filtering effects (same as 1G and 1H), neural networks trained instead with simulated input from the two-microphone array. Localization judgments are front-back folded. Error bars here and in C plot SEM, obtained by bootstrapping across stimuli.*

Although we could not include performance comparisons to the algorithms in the other papers cited by the reviewer, we have added references to them in the revised manuscript.

9/ Please clarify how human listeners' responses were recorded - were the humans able to see (or informed of) the 11 loudspeakers? How long were the stimuli?

We added a section in the Methods to describe the details of this experiment. The humans were able to see the loudspeakers, and the stimuli were 200 ms. The added text reads:

"To provide an example of free-field human sound localization, Fig. 1F plots the results of an experiment by Yost and colleagues[59]. In that experiment, humans were presented with noise bursts (lowpass filtered white noise with a cutoff of 6 kHz, 200ms in duration, with 20ms cosine onset and offset ramps) played from one of 11 speakers in an anechoic chamber. The speakers were spaced every 15

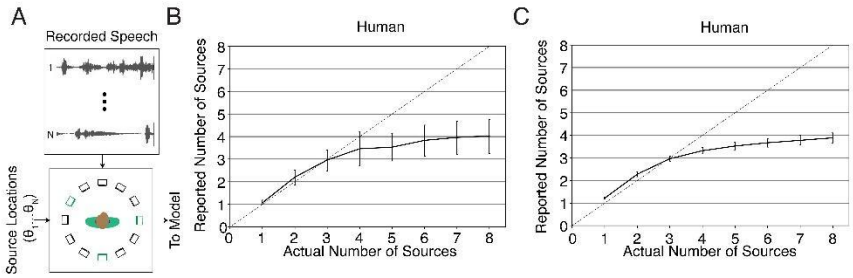
***degrees, with the array centered on the midline. Speakers were visible to participants. Participants indicated the speaker from which the sound was played by entering a number corresponding to the speaker.*" (lines 1144-1149)**

10/ Figure 4B: what is the Y-axis unit? Although the proposed model shows a V-shape result pattern across azimuth, the localization errors at the lateral positions seem very large (30-40 degrees) compared to human data.

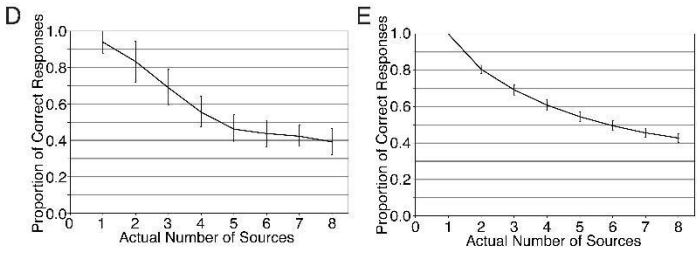
We have clarified in the caption that the human graph plots discriminability (d') of pairs of noises separated by a fixed spatial angle. The units on the human and model graphs are not the same, because the tasks were not exactly the same (for simplicity, we measured absolute localization error of the model). The large error probably reflects the fact that the stimuli in this experiment were very brief.

11/ While it is interesting to see the proposed model is able to replicate many psychoacoustic experiments, all the experiments are focused on single-source localization and the authors did not include any multi-source localization experiments. In realistic listening conditions there are often multiple sources present and indeed this is the setting where the authors trained the neural networks. There are some studies in human sound localization in such settings, e.g.
* N. Kopco, V. Best, and S. Carlile, "Speech localization in a multitalker mixture," J. Acoust. Soc. Amer., vol. 127, no. 3, pp. 1450--1457, 2010.

We have added a multi-source localization experiment that measured both the number of heard sources as well as the accuracy of their localization (Zhong and Yost, 2017). The model reproduces the results rather well. The number of estimated sources shows a very similar dependence on the actual number of sources, despite not being fit to match the human data:



And the dependence of localization accuracy on number of sources is also reproduced quite well:



The revised text now describes this additional experiment:

"Multi-source localization
Humans are able to localize multiple concurrent sources, but only to a point[95-97]. The reasons for the limits on multi-source localization are unclear[96]. These limitations could reflect human-specific cognitive constraints. For instance, reporting a localized source might require attending to it, which could be limited by central factors not specific to localization. Alternatively, localization could be fundamentally limited by corruption of binaural cues by concurrent sources or other ambiguities intrinsic to the localization problem.

To assess whether the model would exhibit limitations like those observed in humans, we replicated an experiment[97] in which humans judged both the number and location of a set of speech signals played from a subset of an array of speakers (Fig. 6A). To enable the model to report multiple sources we fine-tuned the final fully-connected layer (freezing all weights in earlier layers) to indicate the probability of a source at each of the location bins, and set a probability criterion above which we considered the model to report a sound at the corresponding location (see Methods). We then tested the model on the experimental stimuli.

Humans accurately report the number of sources up to three, after which they undershoot, only reporting about four sources in total regardless of the actual number (Fig. 6B). The model reproduces this effect, also being limited to approximately four sources (Fig. 6C). Human localization accuracy also systematically drops with the number of sources (Fig. 6D); the model again quantitatively reproduces this effect (Fig. 6E). The model-human similarity suggests that these limits on sound localization are intrinsic to the constraints of the localization problem, rather than reflecting human-specific limitations." (lines 386-407)

We were excited to see such a close match to human data in a domain that to our knowledge has not previously been modeled or understood at a theoretical level, and we appreciate you having provided the impetus to try this.

12/ Lines 401-403: The reviewer finds it difficult to accept this explanation. The neural networks are very easy to be trained to work well in a match condition. When compared to a generative model, neural networks are prone to overfitting and often fail to generalize to unseen conditions. Thus it is important to include "noise" during training of neural networks. The increased dissimilarity is probably due to the mismatch between training and testing conditions, rather than "human-like spatial hearing emerged from task optimization only for naturalistic training conditions". The reviewer suspects if the human listeners did the tests in an anechoic room, and anechoic test data was used for the neural network trained in the anechoic condition, the similarity between humans / machines is properly still there.

We believe this comment reflects a misunderstanding. The models were in fact tested in the same conditions that humans were tested in. So for experiments where the humans were in an anechoic chamber (e.g. for the precedence effect), so was the model. We have clarified this in the revised text:

"We replicated the conditions of the original experiments as best possible (e.g. when humans were tested in anechoic conditions, we rendered experimental stimuli in an anechoic environment)." (lines 196-198)

Also, many studies (e.g. May et al. (2011), Woodruff and Wang (2012), Ma et al. (2017)) have shown that by including white noise during training in the anechoic condition will significantly increase the robustness of a machine localization system to reververantions.

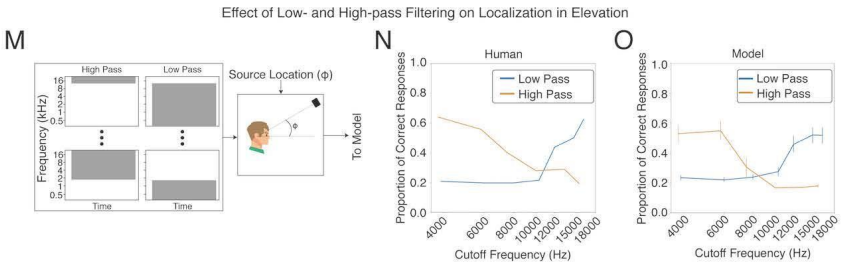
We have added a note to point out that the improved localization performance in the model trained in background noise is consistent with prior work in the engineering literature, citing the papers listed by the reviewer:

"This finding is consistent with the common knowledge in engineering that training systems in noisy and otherwise realistic conditions aids performance[35,40,42,99]." (lines 485-487)

13/ It is also a shame that the authors did not include any experiments in elevation localization, despite that the neural network was trained to do so. It is

understandable that COVID-19 creates a huge challenge to run more listening tests, but perhaps the authors could verify the machine performance against the previous human listening results?

We note that we did model two experiments that involve localization in elevation (shown in Figure 5 in the original submission; Figure 4A-L in the revised manuscript). However, in response to this comment we modeled an additional experiment measuring localization in elevation (Hebrank and Wright, 1974). This experiment measured the effect of low-pass and high-pass filtering on localization accuracy in elevation. We again found that the model reproduced the human results to what we consider to be a fairly compelling extent:



We have added this new experiment to the revised manuscript. Here is the new section of the text describing the new experiment:

"Dependence on high-frequency spectral cues to elevation
The cues used by humans for localization in elevation are primarily in the upper part of the spectrum[87,88]. To assess whether the trained networks exhibited a similar dependence, we replicated an experiment measuring the effect of high-pass and low-pass filtering on the localization of noise bursts[89] (Fig. 4M). Model performance varied with the frequency content of the noise in much the same way as human performance (Fig. 4N&O)." (lines 312-317)

14/ Although the authors have mentioned the front-back confusion and the head movements during human sound localization, this is not examined further. There are classical studies examining this aspect, e.g.:
H. Wallach (1940) "The role of head movements and vestibular and visual cues in sound localization," Journal of Experimental Psychology, vol. 27, no. 4

Although front-back confusions are widely discussed in the spatial hearing literature, we are not aware of experiments measuring front-back confusions in settings where the listener cannot move their head, as would be analogous to the setting in which our model must

operate. We thus showed that the model makes front-back confusions (Figure 1H) but could not include a quantitative comparision with human listeners. Incorporating head movements is an exciting direction for extensions of our work, and one application would be to explain the resolution of front-back confusions as found in humans.

We have expanded the discussion of this future direction in the Discussion section, referencing the Wallach paper mentioned by the reviewer:

"One natural extension of our model would be to incorporate moving sound sources and head movements. We modeled sound localization in static conditions because the vast majority of experimental data has been collected in this setting. But in real-world conditions sound sources often move relative to the listener, and listeners move their head[127,128], often to better disambiguate front from back[61] and more accurately localize. Our approach could be straightforwardly expanded to moving sound sources in the virtual training environment, and a network that can learn to move its head[40], potentially yielding explanations of auditory motion perception[129-131]. The ability to train models that can localize in realistic conditions also underscores the need for additional measurements of human localization behavior – front-back confusions, localization of natural sounds in actual rooms, localization with head movements etc. – with which to further evaluate models." (lines 673-682)

Decision Letter, first revision:

Our ref: NATHUMBEHAV-200711759A

18th August 2021

Dear Dr. McDermott,

Thank you for submitting your revised manuscript "Deep neural network models of sound localization reveal how perception is adapted to real-world environments" (NATHUMBEHAV-200711759A). It has now been seen by the original referees and their comments are below. As you can see, the reviewers find that the paper has improved in revision. We will therefore be happy in principle to publish it in Nature Human Behaviour, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements in about a week. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Please do not hesitate to contact me if you have any questions.

Sincerely,

Samantha Antusch

Samantha Antusch, PhD
Editor
Nature Human Behaviour

Reviewer #1 (Remarks to the Author):

The authors revised the manuscript extensively. The focus changed, being now more inline with the idea of using neural networks to predict human behavior. I like the content, especially, I appreciate the many experiments showing how well the proposed network can replicate human responses in various tasks. I also appreciate the effort to provide all the material to reproduce authors' work by others. I see that the precomputed training data are stored on a commercial cloud. I don't know how persistent this is, and it might be more persistent and transparent to have stored it on systems such as Zenodo, which would provide a DOI to the data.

The only major issues I have are the abstract and the title. They seem to be not adapted to the revised version and they do not seem to reflect the new content.

The title promises an understanding of "how perception is adapted", but the content actually shows the ability of the proposed network to replicate the perception – under many conditions, not only

sound localization, thus “models of sound localization” is misleading. Hence, I suggest to change the title to “Deep neural networks are able to replicate spatial auditory perception”, or similar.

The abstract has a similar problem as the title. First “But when trained in unnatural environments without either reverberation, noise, or natural sounds, these performance characteristics deviated from those of humans. “ is a tautology because if we want to mimic human behavior in natural environment, we need to train a network on data reflecting those natural environments. I suggest to remove this sentence. Second, “The results show how biological hearing is adapted to the challenges of real-world environments [..]” is actually not reflecting the manuscript content because the results do not show how hearing is adapted – they rather show the ability of replicating aspects of biological hearing. Thus, I suggest to rephrase the last sentence of the abstract to e.g., “The results show how artificial neural networks can replicate human spatial hearing and extend traditional ideal observer models to real-world domains.”

With those changes, I congratulate the authors on the great work and highly recommend to have the manuscript published in Nature Human Behavior.

Reviewer #2 (Remarks to the Author):

I viewed this article very favorably the first time around. I have evaluated this revision, together with the other reviewer comments and rebuttal letter. I feel the substantial revisions and clarifications now better highlight the fundamental implications of the findings to the study of human behavior. I remain strongly positive about this manuscript.

Reviewer #3 (Remarks to the Author):

The authors have done a thorough job revising the manuscript and addressing the issues and concerns raised by the reviewer.

In particular, the additional experiments (e.g. multisource localization) have added a great value to the original version in validating the model in more natural listening conditions. The reviewer would like to double check if the deep neural networks were retrained for these new tasks, or they were the same networks used to model the other tasks, i.e. the model was used in a multi-task manner. It is important to clarify this as it's one of the major contributions of this work.

The objectives and contributions have been made clearer in the introduction.

Head movements are not included in the current work, despite that they are natural for human listeners. However, the model can still match a large set of human data without head movements. Incorporating head movements is now discussed in the future direction. It would be interesting to see what impact the head movements would have on the model output.

It is nice to see that the code and data used in this manuscript have been made available, allowing results to be reproduced. This is in itself of great value to the scientific community.

I believe this manuscript will be of broad interests to and beyond the spatial hearing research community and would therefore recommend its publication.

Decision letter, final requests:

** Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your co-authors. **

Our ref: NATHUMBEHAV-200711759A

16th September 2021

Dear Dr. McDermott,

Thank you for your patience as we've prepared the guidelines for final submission of your Nature Human Behaviour manuscript, "Deep neural network models of sound localization reveal how perception is adapted to real-world environments" (NATHUMBEHAV-200711759A). Please carefully follow the step-by-step instructions provided in the attached file, and add a response in each row of the table to indicate the changes that you have made. Ensuring that each point is addressed will help to ensure that your revised manuscript can be swiftly handed over to our production team.

We would hope to receive your revised paper, with all of the requested files and forms within two-three weeks. Please get in contact with us if you anticipate delays.

When you upload your final materials, please include a point-by-point response to any remaining reviewer comments.

If you have not done so already, please alert us to any related manuscripts from your group that are under consideration or in press at other journals, or are being written up for submission to other journals (see: https://www.nature.com/nature-research/editorial-policies/plagiarism#policy-on-duplicate-publication for details).

Nature Human Behaviour offers a Transparent Peer Review option for new original research manuscripts submitted after December 1st, 2019. As part of this initiative, we encourage our authors to support increased transparency into the peer review process by agreeing to have the reviewer comments, author rebuttal letters, and editorial decision letters published as a Supplementary item. When you submit your final files please clearly state in your cover letter whether or not you would like to participate in this initiative. Please note that failure to state your preference will result in delays in accepting your manuscript for publication.

In recognition of the time and expertise our reviewers provide to Nature Human Behaviour's editorial process, we would like to formally acknowledge their contribution to the external peer review of your manuscript entitled "Deep neural network models of sound localization reveal how perception is adapted to real-world environments". For those reviewers who give their assent, we will be publishing their names alongside the published article.

Cover suggestions

As you prepare your final files we encourage you to consider whether you have any images or illustrations that may be appropriate for use on the cover of Nature Human Behaviour.

Covers should be both aesthetically appealing and scientifically relevant, and should be supplied at the best quality available. Due to the prominence of these images, we do not generally select images featuring faces, children, text, graphs, schematic drawings, or collages on our covers.

We accept TIFF, JPEG, PNG or PSD file formats (a layered PSD file would be ideal), and the image should be at least 300ppi resolution (preferably 600-1200 ppi), in CMYK colour mode.

If your image is selected, we may also use it on the journal website as a banner image, and may need to make artistic alterations to fit our journal style.

Please submit your suggestions, clearly labeled, along with your final files. We'll be in touch if more information is needed.

ORCID

Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance:
https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Nature Human Behaviour has now transitioned to a unified Rights Collection system which will allow our Author Services team to quickly and easily collect the rights and permissions required to publish your work. Approximately 10 days after your paper is formally accepted, you will receive an email in providing you with a link to complete the grant of rights. If your paper is eligible for Open Access, our Author Services team will also be in touch regarding any additional information that may be required to arrange payment for your article. Please note that you will not receive your proofs until the publishing agreement has been received through our system.

Please note that Nature Human Behaviour is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

<B>Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. For submissions from January 2021, if your research is supported by a funder that requires immediate open access (e.g.

according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route our standard licensing terms will need to be accepted, including our self-archiving policies. Those standard licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

For information regarding our different publishing models please see our Transformative Journals page. If you have any questions about costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com.

Please use the following link for uploading these materials:
[REDACTED]

If you have any further questions, please feel free to contact me.

Best regards,
Chloe Knight
Editorial Assistant
Nature Human Behaviour

On behalf of

Samantha Antusch

Samantha Antusch, PhD
Editor
Nature Human Behaviour

Reviewer #1:
Remarks to the Author:
The authors revised the manuscript extensively. The focus changed, being now more inline with the idea of using neural networks to predict human behavior. I like the content, especially, I appreciate the many experiments showing how well the proposed network can replicate human responses in various tasks. I also appreciate the effort to provide all the material to reproduce authors' work by others. I see that the precomputed training data are stored on a commercial cloud. I don't know how persistent this is, and it might be more persistent and transparent to have stored it on systems such as Zenodo, which would provide a DOI to the data.

The only major issues I have are the abstract and the title. They seem to be not adapted to the revised version and they do not seem to reflect the new content.

The title promises an understanding of "how perception is adapted", but the content actually shows

the ability of the proposed network to replicate the perception – under many conditions, not only sound localization, thus "models of sound localization" is misleading. Hence, I suggest to change the title to "Deep neural networks are able to replicate spatial auditory perception", or similar.

The abstract has a similar problem as the title. First "But when trained in unnatural environments without either reverberation, noise, or natural sounds, these performance characteristics deviated from those of humans. " is a tautology because if we want to mimic human behavior in natural environment, we need to train a network on data reflecting those natural environments. I suggest to remove this sentence. Second, "The results show how biological hearing is adapted to the challenges of real-world environments [..]" is actually not reflecting the manuscript content because the results do not show how hearing is adapted – they rather show the ability of replicating aspects of biological hearing. Thus, I suggest to rephrase the last sentence of the abstract to e.g., "The results show how artificial neural networks can replicate human spatial hearing and extend traditional ideal observer models to real-world domains."

With those changes, I congratulate the authors on the great work and highly recommend to have the manuscript published in Nature Human Behavior.

Reviewer #2:
Remarks to the Author:
I viewed this article very favorably the first time around. I have evaluated this revision, together with the other reviewer comments and rebuttal letter. I feel the substantial revisions and clarifications now better highlight the fundamental implications of the findings to the study of human behavior. I remain strongly positive about this manuscript.

Reviewer #3:
Remarks to the Author:
The authors have done a thorough job revising the manuscript and addressing the issues and concerns raised by the reviewer.

In particular, the additional experiments (e.g. multisource localization) have added a great value to the original version in validating the model in more natural listening conditions. The reviewer would like to double check if the deep neural networks were retrained for these new tasks, or they were the same networks used to model the other tasks, i.e. the model was used in a multi-task manner. It is important to clarify this as it's one of the major contributions of this work.

The objectives and contributions have been made clearer in the introduction.

Head movements are not included in the current work, despite that they are natural for human listeners. However, the model can still match a large set of human data without head movements. Incorporating head movements is now discussed in the future direction. It would be interesting to see what impact the head movements would have on the model output.

It is nice to see that the code and data used in this manuscript have been made available, allowing results to be reproduced. This is in itself of great value to the scientific community.

I believe this manuscript will be of broad interests to and beyond the spatial hearing research community and would therefore recommend its publication.

Author Rebuttal, first revision:

Please note that line numbers refer to those in the article pdf that we uploaded (they seem to fluctuate across Word versions, so we printed a pdf for a fixed reference).

Reviewer #1 (Remarks to the Author):

The authors revised the manuscript extensively. The focus changed, being now more inline with the idea of using neural networks to predict human behavior. I like the content, especially, I appreciate the many experiments showing how well the proposed network can replicate human responses in various tasks. I also appreciate the effort to provide all the material to reproduce authors' work by others. I see that the precomputed training data are stored on a commercial cloud. I don't know how persistent this is, and it might be more persistent and transparent to have stored it on systems such as Zenodo, which would provide a DOI to the data.

> **We looked into Zenodo and they have a 50 GB limit that our data sets exceed. We thus think the cloud server is the best option available.**

The only major issues I have are the abstract and the title. They seem to be not adapted to the revised version and they do not seem to reflect the new content.

> **We view the new content as clarifying the scientific objectives that are expressed in the title and abstract of the paper (and that were always intended as the focus of the work). The other reviewers indicate that these objectives are now much clearer. We have adjusted the messaging in the introduction to try to further clarify this:**
>
> ***"Human behaviors that emerge from machine learning under a set of naturalistic environmental constraints, but not under alternative constraints, are plausibly a consequence of optimization for those natural constraints (i.e., adapted to the natural environment) (Fig. 1A)." (lines 47-49)***

We note that the logic and purpose of the paper is made explicit in the first and last paragraphs of the introduction, and is consistent with the message in the title and abstract:

"Why do we see or hear the way we do? Perception is believed to be adapted to the world – shaped over evolution and development to help us survive in our ecological niche. Yet adaptedness is often difficult to test. Many phenomena are not obviously a consequence of adaptation to the environment, and perceptual traits are often proposed to reflect implementation constraints rather than the consequences of performing a task well." (lines 30-34)

"When tested on stimuli from classic laboratory experiments, the resulting model replicated a large and diverse array of human behavioral characteristics. We then trained models in unnatural conditions to simulate evolution and development in alternative worlds. These alternative models deviated notably from human-like hearing. The results suggest that the characteristics of human hearing are indeed adapted to the constraints of real-world localization, and that the rich panoply of sound localization phenemona can be explained as consequences of this adaptation. The approach we employ is broadly applicable to other sensory modalities, providing a way to test the adaptedness of aspects of human perception to the environment and to understand the conditions in which human-like perception arises." (lines 86-93)

The title promises an understanding of "how perception is adapted", but the content actually shows the ability of the proposed network to replicate the perception – under many conditions, not only sound localization, thus "models of sound localization" is misleading. Hence, I suggest to change the title to "Deep neural networks are able to replicate spatial auditory perception", or similar.

We respectfully disagree. As noted above, the point of the paper is to test whether perception is adapted to the natural environment, by assessing whether systems optimized under natural, but not unnatural, conditions exhibit human-like traits. And all the traits in question pertain to sound localization. We thus think the title is appropriate, and have opted to leave it as is. We hope the tweaks to the introduction help to further clarify why this title is appropriate.

The abstract has a similar problem as the title. First "But when trained in unnatural environments without either reverberation, noise, or natural sounds, these performance characteristics deviated from those of humans. " is a tautology because if we want to mimic human behavior in

natural environment, we need to train a network on data reflecting those natural environments. I suggest to remove this sentence.

> **We respectfully disagree. It is not a foregone conclusion that training in natural conditions is necessary to reproduce human behavior in natural environments. In principle, and as we point out in the introduction, human behavior could be strongly influenced by neural implementation constraints. If this were the case, the behavioral phenotype of a system optimized in natural environments might not resemble that of humans because it lacks the same neural implementation constraints. However, we also note that with the exception of Figure 1, we are not evaluating human behavior in natural environments, but rather in laboratory conditions intended to probe particular characteristics of spatial hearing. So it is highly nonobvious whether training in natural conditions would suffice to reproduce all the behavioral traits that we examined. One of the primary purposes of the paper is to vary the training conditions and test whether human traits are specific to systems optimized in natural conditions.**

Second, “The results show how biological hearing is adapted to the challenges of real-world environments [..]” is actually not reflecting the manuscript content because the results do not show how hearing is adapted – they rather show the ability of replicating aspects of biological hearing. Thus, I suggest to rephrase the last sentence of the abstract to e.g., “The results show how artificial neural networks can replicate human spatial hearing and extend traditional ideal observer models to real-world domains.”

> **The point of the paper is not merely to show that neural networks can replicate aspects of biological hearing but also to demonstrate that the behavioral traits in question are adapted to natural environments. One way to test whether a behavioral trait is adapted to the natural environment is to assess whether systems optimized under natural, but not unnatural, conditions exhibit the behavioral trait of interest. Based on the experiments and results we have presented, we think the content of the abstract is appropriate, and important in conveying our message.**

With those changes, I congratulate the authors on the great work and highly recommend to have the manuscript published in Nature Human Behavior.

> **Thank you.**

Reviewer #2 (Remarks to the Author):

I viewed this article very favorably the first time around. I have evaluated this revision, together with the other reviewer comments and rebuttal letter. I feel the substantial revisions and clarifications now better highlight the fundamental implications of the findings to the study of human behavior. I remain strongly positive about this manuscript.

Thank you.

Reviewer #3 (Remarks to the Author):

The authors have done a thorough job revising the manuscript and addressing the issues and concerns raised by the reviewer.

In particular, the additional experiments (e.g. multisource localization) have added a great value to the original version in validating the model in more natural listening conditions. The reviewer would like to double check if the deep neural networks were retrained for these new tasks, or they were the same networks used to model the other tasks, i.e. the model was used in a multi-task manner. It is important to clarify this as it's one of the major contributions of this work.

The neural networks were not retrained. We just added a single-layer decision stage that was retrained to enable the model to report multiple sound sources. We have clarified this in the relevant section of the results:

"The weights in all earlier layers were "frozen" during this fine-tuning, such that all other stages of the model were identical to those used in all other experiments." (lines 304-305)

The objectives and contributions have been made clearer in the introduction.

Head movements are not included in the current work, despite that they are natural for human listeners. However, the model can still match a large set of human data without head movements. Incorporating head movements is now discussed in the future direction. It would be interesting to see what impact the head movements would have on the model output.

It is nice to see that the code and data used in this manuscript have been made available, allowing results to be reproduced. This is in itself of great value to the scientific community.

I believe this manuscript will be of broad interests to and beyond the spatial hearing research community and would therefore recommend its publication.

Thank you.

Final Decision Letter: