Peer Review File

Manuscript Title: High-performance brain-to-text communication via handwriting

Editorial Notes: *none*

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referee #1 (Remarks to the Author):

A. Summary of the key results
This paper represents a truly novel approach to restoring communication with a brain-computer interface. Previous approaches have used point-and-click cursor control to enable communication with an onscreen keyboard and have demonstrated very good performance that enables functional performance. Here, the authors instead try to decode handwriting movements in order to predict individual letters as the brain-computer interface (BCI) users imagines writing words and sentences. Impressively, online BCI performance was more than twice as fast as previously demonstrated and approaches smartphone typing speeds. Further, the authors demonstrate that the temporal variability associated with handwriting trajectories is a major contributor to the high level of performance that was shown, which has implications for BCIs in general as it may be advantageous to try to decode complex and dexterous movements.

B. Originality and significance:
This study takes a new and original approach to BCI-controlled communication by decoding attempted handwriting movements in order to enable computer-based communication. This approach is unique because rather than decoding the movement trajectory directly (although they demonstrate that this is possible), they implement a two-step classification process using an RNN to identify when the user is attempting to write a character and then determining which character the user is trying to write. The decoding approach relies on both the spatial and temporal variability of the attempted movements to boost performance far beyond what has previously been demonstrated for BCI-based communication.

This work provides evidence that an intracortical BCI can enable fast rates of communication based on decoded handwriting patterns. This work is therefore of interest to scientists and engineers developing neural interfaces to restore communication as well as clinicians working with patients with communication impairments.

C. Data & methodology:

1) This paper is well written and clearly describes the key details and decision points that were used to implement the RNN-based decoding approach. The figures highlight key methodologicial elements and results. A rigorous approach was taken to investigate the impact of various optimization parameters, data quantity, and data quality (vs. noise). All data and code will be made publicly available providing an extremely valuable resources for the research community as well as transparency in reporting.

2) Performance metrics are appropriate and the details of how each was calculated are included.

D. Appropriate use of statistics and treatment of uncertainties:

1) All data are presented from a single subject across multiple data sessions. This is appropriate given the limited number of human participants that have been implanted with an intracortical BCI, the rigor of the approach, and the importance of the findings.

2) Statistical tests should be performed to compare between the character and lines conditions for data shown in Figure 3C and E and reflected in the manuscript and figure.

3) While many comparisons are made based on qualitative results or comparisons of confidence intervals, the effects and improvements over previous methods are large and robust. Further statistical analysis is not needed to support the conclusions.

E. Conclusions:

1) The major conclusions are robustly supported by the presented data with consistent performance achieved across multiple sessions

2) The limitations section should mention that this work comes from a single subject who had the ability to write prior to his injury.

3) The authors conclude that a handwriting BCI is the first type of BCI that has the potential to work in people with visual impairments. This was not evaluated in the present study. While the subject did not have feedback of BCI performance until after each letter was selected, this did provide feedback that could be accumulated over the course of the session. Additionally, the importance of this was not made clear. Other forms of feedback (auditory, tactile, etc.) could be used to convey information to a person with visual impairments. Further, it is a very small population that is impacted by both visual and communication impairments.

F. Suggested improvements and comments:

1) Results, line 45: specify that the participant had a cervical spinal cord injury and be more precise in the description of residual movement abilities.

2) Results, line 60: why was a non-linear approach (t-SNE) selected for data visualization and separability analysis given that PCA allowed for accurate trajectory reconstruction. Readability would be improved by understanding the intuition that guided this decision.

3) Results, line 63: Please provide a confidence interval (or similar measure of variability) for the k-nearest neighbor classification result.

4) Results, lines 95-106: It is important to note that a large amount of training data needed to be collected each day. In addition to reporting the number of sentences, the authors should report the number of characters and duration of data collection in the main text. It is noted that this information is included in the Supplemental Material. Additionally it wasn't clear from the main text that "…data was cumulatively added to the training dataset…" referred to data collected prior to BCI control, rather than just adding in data as it was collected during BCI assessment.

5) Results, figure 2C. It is interesting that day 1237 seems to have a higher character error rate that interrupts what appears to be a linear increase in error rate that is mirrored by an increase in characters per minute. Is there a reason for this? Across the 5 sessions, did the participant have a change in strategy (e.g. to go faster with less regard for error?).

6) Results, Table 1: For clarity, I suggest renaming the second row "online output + offline language model".

7) Results, Lines 166-174: How do the values chosen for simulated neural noise compare the variability in feature means that were observed in the experiment?

8) Results, Lines 178-183 & Figure 3E: While the effect of temporal dimensionality is more striking, spatial dimensionality is also likely statistically different between the characters and straight lines. This statement may therefore be too strong: "We found that the spatial dimensionality was similar for straight-lines and characters (Fig. 3E)."

9) Results- suggestions for additional data presentation:
a) Did the subject provide any subjective feedback about ease of use, training duration, suggestions for improvements, etc?
b) Had the subject previously used a point-and-click communication BCI?
c) Was there any notable change in performance within a session?
d) The authors state that the language model is capable of running in real-time. If this is the case, why wasn't this done? With the data presented, the major outcome that should be reported in the abstract is the fully-online performance with notes about how this can be improved offline.

10) Discussion, line 252: This sentence states that the subject's hand never movement, but a video is shown to highlight the micromotions. Was the subject intending to trace the letter trajectory, even if his injury likely limited his ability to do so accurately?

11) Methods, lines 689-692: Additional detail about the linear transformation and process of fitting separate input layers each day should be stated here, or clearly linked to the supplemental methods. The supplemental figure alone is not sufficient for understanding these steps.

G. References:
References are appropriate. The only comment is with regard to Reference 24 that is cited to show that EEG-BCI has achieved speeds of 60 characters per minute. This is a generous statement and other limitations could be noted given that that level of performance is not typical and was obtained from some healthy subjects due cued typing. This is a minor point.

H. Clarity and context:
1) In the abstract, results, and discussion, the authors refer to the subject as being completely paralyzed below the neck and that he performed "imagined" hand movements. However, they note that the subject retained some movement of his shoulders and that he had micromotions of his hand during the handwriting task. It would be more appropriate to describe any residual function in the subject's arm and hand. Additionally, the authors should clarify if the subject was imagining the movements or attempting them (resulting in micromotions). See for example previous work from this group: Rastogi, A., Vargas-Irwin, C.E., Willett, F.R. et al. Neural Representation of Observed, Imagined, and Attempted Grasping Force in Motor Cortex of Individuals with Chronic Tetraplegia. Sci Rep 10, 1429 (2020).

2) The abstract should report the typing speeds and accuracy that were achieved completely online without the language model since that is most representative of actual performance. It would be appropriate to also include results with offline enhancements as these would be acceptable in many contexts (such as writing an email).

Referee #2 (Remarks to the Author):

Willett et al. present an intracortical BCI (iBCI) decoding approach for classifying many characters to enable rapid typing. Their approach uses an RNN architecture to perform classification on neural activity as the subject imagines writing letters/words/sentences. They achieve typing speeds up to

90 characters per minute with above 94.5% accuracy in one subject, which significantly outperforms previous communication iBCIs. They demonstrate the system works across several sessions and both for copying text and free expression. The authors further provide analyses to provide intuition for why their approach succeeds--they achieve high classification accuracy by having the user perform a task that generates highly discriminable neural activity.

Overall, the manuscript is very well written and represents a clear and important advance in the field of BCIs. The technical innovations of the paper include 1) methods for creating training datasets when there is minimal available information (since the subject imagined moving) and 2) methods for leveraging the power of RNNs even with relatively limited data. The approaches for challenge 2 primarily use techniques common in ANNs (data augmentation) and techniques previously shown to be useful in animal studies (adding external noise to increase robustness of the networks). The solutions to challenge 1 appear relatively novel, and are certainly new to the field of BCIs. The approach/conceptual innovation of the paper is a shift away from decoding continuous control towards a method that provides accurate classification even for a relatively large 31 character set. To my knowledge this is a notable departure from prior work.

My primary concern with the manuscript is how the author's frame the work's overall approach which should more clearly emphasize the shift towards classification. As their work demonstrates, this shift can be powerful but it is also very specialized to this task. The manuscript's current comparisons to previous state-of-the-art (Pandarinath et al.) and figure 3 fail to fully make the distinction between continuous decoding of a cursor for selecting keys on a keyboard from their BCI performing a 31-way classification. Figure 3, for instance, almost implies that Pandarinath and prior BCIs were trying to classify straight line movements, which they did not. The authors' point that discriminability of the neural activity patterns directly impacts classifier performance is well taken. And provides an intuition for why having users imagine writing letters enabled their advance. But the manuscript needs to be very clear that in and of itself does not explain why they achieve higher performance. It explains why they were able to classify a large alphabet successfully for the first time. They then achieve higher performance compared to prior work because their classifier can predict letters more quickly than the average translation + click time of continuous control cursor tasks. The primary reason I emphasize this distinction is that their classification approach solves the problem of typing quite well, but does not provide a mechanism that necessarily generalizes to other tasks that are more continuous in nature like controlling a robotic limb (the authors do not claim this, but I think it's important the paper itself makes this distinction more clearly).

Specific points:
Is this the same T5 patient from Pandarinath et al. 2017? If so, it would strengthen the manuscript's claims to highlight this direct comparison (where they are also potentially at a disadvantage if studies were performed later with likely lower quality neural recordings).

If this is the same patient T5, the manuscript should mention that this subject did have the best performance of the 3-subject cohort in that prior study. While the performance advantages of their decoder are clear, given the single subject demonstration this potential subject-to-subject variability should be discussed.

The increase in characters per minute (Figure 2C) should be discussed. In addition to being more accurate over time (which may be attributed to the addition of previous day's data to the RNN training dataset), there is also an observed increase in typing speed (characters per minute). Is this also due to additional training data or other phenomena? A retrospective analysis with decoder performance on a single day's data would be useful information.

The experimental setup for real-time decoding should be clarified. Did the subject see the raw outputs during the task?

The authors nicely isolate the effect of the RNN from the more discriminable neural activity (supplemental table S4). Though I think they somewhat overstate the importance of the RNN compared to HMM in the main manuscript methods, since the RNN's main advantage is it's robustness against noise (by the authors design with noise-training for the RNN). It's actually quite noteworthy that the neural activity differences alone still lead to solid performance in a 31-way classification task with a linear HMM.

The "character stretch factor" is not well explained in the supplements. What does this factor represent?

Figure S3C and D -- are these differences statistically significant? More quantification rather than just "substantially improved" would be useful.

I'm left with an impression that many design choices in the machine learning algorithms were hand tailored. This is fine, especially for initial proof of concept. But the discussion might benefit from mentioning that methods for more automated algorithm development/training will be needed for wider utility.

Referee #3 (Remarks to the Author):

A. Summary of the key results
The work reports a single subject's performance using an intracortical BCI that can decode imagined handwriting movements from neural activity in motor cortex and map it to text in real-time. Overall the work fits within the growing body of literature intended to demonstrate faster and more accurate BCIs with improved understanding of movement encoding and more sophisticated decoding methods.
Outstanding features of the work are:
• Typing speeds of on-screen prompt at 90 characters per minute at 99% error rate with the use of a general-purpose autocorrect and 73.8 characters at 8.54% error rate for self-generated sentences (2.25% with a language model) are significant advances over the highest reported point and-click typing with a 2D computer cursor, peaking at 40 characters per minute. Results open a new approach for BCIs and demonstrate the feasibility of accurately decoding imagined handwriting movements years after losing ability to move and speak.
• The combination of probabilistic and modeling frameworks forming a hierarchical decoding approach with multiple time scales to combat neural signal variability.
• An interesting theoretical principle is proposed in which point-to-point movements may be harder to decode from one another compared to handwritten letters. Authors attribute this to the idea that temporally complex movements, such as handwriting, may be fundamentally easier to decode than point-to-point movements.

B. Originality and significance:
The paper draws upon handwriting or touch typing as a faster means to communicate by a specific population of neurologically impaired subjects. The work is an extension to this group's past contributions on BCIs for communications to the 'locked-in' population. Results presented here would be of interest to people in the BCI community who are working on restoring communication to these people who cannot move or speak.
Overall, the work is significant and original but can be better articulated. First, authors should cite the prevalence of such conditions to put this contribution in the right context.
Second, the primary performance metric is typing speed. However, on numerous occasions, the authors attempt to give the impression that this is the primary metric that could be the sole determinant for adopting the technology. While this metric is undoubtedly critical, I think the authors should reframe this argument differently, in that it is the combination of a number of

factors —one of which is typing speed— that would ultimately make the technology a first choice for the intended population. For example, the recalibration of decoders is another such factor, and while it is acknowledged by the authors that their approach is quite extensive, it is unclear how much time and resources the recalibration process takes (see detailed comments below). Another factor is the integrity of the signals over the longevity of the implant, which is a prime issue with all invasive technology (see detailed comments below).
Third, given the paper's emphasis on how the character and word decoding rates surpass existing state of the art, the data may actually have much more information about the nature of neural representation of attempted handwriting that could benefit a broader audience (particularly the neurobiology and neurophysiology communities), but this is not emphasized in the current version of the paper. As such, it is unclear if the work will be of immediate interest to many people from several disciplines.
Fourth, direct comparison to behaviors requiring dexterous movements such as typing at speeds of 120 characters per minute for intact subjects is somewhat irrelevant since the ability to modulate brain signals to become a reliable source of control of these assistive devices vary considerably among human subjects who cannot move or speak. For example, it is unclear that the achieved speed/error rates will generalize to other subjects with similar impairment. In other occasions, they draw comparison to speech-decoding BCIs for restoring verbal communication, but this technology is at a very early stage to be compared to the current approach.
Taken together, the authors should present their findings within the broader context in which the population of potential beneficiaries need to opt for a brain surgery with unknown longevity of the implanted device and a relatively long calibration process to gain additional typing speeds (extra 33 characters/min as I consider the self-paced performance reported here to be the real use case of such communication technology).

C. Data & methodology:
General comments:
The presentation is clear, logical and readable to general audience. The reporting of data and methodology is sufficiently detailed to enable reproducing the results. They state that they will share the data and code to enable reproducibility.
Major Comments:
The authors state that they 'linearly decoded pen tip velocity from neural activity'. Arguably, this variable varies considerably among different people depending on their handwriting style, accuracy, appearance, readability, etc. Did the authors have a sample handwriting from the subject before injury so they can be compared to the ones they decoded? If so, could they analyze such data to infer the pen tip speed profiles the subject likely used to better understand if the observed neural activity correlated with the character shapes? it would be more helpful if the work attempts to provide some understanding of the extent to which the dynamics of the ensemble neural activity do actually reflect this critical behavioral parameter. Also, the authors should demonstrate the extent to which character encoding might have changed as a function of trials/sentences/sessions, particularly during times when the subject was observing the prompted text, the decoded text, and when the subject was asked to write from memory. This characterization is also needed to provide credence for the claim made in the conclusion that this is a BCI without visual feedback.
It is unclear if the authors have characterized the performance long enough (beyond the stated 10 sessions) to report how nonstationarity in the neural signals can potentially deteriorate the performance reported. In fact, with the exception of the first couple of sessions that were spaced almost a month apart, the remaining 9 sessions took place almost 6 months afterwards and were closely spaced, happening within the span of 7-8 weeks. From the extensive calibration protocol described, there seems to be substantial variability in these signals.
Specific comments:
Line 93: Why did the subject write 'periods as '~' and spaces as '>'?
Line 100: Clarify if the statement 'After each new day of decoder evaluation,' refers to offline or online decoding.
Line 112: How did the authors know the exact timing of completion of each letter by the subject in

real time to be able to display it after it was completed? It is stated that visual feedback about the decoder output was 'estimated to be between 0.4-0.7'. The supplementary material explains how they arrived at these estimates, but this inherently assumes that the character was 'completed' when the start of a new one was detected. One can argue that natural handwriting of a word does not entail separating in time the representation of characters — they are all 'connected'. One can also argue that their approach (delaying the decoder output by 1 sec and adding the filter kernel widths to the total interval) prevents visual feedback about the state of neural activity until a complete character is encoded by the subject, but the reality is that the subject can 'covertly' infer information from the structure of the word being typed (self-generated case) and visual feedback from the screen (on-prompt case).
Line 115: How did neural activity look like when an error was made? and when the subject was provided visual feedback about the language model's autocorrection of that error? Did the subject stop modulating, eventually relying on the model to autocorrect, or did he continue to modulate neural activity to correct the typo? Was the decoder 'disengaged' in those instances? did the neural activity occupy different regions of the state space relative to the intended character or the corrected character?
Line 118: It is stated that the raw decoder output plateaued at 90 characters per minute with a 5.4% character error rate. But the comparison drawn in the sentence that followed argues that the 'word error rate' decreased to 3.4% average across all days. The authors should provide the reduction in 'character error rate' not 'word error rate' with the use of the language model to make this comparison objective. Arguably, many words share the same characters and understanding of words depends on the sentence context.
Line 120: it is stated that 'a new RNN was trained using all available sentences to process an entire sentence'. This means that offline decoding of an entire sentence achieved the stated 0.17% character error rate. As stated this decoder has not been used by the subject in real time to see if this newly trained decoder will be able to display an entire sentence at the end of a neural activity modulation epoch by the subject in the absence of the delayed character-by-character feedback as in the online case. As such, what is the significance of this result?
Table 1: Can the authors explain why the word error rate is so high (25.1%) in the raw online output case despite a character error rate of 5.9%?
Supplementary material:
Line 427: it is stated that "some micromotions of the right hand were visible during attempted handwriting (see 10 for neurologic exam results and SVideo 4 for hand micromotions" Have authors quantified the extent of variance in the neural data that could be explained by this potential confound?

Line 491: It would be informative for the authors to comment on how did the neural activity differ between repetitions of each character individually and when they are within a word or a sentence.

D. Appropriate use of statistics and treatment of uncertainties:
Figures are well illustrated. Probability values and error bars are explained. There were no statistical significance tests performed.

Line 178: Authors should provide more explanation for "the participation ratio (PR), which quantifies approximately how many spatial or temporal axes are required to explain 80% of the variance in the neural activity patterns" in this section. Readers have to refer to the supplementary methods section to understand this metric.

Line 192 Figure 3: The authors find that increased temporal complexity in neural state space trajectories could make movements easier to decode compared to trajectories that do not have such complexity, or have only spatial complexity. They then present a toy example in Figure 3 to make this point. I would partly disagree with their assessment and argument for the following reasons:

i) In the toy example in (Figure 3F) they increased variations of neural trajectories over time to illustrate that this increases separability (measured by nearest neighbor distance) compared to the case where the neurons' activity is constrained to a single spatial dimension, the unity diagonal). But the example lacks inclusion of noise, the temporal characteristics of which can easily 'fool' the classifier, making it think there is more temporal complexity in the trajectories than really is.
ii) The nearest neighbor distance and consequently classifier performance should be characterized when noise is present in this toy example, with a parameter that controls the amount of temporal complexity in noisy neural trajectories. Directions of fluctuations around these trajectories are likely to influence the conclusion made, both in the straight line as well as the handwritten characters cases.

Line 244: Authors state that "One unique advantage of our handwriting BCI is that, in theory, it does not require vision (since no feedback of the imagined pen trajectory is given to the participant, and letters appear only after they are completed)." I would argue against that, partially because this claim is contingent on: 1) exact knowledge of the length of time interval where each decoded character is fully known and, 2) the instructed text was always present on the screen in the on-prompt case. To my understanding this was estimated (see my comment on Line 112 above) based on approximations made by the delayed decoder training and time warping algorithm (1.4 sec delay), which was used offline to build spatiotemporal neural "templates" of the characters.

Line 534: Please clarify what is a 'single movement condition'. Is it a character, a word or a sentence? From line 801 it seems it corresponds to character but the earlier sentence needs clarification.

Line 553: Authors used character templates drawn by a computer mouse in the same way as T5 described writing the character. This description provides a shape of the character but it is unclear how this information was translated into pen velocity to train the decoder.

Line 577: "the criteria for excluding data points from display in Figure 1E is not clear. It is stated that these data labeled as "outliers in each class" were excluded "To make the t-SNE plot clearer". While it is stated that this resulted in removing 3% of data points, the explanation that these "were likely caused by lapsed attention by T5" is not convincing. How did the authors ascertain that this was the case?

Supp Fig 2 and lines 642-667: The authors use a technique from automatic speech recognition literature called forced alignment labeling with HMMs in which they augmented the data via synthetic sentence generation to cope with the limited data size. This section needs improvement regarding how the method works. For example, creating snippets to make synthetic sentences assumes the neural data correspouding to each snippet is independent of the others. How it is then integrated into a new synthetic sentence that is then labeled by the HMM? How 'one-hot representation' is defined based on the heatmaps generated in SF-2D?

E. Conclusions

The conclusions are generally based on findings in the work performed in One subject. At times though there are some overstatements about the far reaching ability of the technology which should be scaled down. For example, I did not find the conclusion that this is a BCI without visual feedback to be convincing. If it were, then how can the authors explain the difference in performance between the on-prompt typing and self-paced typing? It is unclear whether there was any type of eye tracking to determine the type of visual feedback the subject was receiving at each moment. For example, was the subject always staring at the text prompt, or was the subject always looking to the decoded characters? Or a combination of both? unless they have an objective measure of visual feedback, it is unclear whether the BCI was truly operating without vision as claimed.

F. Suggested improvements:
In addition to the above, I think a critical experiment/analysis to be performed is one in which the authors characterize the longevity and stability of representation of neural signals of the decoded variable(s). The extensive calibration process indicates that the data is highly nonstationary but none of this is characterized. Based on a few published studies, it is reasonably expected that the implanted device can leverage single cell resolution of neural spiking signals within the first year of implant. However, authors used multiunit activity (binned threshold crossing), implying the activity could not be spike sorted to reveal individual neuronal activity encoding of the pen tip velocity. More explanation should be provided on how the nonuniform distribution of session dates affected the data quality. Authors explain in the supplementary material that this approach allowed them to "leverage information from more electrodes, since many electrodes recorded activity from multiple neurons that could not be precisely spike-sorted into single units." Although they cite a paper from their group that demonstrated that neural population structure can be accurately estimated from threshold crossing rates alone, it is unclear if sorting spikes from a lower number of electrodes (which they did not state) on which single units could be identified would provide similar results.

G. References: appropriate credit to previous work?
Mostly relevant and appropriate. The work could benefit from a few more citations that documented the idea of training decoders from 'desired' behavioral templates when overt movements could not be performed.

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions
No issues.

Author Rebuttals to Initial Comments:

Reply to Reviewers

Note: reviewers' comments appear in **black text**. Our replies appear in **blue text**, and revised manuscript text appears indented (with old text shown in **black** and new edits in **red**).

Overview:

We thank the reviewers for their careful read of the manuscript and their insightful and helpful suggestions. Most of the questions raised were requests for clarification, additional statistics, and/or reframing of certain results. We have addressed all these suggestions, which we believe has improved the presentation and rigor of the work. Point-by-point responses to each reviewer suggestion appear below this higher-level "Overview" section.

We appreciate the reviewers' unanimous recognition that this is a truly different and novel approach, with a substantial performance gain that is important for the field (and, one day hopefully, for patients as well). Three brief snippets might be helpful as it has been a while since reviewing the manuscript. Reviewer 1 (R1), "This paper represents a truly novel approach to restoring communication with a brain-computer interface." R2, "Overall, the manuscript is very well written and represents a clear and important advance in the field of BCIs." R3, "Outstanding features of the work are: Typing speeds of on-screen prompt at 90 characters per minute at 99% error rate with the use of a general-purpose autocorrect and 73.8 characters at 8.54% error rate for self-generated sentences (2.25% with a language model) are significant advances over the highest reported point and-click typing with a 2D computer cursor, peaking at 40 characters per minute. Results open a new approach for BCIs and demonstrate the feasibility of accurately decoding imagined handwriting movements years after losing ability to move and speak."

The most involved questions were raised by R3 with regards to the longevity and robustness of the intracortical BCI (iBCI). In particular, how long the neural signals can be expected to last and whether the neural signals change so quickly over time that extensive decoder retraining is required each day. With our new additions, we believe that we have addressed this question thoroughly and at a high standard, with the result being that our handwriting iBCI is right in line with other state-of-the-art iBCIs in terms of longevity and robustness. We outline below how we have addressed the longevity and robustness concerns; our additions include new discussion points as well as new data analyses that show the feasibility of achieving high-performance without requiring extensive daily decoder retraining.

Scope. Before explaining our new additions, we think it is important to first delineate what we see as the scope of this work. Any effective manuscript must have a well-defined (and necessarily limited) scope of investigation. We see this paper as being primarily focused on demonstrating the feasibility of decoding attempted handwriting movements from a person with tetraplegia well enough to substantially increase (i.e., double) communication rates while also maintaining high accuracy. Doing so opens the door to a promising new approach for iBCIs, as this is the first study to propose the fundamental idea of decoding attempted handwriting and to demonstrate that rapid sequences of attempted dexterous movements can be accurately decoded in a person who has been paralyzed for several years. However, by no means does our iBCI yet constitute a 'complete product' that would be appropriate for immediate clinical adoption, and we believe that meeting such a standard lies outside the scope of this work. Subsequent research in academia will be needed to further advance this system (e.g., just as several studies needed to follow the original Hochberg et al. *Nature* 2006 paper) and,

importantly, a truly corporate effort would be needed to fully ruggedize this, or any other, system for commercial medical use.

As suggested by R3, a final product would require systematic clinical trials that demonstrate both the longevity of the intracortical microelectrode arrays as well as decoder training algorithms that minimize (or eliminate) the need for daily decoder recalibration. To our knowledge, both "longevity" (the need to demonstrate device functionality over many years) and "robustness" (the need for less decoder retraining) are longstanding issues for intracortical BCIs, which no published manuscript has yet fully solved (but see below for reasons to be optimistic). As such, we see our work as providing important, and hopefully intellectually creative, motivation for academic researchers and companies to continue improving the longevity of intracortical arrays and designing new methods for minimizing decoder (re)calibration time. However, we do not see the complete resolution of these issues as within the scope of this study.

That said, we now outline how we have conducted extensive new data analyses and changed the manuscript to address the longevity and robustness issues to the best of our ability. We too are deeply interested in understanding these limits, so as to be most helpful to subsequent efforts.

Longevity. As R3 has noted, array longevity is a critical issue for any intracortical BCI. Before a product is taken to market, a systematic study must be conducted which demonstrates longevity across many subjects. While no such study has yet been published, preliminary results from several studies – including our own BrainGate clinical trials (NCT00912041) spanning 15 years and 14 participants – indicate that (Utah) arrays retain their functionality for several years in people; there are multiple examples of retained functionality for 1000+ days (Bullard et al., 2020; Simeral et al., 2011). Importantly, in the present study high performance was obtained 1200+ days post-implant. We added a new supplemental figure (now SFig. 6) to demonstrate that high-quality spiking activity is still present on many of the electrodes (on average 82 out of 192). In the Discussion, we now highlight the array longevity issue as well as reasons to be optimistic about array longevity.

Robustness. Second, as R3 and other reviewers have noted, minimizing decoder recalibration time is also an important problem for iBCIs (as well as non-invasive BCIs). This issue must also be addressed before a viable product can be taken to market, since users are not likely to tolerate long recalibration procedures each day. However, we see minimizing calibration time as a deep topic in and of itself, which has been the sole focus of several recent studies (Jarosiewicz et al., 2015; Dyer et al., 2017; Degenhart et al., 2020). Additionally, to our knowledge, daily decoder recalibration is still standard practice in the iBCI field and many important papers have used this method (e.g., Hochberg et al., 2006, 2012; Collinger et al., 2013; Bouton et al., 2016; Ajiboye et al., 2017). We think it is therefore reasonable to leave this aspect of handwriting decoding to be more fully investigated in future work. *Nevertheless*, we agree that it is important to both (1) more clearly highlight this issue in the manuscript, and (2) do whatever analyses we can to address it while still remaining reasonably within scope.

To that end, **we have added a new figure to the main text (now Fig. 3)** that reports results from offline analyses estimating how much data was actually needed for daily decoder retraining. Encouragingly, the results suggest that high performance would have been possible with only 10 sentences of data per day (as opposed to the 50 sentences per day that were originally used). We also report promising results from a new unsupervised method, that we

introduce here in the revised manuscript thanks to the Reviewers' questions, that uses a language model to retrain the decoder without requiring any explicit data labels. This could enable decoder retraining to occur in the background, as a parallel computational process making use of newly incoming data, without interrupting the person's iBCI use. We believe these analyses show promise that it should be possible to achieve high performance with unsupervised retraining, combined with smaller amounts of supervised data after long periods of not using the iBCI. This points the way towards a handwriting iBCI that can achieve high performance while minimizing user interruptions.

We also added a new supplemental figure (now SFig. 4) that assess the stability of the neural patterns associated with each character over time, since this is a critical issue that ultimately determines how much data is needed for daily decoder recalibration. We found high short-term stability (mean correlation = 0.85 when 7 days apart or less), and neural changes that seemed to accumulate at a steady and predictable rate. Again, this is promising for the possibility of recalibrating decoders with limited amounts of data (or even in an unsupervised manner without interrupting the user).

Future Work. Although we cannot fully resolve the longevity and robustness issues in this current manuscript, we do want the Reviewers and Editors to know that we appreciate the importance of these issues in general. As such, we thought it might be helpful to share that we are currently in the process of writing a separate manuscript summarizing array safety and longevity data from all 14 participants of the BrainGate pilot clinical trial (collected over a span of 15 years), which will be the first systematic study of its kind in people. We think that this kind of a study is a better forum for more fully resolving these issues than what this current manuscript can do, which we think should remain focused on laying out and demonstrating an entirely new kind of iBCI and associated methods.

References

Ajiboye, A.B., Willett, F.R., Young, D.R., Memberg, W.D., Murphy, B.A., Miller, J.P., Walter, B.L., Sweet, J.A., Hoyen, H.A., Keith, M.W., Peckham, P.H., Simeral, J.D., Donoghue, J.P., Hochberg, L.R., Kirsch, R.F., 2017. Restoration of reaching and grasping movements through brain-controlled muscle stimulation in a person with tetraplegia: a proof-of-concept demonstration. The Lancet 389, 1821–1830. https://doi.org/10.1016/S0140-6736(17)30601-3

Bouton, C.E., Shaikhouni, A., Annetta, N.V., Bockbrader, M.A., Friedenberg, D.A., Nielson, D.M., Sharma, G., Sederberg, P.B., Glenn, B.C., Mysiw, W.J., Morgan, A.G., Deogaonkar, M., Rezai, A.R., 2016. Restoring cortical control of functional movement in a human with quadriplegia. Nature 533, 247–250. https://doi.org/10.1038/nature17435

Bullard, A.J., Hutchison, B.C., Lee, J., Chestek, C.A., Patil, P.G., 2020. Estimating Risk for Future Intracranial, Fully Implanted, Modular Neuroprosthetic Systems: A Systematic Review of Hardware Complications in Clinical Deep Brain Stimulation and Experimental Human Intracortical Arrays. Neuromodulation Technol. Neural Interface 23, 411–426. https://doi.org/10.1111/ner.13069

Collinger, J.L., Wodlinger, B., Downey, J.E., Wang, W., Tyler-Kabara, E.C., Weber, D.J., McMorland, A.J., Velliste, M., Boninger, M.L., Schwartz, A.B., 2013. High-performance neuroprosthetic control by an individual with tetraplegia. The Lancet 381, 557–564. https://doi.org/10.1016/S0140-6736(12)61816-9

Degenhart, A.D., Bishop, W.E., Oby, E.R., Tyler-Kabara, E.C., Chase, S.M., Batista, A.P., Yu, B.M., 2020. Stabilization of a brain–computer interface via the alignment of low-dimensional spaces of neural activity. Nat. Biomed. Eng. 1–14. https://doi.org/10.1038/s41551-020-0542-9

Dyer, E.L., Gheshlaghi Azar, M., Perich, M.G., Fernandes, H.L., Naufel, S., Miller, L.E., Körding, K.P., 2017. A cryptography-based approach for movement decoding. Nat. Biomed. Eng. 1, 967–976. https://doi.org/10.1038/s41551-017-0169-7

Hochberg, L.R., Bacher, D., Jarosiewicz, B., Masse, N.Y., Simeral, J.D., Vogel, J., Haddadin, S., Liu, J., Cash, S.S., Smagt, P. van der, Donoghue, J.P., 2012. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. Nature 485, 372–375. https://doi.org/10.1038/nature11076

Hochberg, L.R., Serruya, M.D., Friehs, G.M., Mukand, J.A., Saleh, M., Caplan, A.H., Branner, A., Chen, D., Penn, R.D., Donoghue, J.P., 2006. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. Nature 442, 164–171. https://doi.org/10.1038/nature04970

Jarosiewicz, B., Sarma, A.A., Bacher, D., Masse, N.Y., Simeral, J.D., Sorice, B., Oakley, E.M., Blabe, C., Pandarinath, C., Gilja, V., Cash, S.S., Eskandar, E.N., Friehs, G., Henderson, J.M., Shenoy, K.V., Donoghue, J.P., Hochberg, L.R., 2015. Virtual typing by people with tetraplegia using a self-calibrating intracortical brain-computer interface. Sci. Transl. Med. 7, 313ra179-313ra179. https://doi.org/10.1126/scitranslmed.aac7328

Simeral, J.D., Kim, S.-P., Black, M.J., Donoghue, J.P., Hochberg, L.R., 2011. Neural control of cursor trajectory and click by a human with tetraplegia 1000 days after implant of an intracortical microelectrode array. J. Neural Eng. 8, 025027. https://doi.org/10.1088/1741-2560/8/2/025027

Point-by-point responses to referee #1

A. Summary of the key results
This paper represents a truly novel approach to restoring communication with a brain-computer interface. Previous approaches have used point-and-click cursor control to enable communication with an onscreen keyboard and have demonstrated very good performance that enables functional performance. Here, the authors instead try to decode handwriting movements in order to predict individual letters as the brain-computer interface (BCI) users imagines writing words and sentences. Impressively, online BCI performance was more than twice as fast as previously demonstrated and approaches smartphone typing speeds. Further, the authors demonstrate that the temporal variability associated with handwriting trajectories is a major contributor to the high level of performance that was shown, which has implications for BCIs in general as it may be advantageous to try to decode complex and dexterous movements.

B. Originality and significance:
This study takes a new and original approach to BCI-controlled communication by decoding attempted handwriting movements in order to enable computer-based communication. This approach is unique because rather than decoding the movement trajectory directly (although they demonstrate that this is possible), they implement a two-step classification process using an RNN to identify when the user is attempting to write a character and then determining which character the user is trying to write. The decoding approach relies on both the spatial and temporal variability of the attempted movements to boost performance far beyond what has previously been demonstrated for BCI-based communication.

This work provides evidence that an intracortical BCI can enable fast rates of communication based on decoded handwriting patterns. This work is therefore of interest to scientists and engineers developing neural interfaces to restore communication as well as clinicians working with patients with communication impairments.

We are gratified that the reviewer expresses that this is a truly novel approach that makes significant gains in intracortical brain-computer interface (iBCI) performance. We thank the reviewer for their thorough read of the manuscript and insightful and helpful questions and suggestions.

C. Data & methodology:

1) This paper is well written and clearly describes the key details and decision points that were used to implement the RNN-based decoding approach. The figures highlight key methodologicial elements and results. A rigorous approach was taken to investigate the impact of various optimization parameters, data quantity, and data quality (vs. noise). All data and code will be made publicly available providing an extremely valuable resources for the research community as well as transparency in reporting.

Thank you for noting the methodological rigor and the value of the data & code release, which we too think will help the research community improve upon what we have done and apply our methods to new problems.

2) Performance metrics are appropriate and the details of how each was calculated are included.

D. Appropriate use of statistics and treatment of uncertainties:

1) All data are presented from a single subject across multiple data sessions. This is appropriate given the limited number of human participants that have been implanted with an intracortical BCI, the rigor of the approach, and the importance of the findings.

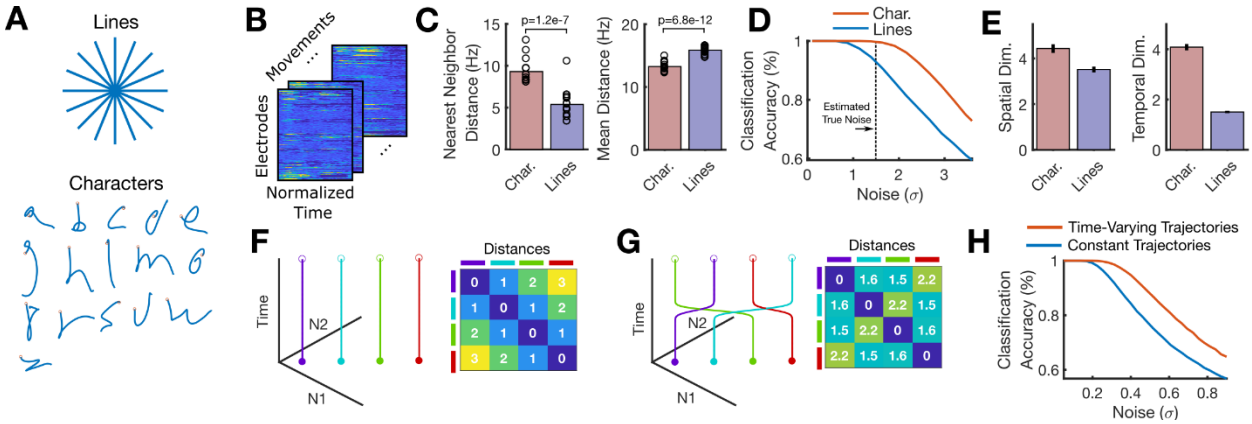
Thank you for explicitly noting that one subject is appropriate for this type of study. We too believe this to be the case.

2) Statistical tests should be performed to compare between the character and lines conditions for data shown in Figure 3C and E and reflected in the manuscript and figure.

Thank you for this suggestion. We now report 95% confidence intervals for the effects shown in Fig. 3C and 3E (which is now Fig. 4). We believe that confidence intervals keep the focus on the effect sizes, while also demonstrating statistical significance. Confidence intervals were generated by jackknife. Below is a snippet from the main text where the confidence intervals were added:

> First, we analyzed the pairwise Euclidean distances between each neural activity pattern. We found that the nearest-neighbor distances for each movement were ~~almost twice as large~~ 72% larger for characters as compared to straight lines (95% CI = [60%, 86%])~~(72% larger)~~, making it less likely for a decoder to confuse two nearby characters (Fig. 4~~3~~C).
>
> …
>
> We found that ~~the~~ spatial dimensionality was ~~similar for straight lines and characters~~only modestly larger for characters (~~Fig. 3E~~1.24 times larger; 95% CI = [1.11, 1.37]), but that the temporal dimensionality was much greater (~~more than twice~~ 2.65 times larger; ~~as large for characters~~ 95% CI = [2.63, 2.68]), suggesting that the increased variety of temporal patterns in letter writing drives the increased separability of each movement (Fig. 4E).

Using two-sample t-tests, we also now report p-values for Fig. 4C (see below). The t-tests compare the means of the distributions shown (n=16).



Finally, it does not seem straightforward to apply hypothesis testing to 4E, since temporal and spatial dimensionalities are complex functions of the data that do not appear to have standard tests or null distributions. We believe that the 95% confidence intervals shown on the bars (computed via jackknife), as well as the new confidence intervals for the dimensionality ratios mentioned above, sufficiently demonstrate statistical significance.

3) While many comparisons are made based on qualitative results or comparisons of confidence intervals, the effects and improvements over previous methods are large and robust. Further statistical analysis is not needed to support the conclusions.

Thank you.

E. Conclusions:

1) The major conclusions are robustly supported by the presented data with consistent performance achieved across multiple sessions

Thank you.

2) The limitations section should mention that this work comes from a single subject who had the ability to

write prior to his injury.

Thank you, we have now added this limitation to the Discussion section:

> Finally, it is important to recognize that ~~our~~ the current system is a proof-of-concept that a high-performance handwriting BCI is possible (in a single participant capable of handwriting prior to his injury); it is not yet a complete, clinically viable system.

3) The authors conclude that a handwriting BCI is the first type of BCI that has the potential to work in people with visual impairments. This was not evaluated in the present study. While the subject did not have feedback of BCI performance until after each letter was selected, this did provide feedback that could be accumulated over the course of the session. Additionally, the importance of this was not made clear. Other forms of feedback (auditory, tactile, etc.) could be used to convey information to a person with visual impairments. Further, it is a very small population that is impacted by both visual and communication impairments.

Thank you for raising these important limitations. We agree, and now no longer discuss our iBCI's potential to work in people with visual impairments. While we did collect some data demonstrating good performance with his eyes closed that could be added, it is not a major point and we believe that it is better to remove it to help the manuscript stay focused.

F. Suggested improvements and comments:

1) Results, line 45: specify that the participant had a cervical spinal cord injury and be more precise in the description of residual movement abilities.

The description now reads:

> T5 has a high-level spinal cord injury (C4 AIS C) and was paralyzed from the neck down; his hand movements were entirely non-functional and limited to twitching and micromotion.

Also, note that in the Methods section we refer to T5's neurologic exam data that was recently published as part of a different paper (Willett et al. *Cell* 2020, cited below). We have added more detail to the Methods section which now reads as follows:

> Below the injury, T5 retained some very limited voluntary motion of the arms and legs that was largely restricted to the left elbow; however, some micromotions of the right hand were visible during attempted handwriting (see [12] for full neurologic exam results and SVideo 4 for hand micromotions). T5's neurologic exam findings were as follows for muscle groups controlling the motion of his right hand: Wrist Flexion=0, Wrist Extension=2, Finger Flexion=0, Finger Extension=2 (MRC Scale: 0=Nothing, 1=Muscle Twitch but no Joint Movement, 2=Some Joint Movement, 3=Overcomes Gravity, 4=Overcomes Some Resistance, 5=Overcomes Full Resistance).

[12] Willett FR, Deo DR, Avansino DT, Rezaii P, Hochberg LR, Henderson JM, Shenoy KV (2020) Hand Knob Area of Premotor Cortex Represents the Whole Body in a Compositional Way. Cell 181:396–409.

2) Results, line 60: why was a non-linear approach (t-SNE) selected for data visualization and separability analysis given that PCA allowed for accurate trajectory reconstruction. Readability would be improved by understanding the intuition that guided this decision.

Thank you for raising this lack of clarity. The difference is that the trajectory reconstruction was performed on the trial-averaged data (which averages and thereby reduces noise), while t-SNE was applied to single trials (which inevitably have considerable noise). The advantage of t-SNE (as compared to a method like PCA) is that t-SNE is designed to accurately portray the separability of high-dimensional clusters in the presence of noise, while PCA on single trials will often show highly overlapping clusters in low-

dimensional space that are highly separable in the full-dimensional space. To make this clearer, we now emphasize more clearly that the trajectory reconstruction was performed on trial-averaged data, while t-SNE was applied to single trials (originally this was spelled out only in the figure legend):

> To see if the neural activity encoded the pen movements needed to draw each character's shape, we attempted to reconstruct each character by linearly decoding pen tip velocity from the trial-averaged neural activity (Fig. 1D). Readily recognizable letter shapes confirm that pen tip velocity is robustly encoded.
> ….
> Finally, we used a nonlinear dimensionality reduction method (t-SNE) to produce a 2-dimensional visualization of each single trial's neural activity recorded during a 1 s window after the 'go' cue was given (Fig. 1E).

3) Results, line 63: Please provide a confidence interval (or similar measure of variability) for the k-nearest neighbor classification result.

A confidence interval is now provided (binomial proportion confidence interval, Clopper-Pearson method):

> Using a k-nearest neighbor classifier applied to the neural activity, we could classify the characters with 94.1% accuracy (95% CI = [92.6, 95.8], chance level = 3.2%).

4) Results, lines 95-106: It is important to note that a large amount of training data needed to be collected each day. In addition to reporting the number of sentences, the authors should report the number of characters and duration of data collection in the main text. It is noted that this information is included in the Supplemental Material. Additionally it wasn't clear from the main text that "…data was cumulatively added to the training dataset…" referred to data collected prior to BCI control, rather than just adding in data as it was collected during BCI assessment.

Thank you for raising this important point. We have added the above-requested details and rephrased the main text to clarify how the training data were used. The description now reads:

> Prior to the first day of real-time use described here, we collected a total of 242 sentences across 3 days that were combined to train the RNN (sentences were selected from the British National Corpus). On each day of real-time use, additional training data were collected to retrain the RNN prior to real-time evaluation, yielding a combined total of 572 training sentences by the last day (comprising 7.3 hours and 30.4k characters). ~~After each new day of decoder evaluation, that day's data was cumulatively added to the training dataset for the next day (yielding a total of 572 sentences by the last day).~~

In addition, based on feedback from the other reviewers, we have added new offline analyses that estimate how much data were actually needed for daily decoder retraining. The results suggest that high performance is possible with only 10 sentences of data per day (as opposed to the 50 sentences per day that were originally used). We also report results from a new unsupervised method that uses a language model to retrain the decoder without requiring any explicit data labels; this could enable decoder retraining to occur in the background without interrupting the user for data collection. We believe these analyses both (1) draw important attention to this issue, and (2) show promise that it may be possible to achieve high performance with unsupervised retraining (combined with smaller amounts of supervised data after long periods of not using the iBCI). This points the way towards a handwriting iBCI that can achieve high performance while minimizing user interruptions.

For convenience, this new Results section is reproduced below:

> Following standard practice for BCIs (e.g. [1,2,19,4,5]), we retrained our handwriting decoder each day before evaluating it, with the help of "calibration" data collected at the beginning of each day. Retraining helps account for changes in neural recordings that accrue over time. Ideally, to reduce the burden on the user, little or no calibration data would be required. In a retrospective analysis of the copy typing data reported

above in Fig. 2, we assessed whether high performance could still have been achieved using less than the original 50 calibration sentences per day (Fig. 3A). We found that 10 sentences (8.7 minutes) were enough to achieve a raw error rate of 8.5% (1.7% with a language model), although 30 sentences (26.1 minutes) were needed to match the raw online error rate of 5.9%.

However, our copy typing data were collected over a 28-day time span, possibly allowing larger changes in neural activity to accumulate. We therefore tested whether more closely-spaced sessions reduce the need for calibration data (Fig. 3B), using an offline analysis of copy typing data across 8 sessions. We found that when only 2-7 days passed between sessions, performance was reasonable with *no* decoder retraining (11.1% raw error rate, 1.5% with a language model). Finally, we tested whether decoders could be retrained in an unsupervised manner by using a language model to error-correct and retrain the decoder, thus bypassing the need to interrupt the user for calibration (i.e. by recalibrating automatically during normal use). Encouragingly, unsupervised retraining achieved a 7.3% raw error rate (0.84% with a language model) when sessions were separated by 7 days or less (see Methods & Supplemental Methods for details). Ultimately, whether decoders can be successfully retrained with minimal recalibration data depends on how quickly the neural activity changes over time. We assessed the stability of the neural patterns associated with each character and found high short-term stability (mean correlation = 0.85 when 7 days apart or less), and neural changes that seemed to accumulate at a steady and predictable rate (SFig. 4 provides a quantitative visualization). The above results are promising for clinical viability, as they suggest that unsupervised decoder retraining, combined with more limited supervised retraining after longer periods of inactivity, may be sufficient to achieve high performance.

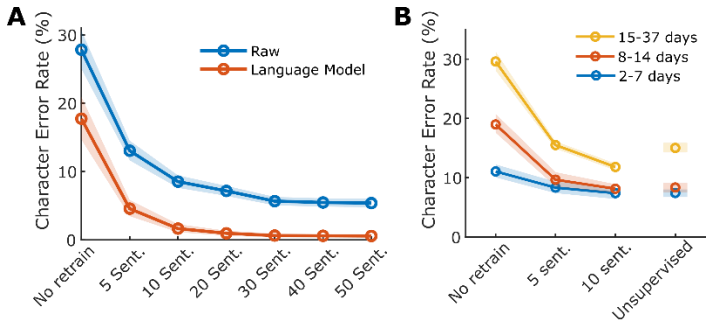


Figure 3. Performance remains high when decoder retraining is limited or omitted. (A) To account for neural activity changes that accrue over time, we retrained our handwriting decoder each day before evaluating it. Here, we simulate offline what the decoding performance shown in Fig. 2 would have been if less than 50 calibration sentences were used. Lines show the mean error rate across all data and shaded regions indicate 95% CIs (computed via bootstrap resampling of single trials, N=10,000). **(B)** Copy typing data from eight sessions were used to assess whether less calibration data are required if sessions occur closer in time. All session pairs (X, Y) were considered. Decoders were first initialized using training data from session X and earlier, and then evaluated on session Y under different retraining methods (no retraining, retraining with limited calibration data, or unsupervised retraining). The average raw character error rate is plotted for each category of time elapsed between sessions X and Y, and for each retraining method. Shaded regions indicate 95% CIs.

5) Results, figure 2C. It is interesting that day 1237 seems to have a higher character error rate that interrupts what appears to be a linear increase in error rate that is mirrored by an increase in characters per minute. Is there a reason for this? Across the 5 sessions, did the participant have a change in strategy (e.g. to go faster with less regard for error?).

Day 1237 does indeed seem to be an outlier, but we don't have a strong reason to suspect a particular cause. T5 was always instructed to go as fast as possible; anecdotally, he reported becoming more comfortable with going faster over time, as he gained confidence that the system would work accurately at higher speeds. There is indeed a modest increase in error rate over time (we observed an error rate of 4.3% on the first day and 5.4% on the last day). We speculate that it is easier to classify at slower speeds

because of an increased amount of neural data per character that can be used to decide that character's identity, effectively increasing the overall SNR of the data available to the decoder.

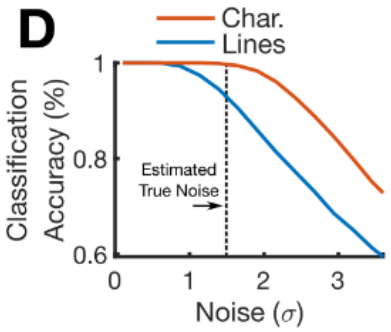
6) Results, Table 1: For clarity, I suggest renaming the second row "online output + offline language model".

Thank you for this suggestion. We have reformatted the table as follows:

	Character Error Rate (%) [95% CI]	Word Error Rate (%) [95% CI]
Online output	5.9 [5.3, 6.5]	25.1 [22.5, 27.4]
Online output + offline language model	0.89 [0.61, 1.2]	3.4 [2.5, 4.4]
Offline bidirectional RNN + language model	0.17 [0, 0.36]	1.5 [0, 3.2]

7) Results, Lines 166-174: How do the values chosen for simulated neural noise compare the variability in feature means that were observed in the experiment?

We have now annotated the figure with an estimate of the true noise level in the recorded neural features:



This estimate was generated by computing the neural population distance of each single trial from the class mean, along neural dimensions that connect each class to each other class (thus ignoring dimensions that are irrelevant for classification).

8) Results, Lines 178-183 & Figure 3E: While the effect of temporal dimensionality is more striking, spatial dimensionality is also likely statistically different between the characters and straight lines. This statement may therefore be too strong: "We found that the spatial dimensionality was similar for straight-lines and characters (Fig. 3E)."

Thank you for pointing this out. We now report the results as follows, which do indeed reveal a modest (but statistically significant) increase in spatial dimensionality for characters:

> We found that ~~the~~ spatial dimensionality was ~~similar for straight-lines and characters~~only modestly larger for characters (~~Fig. 3E~~1.24 times larger; 95% CI = [1.11, 1.37]), but that the temporal dimensionality was much greater (~~more than twice~~ 2.65 times larger; ~~as large for characters~~ 95% CI = [2.63, 2.68]), suggesting that the increased variety of temporal patterns in letter writing drives the increased separability of each movement (Fig. 4E).

9) Results- suggestions for additional data presentation:

a) Did the subject provide any subjective feedback about ease of use, training duration, suggestions for improvements, etc?

One of the most interesting things T5 described to us was his own experience of what it felt like to 'attempt' to handwrite. T5 imagined that he was holding a pen in his hand. As he attempted to write each letter, he reported having the subjective experience of feeling as though the pen was actually moving and tracing out the letter shapes (even though the actual motion of his hand was very limited, and he was not holding a pen). He sometimes reported being reticent about writing more quickly, because this could cause the subjective experience to lose clarity. He also reported that this experience seemed to follow physical constraints, because he was able to 'write' more quickly if he attempted to write in a smaller font. We now mention this subjective experience in the Results section, which reads:

> We instructed T5 to 'attempt' to write as if his hand was not paralyzed (while imagining that he was holding a pen on a piece of ruled paper). T5 reported having the subjective experience of feeling as though the imaginary pen was moving and tracing out the letter shapes.

Regarding suggestions for improving the BCI, T5 did not have much to say. Mostly, T5 was happy and somewhat amazed that the BCI could figure out what he was writing and show it to him on the screen. T5 reported feeling like he wasn't making 'clear' or 'legible' movements, and so he was surprised at how consistently the BCI could decode what he was trying to write.

b) Had the subject previously used a point-and-click communication BCI?

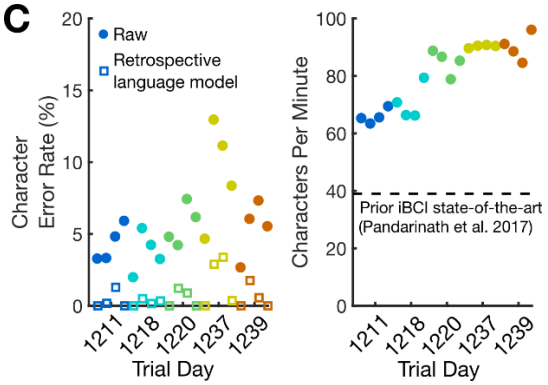
Yes, T5 set the prior record for BCI communication using a point-and-click BCI in one of our prior publications (Pandarinath et al. *eLife* 2017, cited below). We now explicitly mention this in the Results section:

> For intracortical BCIs, the highest performing method has been point-and-click typing with a 2D computer cursor, peaking at 40 characters per minute [7] (this record was also set by participant T5 3 years earlier; see SVideo 3 for a direct comparison).

Pandarinath C, Nuyujukian P, Blabe CH, Sorice BL, Saab J, Willett FR, Hochberg LR, Shenoy KV, Henderson JM (2017) High performance communication by people with paralysis using an intracortical brain-computer interface. Elife 6 Available at: http://dx.doi.org/10.7554/eLife.18554.

c) Was there any notable change in performance within a session?

Thank you for this interesting suggestion. Fig. 2C (reproduced below) shows a relatively flat character error rate within each session (as can be assessed by looking at the four circles from each session, each of which corresponds to a single block of sentences). Nevertheless, there does appear to be a modest (but potentially statistically significant) increase in error near the end of each session.



Since the first and last block contain the same seven 'comparison' sentences which we collected for a direct comparison to prior work (Pandarinath et al. 2017), we can directly compare the difference in error between the first and last block of each session. Pooling all the data together reveals an increase in error

rate of 2.4% (95% CI = [0.63, 4.2]). However, it is difficult to know whether this increase in error is due to small changes in neural activity which accrue over time, or due to the participant's fatigue after a long session. We think that an interesting area of future work could attempt to iteratively adjust the decoder to account for neural changes after each sentence is decoded (using unsupervised retraining), to see if this prevents the error rate from increasing. However, as there is already a lot to discuss in the paper, we think it is best to tackle this issue in a separate study.

d) The authors state that the language model is capable of running in real-time. If this is the case, why wasn't this done? With the data presented, the major outcome that should be reported in the abstract is the fully-online performance with notes about how this can be improved offline.

Although the language model is theoretically quite capable of running in real-time, the software engineering and development needed to implement this would have required a large amount of effort that we felt was not germane to the core scientific questions on which we were focused. Although a real-time demonstration of the language model is potentially compelling as a demonstration, we felt that this portion of an eventual clinical system would best be left to experts in language modeling and future work.

Nevertheless, we do agree that the online results should be reported in the abstract and given precedence. We have changed the abstract to read as follows:

> With this BCI, our study participant, ~~(~~whose hand was paralyzed from spinal cord injury,~~)~~ achieved typing speeds that exceed those of any other BCI yet reported: 90 characters per minute at 94.1% raw accuracy online, and >99% accuracy offline with a general-purpose autocorrect.

10) Discussion, line 252: This sentence states that the subject's hand never movement, but a video is shown to highlight the micromotions. Was the subject intending to trace the letter trajectory, even if his injury likely limited his ability to do so accurately?

Yes, the participant was *attempting* to handwrite each letter (see our longer response to this issue below under "H. Clarity and context"). Note that although the participant could generate twitches/micromotions, his hand was severely paralyzed and retained no useful function. We changed this sentence to read:

> To achieve high performance, we developed new decoding methods to overcome two key challenges: (1) lack of observable behavior during long sequences of self-paced training data (our participant's hand ~~never moved~~was paralyzed), and (2) limited amounts of training data.

11) Methods, lines 689-692: Additional detail about the linear transformation and process of fitting separate input layers each day should be stated here, or clearly linked to the supplemental methods. The supplemental figure alone is not sufficient for understanding these steps.

This section now links clearly to the supplemental methods:

> To account for differences in neural activity across days [11,13], we separately transformed each days' neural activity with a linear transformation that was simultaneously optimized with the other RNN parameters (see Supplemental Methods, "Combining Data Across Days" section).

G. References:
References are appropriate. The only comment is with regard to Reference 24 that is cited to show that EEG-BCI has achieved speeds of 60 characters per minute. This is a generous statement and other limitations could be noted given that that level of performance is not typical and was obtained from some healthy subjects due cued typing. This is a minor point.

Thank you, and indeed we wanted to, if anything, err in the direction of being generous. But we agree, and we have updated this discussion section to be more comprehensive. It now reads as follows:

> Commonly used BCIs for restoring communication to people who can't move or speak are either flashing EEG spellers[14–19] or 2D point-and-click computer cursor-based BCIs for selecting letters on a virtual keyboard [20,13,3]. Typical EEG spellers based on P300s or motor imagery achieve 1-5 characters per minute in people with paralysis [14–16,18,19]. EEG spellers that use visually evoked potentials have achieved speeds of 60 characters per minute [17] in able-bodied people, but have important usability limitations, as they tie up the eyes, are not typically self-paced, and require panels of flashing lights on a screen that take up space and may be fatiguing.

H. Clarity and context:
1) In the abstract, results, and discussion, the authors refer to the subject as being completely paralyzed below the neck and that he performed "imagined" hand movements. However, they note that the subject retained some movement of his shoulders and that he had micromotions of his hand during the handwriting task. It would be more appropriate to describe any residual function in the subject's arm and hand. Additionally, the authors should clarify if the subject was imagining the movements or attempting them (resulting in micromotions). See for example previous work from this group: Rastogi, A., Vargas-Irwin, C.E., Willett, F.R. et al. Neural Representation of Observed, Imagined, and Attempted Grasping Force in Motor Cortex of Individuals with Chronic Tetraplegia. Sci Rep 10, 1429 (2020).

Thank you for raising this lack of clarity. Our participant was *attempting* to handwrite each letter, thus resulting in micromotion of the paralyzed hand. Although the participant was imagining that he was holding a pen over a piece of paper, the movement itself is better described as attempted instead of imagined. We chose to instruct attempted movement instead of imagined movement because, as you note, prior work has demonstrated that attempted movement evokes stronger neural activity than purely imagined movement. The following sentence in the first paragraph of the Results describes the movement as attempted:

> We instructed T5 to 'attempt' to write as if his hand was not paralyzed (while imagining that he was holding a pen on a piece of ruled paper).

We have also substituted all instances of the word 'imagined' with 'attempted' throughout the manuscript. In the title, we have simply removed the word 'imagined'. The title now reads: "High-performance brain-to-text communication via handwriting". We felt that including the phrase "attempted handwriting" in the title may confuse readers who are not in the BCI field since, to our knowledge, "attempted movement" is BCI-specific jargon.

2) The abstract should report the typing speeds and accuracy that were achieved completely online without the language model since that is most representative of actual performance. It would be appropriate to also include results with offline enhancements as these would be acceptable in many contexts (such as writing an email).

Thank you for pointing this out, we agree and have changed the abstract to read as follows:

> With this BCI, our study participant, ~~(~~whose hand was paralyzed from spinal cord injury,~~)~~ achieved typing speeds that exceed those of any other BCI yet reported: 90 characters per minute at 94.1% raw accuracy online, and >99% accuracy offline with a general-purpose autocorrect.

We envision that in a final system, a language model could even be integrated into the BCI itself and thus used for all applications (much like speech recognition systems which rely heavily on language modeling). Thus, we think it is important and relevant to report the results with a language model applied.

Point-by-point responses to referee #2

Willett et al. present an intracortical BCI (iBCI) decoding approach for classifying many characters to enable rapid typing. Their approach uses an RNN architecture to perform classification on neural activity as the subject imagines writing letters/words/sentences. They achieve typing speeds up to 90 characters per minute with above 94.5% accuracy in one subject, which significantly outperforms previous communication iBCIs. They demonstrate the system works across several sessions and both for copying text and free expression. The authors further provide analyses to provide intuition for why their approach succeeds--they achieve high classification accuracy by having the user perform a task that generates highly discriminable neural activity.

Overall, the manuscript is very well written and represents a clear and important advance in the field of BCIs. The technical innovations of the paper include 1) methods for creating training datasets when there is minimal available information (since the subject imagined moving) and 2) methods for leveraging the power of RNNs even with relatively limited data. The approaches for challenge 2 primarily use techniques common in ANNs (data augmentation) and techniques previously shown to be useful in animal studies (adding external noise to increase robustness of the networks). The solutions to challenge 1 appear relatively novel, and are certainly new to the field of BCIs. The approach/conceptual innovation of the paper is a shift away from decoding continuous control towards a method that provides accurate classification even for a relatively large 31 character set. To my knowledge this is a notable departure from prior work.

We are gratified that the reviewer expresses that this is an important advance for BCIs, and that the training methods and approach is genuinely novel. We thank the reviewer for their thorough read of the manuscript and their insightful and helpful suggestions.

My primary concern with the manuscript is how the author's frame the work's overall approach which should more clearly emphasize the shift towards classification. As their work demonstrates, this shift can be powerful but it is also very specialized to this task. The manuscript's current comparisons to previous state-of-the-art (Pandarinath et al.) and figure 3 fail to fully make the distinction between continuous decoding of a cursor for selecting keys on a keyboard from their BCI performing a 31-way classification. Figure 3, for instance, almost implies that Pandarinath and prior BCIs were trying to classify straight line movements, which they did not. The authors' point that discriminability of the neural activity patterns directly impacts classifier performance is well taken. And provides an intuition for why having users imagine writing letters enabled their advance. But the manuscript needs to be very clear that in and of itself does not explain why they achieve higher performance. It explains why they were able to classify a large alphabet successfully for the first time. They then achieve higher performance compared to prior work because their classifier can predict letters more quickly than the average translation + click time of continuous control cursor tasks. The primary reason I emphasize this distinction is that their classification approach solves the problem of typing quite well, but does not provide a mechanism that necessarily generalizes to other tasks that are more continuous in nature like controlling a robotic limb (the authors do not claim this, but I think it's important the paper itself makes this distinction more clearly).

Thank you for pointing out this potential point of confusion. Indeed, the idea of improving classification performance by increasing the temporal dimensionality of the decoded movements is specific to discrete BCIs (and thus does not apply to BCIs that restore continuous motion). Additionally, your point is well taken that the increased decodability of handwritten letters is not necessarily the only reason why the handwriting BCI was able to go twice as fast as a point-and-click BCI. However, we theorize that it is *one* important factor that enabled the speed increase.

Fundamental to our argument is the idea that the speed of a point-and-click BCI is limited primarily by decoding accuracy. During parameter optimization of point-and-click BCIs, the cursor gain (speed scaling parameter) is typically increased as much as possible to increase typing speed, until it reaches a point where the cursor becomes uncontrollable due to decoding errors that push the cursor around randomly [1]. Thus, we do believe it is important to ask: how is it that our handwriting BCI was able to achieve similar levels of decoding accuracy (mid-90s) while going twice as fast? In other words, why couldn't the

point-and-click BCI go twice as fast as it did? Why did accurate point-and-click movement become essentially impossible at only half the speed of the handwriting BCI?

Our explanation is that different handwritten characters are easier to tell apart from each other than different straight-line movements. It is important to note that there is still a large gap between the performance of continuous cursor BCIs and able-bodied movement, suggesting that point-and-click BCIs are still limited primarily by decoding accuracy and not by fundamental behavior limits. Consistent with this idea, data on 1-finger typing on smartphones shows that the average typing rates are much higher than what was achieved in (our) Pandarinath et al. *eLife* 2017 publication (40 characters per minute vs. 120+) [2], further suggesting that point-and-click BCI speeds are not limited by fundamental brain/behavior limits on point-to-point movement and click/selection. Finally, we note that letters have more movement segments than a straight-line movement does (several letters have multiple straight lines in them). Despite this, handwriting movements could be decoded at greater speeds than point-to-point movements, which also suggests that point-and-click BCI speeds have not yet approached the fundamental limit of human behavior.

[1] Willett, Francis R., Brian A. Murphy, William D. Memberg, Christine H. Blabe, Chethan Pandarinath, Benjamin L. Walter, Jennifer A. Sweet, et al. "Signal-Independent Noise in Intracortical Brain–Computer Interfaces Causes Movement Time Properties Inconsistent with Fitts' Law." *Journal of Neural Engineering* 14, no. 2 (2017): 026010. https://doi.org/10.1088/1741-2552/aa5990.

[2] Palin, Kseniia, Anna Maria Feit, Sunjun Kim, Per Ola Kristensson, and Antti Oulasvirta. "How Do People Type on Mobile Devices? Observations from a Study with 37,000 Volunteers." In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, 1–12. MobileHCI '19. Taipei, Taiwan: Association for Computing Machinery, 2019. https://doi.org/10.1145/3338286.3340120.

To better explain this point, we added additional text to the motivating paragraph of the Results section, which now reads as follows:

> To our knowledge, 90 characters per minute is the highest typing rate yet reported for any type of BCI (see Discussion). For intracortical BCIs, the highest performing method has been point-and-click typing with a 2D computer cursor, peaking at 40 characters per minute [7] (this record was also set by participant T5 3 years earlier; see SVideo 3 for a direct comparison). The speed of point-and-click BCIs is limited primarily by decoding accuracy. During parameter optimization, the cursor gain is increased so as to increase typing rate, until the cursor moves so quickly that it becomes uncontrollable due to decoding errors[20]. How is it then that handwriting movements could be decoded more than twice as fast, with similar levels of accuracy?

Importantly, we also now clarify that there are other factors to consider when comparing the handwriting BCI to a point-and-click BCI:

> These results suggest that time-varying patterns of movement, such as handwritten letters, are fundamentally easier to decode than point-to-point movements, and can thus enable higher communication rates (although other important differences between continuous point-and-click BCIs and discrete handwriting BCIs, such as the time taken to execute a click, also contribute to their speed difference).

Additionally, we now explicitly clarify that the concept of increasing the temporal dimensionality of the decoded movements can be applied to improve *discrete* BCIs only:

> The concept of intentionally increasing temporal dimensionality could be applied more generally to improve any discrete (but not continuous) BCI that enables ~~discrete~~ selection between a set of options, ~~(~~by associating these options with time-varying gestures as opposed to simple movements~~)~~.

Finally, we now draw a clearer distinction between this work and prior work on discrete intracortical BCIs:

> Prior discrete BCIs have typically used simple directional movements as opposed to spatiotemporally patterned movement, which may have limited their accuracy and/or the size of the movement set[22,23].

[22]Musallam S, Corneil BD, Greger B, Scherberger H, Andersen RA (2004) Cognitive Control Signalsfor Neural Prosthetics. Science 305.

[23]Santhanam G, Ryu SI, Yu BM, Afshar A, Shenoy KV (2006) A high-performance brain–computer interface. Nature 442:195–198.

Specific points:
Is this the same T5 patient from Pandarinath et al. 2017? If so, it would strengthen the manuscript's claims to highlight this direct comparison (where they are also potentially at a disadvantage if studies were performed later with likely lower quality neural recordings).

Yes, it is the same participant. We now highlight this fact explicitly:

> For intracortical BCIs, the highest performing method has been point-and-click typing with a 2D computer cursor, peaking at 40 characters per minute [3] (this record was also set by participant T5 three years earlier; see SVideo 3 for a direct comparison).

If this is the same patient T5, the manuscript should mention that this subject did have the best performance of the 3-subject cohort in that prior study. While the performance advantages of their decoder are clear, given the single subject demonstration this potential subject-to-subject variability should be discussed.

We now highlight more explicitly in the Discussion that this study was from a single participant, and highlight the potential variability across participants:

> Finally, it is important to recognize that ~~our~~ the current system is a proof-of-concept that a high-performance handwriting BCI is possible (in a single participant capable of handwriting prior to his injury); it is not yet a complete, clinically viable system. More work is needed to demonstrate high performance in additional people, expand the character set (e.*g.* capital letters), enable text editing and deletion, and maintain robustness to changes in neural activity without interrupting the user for decoder retraining. More broadly, intracortical microelectrode array technology is still maturing, and requires further demonstrations of longevity, safety, and efficacy before widespread clinical adoption[33,34]. Variability in performance across participants is also a potential concern that may require improvements in intracortical recording technology to increase consistency (in a prior study, T5 achieved the highest performance of 3 participants[7]).

The increase in characters per minute (Figure 2C) should be discussed. In addition to being more accurate over time (which may be attributed to the addition of previous day's data to the RNN training dataset), there is also an observed increase in typing speed (characters per minute). Is this also due to additional training data or other phenomena? A retrospective analysis with decoder performance on a single day's data would be useful information.

We believe that the increase in speed over time is due to T5 becoming more comfortable with writing quickly. Our handwriting BCI is different from a point-and-click BCI in that there is no gain (speed scaling) parameter that effectively determines the typing rate of the BCI. Instead, the pace is set entirely by the user, who chooses how fast to write each letter (i.e. the BCI does not constrain the user to write more slowly in order to maintain accuracy – the writing speed is entirely up to the user). Although we always instructed T5 to write as quickly as possible, he reported increasing his speed over time as he began to trust that the BCI would maintain high accuracies at faster speeds. We note that there is no way to know, objectively, why T5 decided to increase his writing speed other than the reasoning that he reports to us. We therefore added the following sentence to the Results:

> T5 reported increasing his writing speed over time as he gained confidence that the BCI could maintain its accuracy at high speeds.

Note that although characters per minute increased over time, the accuracy did not. Instead, there was a modest increase in *error rate* over time (we observed an average error rate of 4.3% on the first day and 5.4% on the last day). For convenience, Fig. 2C is reproduced below, which shows a small increase in error over time which is interrupted by trial day 1237 (which appears to be an outlier).

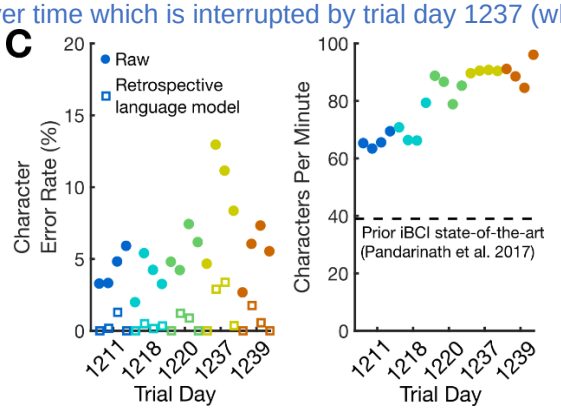
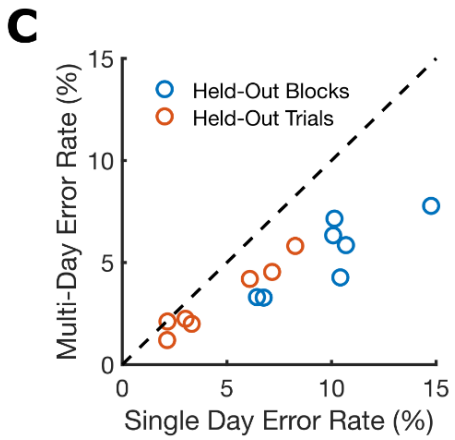


Fig. 2C. Error rates (edit distances) and typing speeds are shown for five days, with four blocks of 7-10 sentences each (each block indicated with a single circle).

We speculate that it is easier to classify at slower speeds because of an increased amount of neural data per character that can be used to decide that character's identity, effectively increasing the overall SNR of the data available to the decoder.

As suggested by the reviewer, it is indeed the case that adding more training data increases the accuracy of the RNN, as shown by Supplementary Figure 2C (reproduced below). This figure panel compares the offline decoding performance of an RNN trained on all days ("Multi-Day Error Rate") to the offline decoding performance of separate RNNs trained on each day alone ("Single Day Error Rate"). Training on all days relative to just a single day reduced the error rate percentage by 4.7 (95% CI = [4.1, 5.3]).



Supp Fig. 2C. Training an RNN with all of the datasets combined improves performance relative to training on each day separately. Each circle shows the performance on one of seven days.

The experimental setup for real-time decoding should be clarified. Did the subject see the raw outputs during the task?

Yes, T5 saw the raw outputs (i.e., each letter, whether correct or incorrect) appear on the screen during the task. In the Results section, we offer the following description:

> T5 copied each sentence from an onscreen prompt, attempting to handwrite it letter by letter, while the decoded characters appeared on the screen in real-time as they were detected by the RNN (SVideo 1, Table S2).

To make this clearer, we amended the legend of Figure 2 to now state the following:

> Finally, the character probabilities were thresholded to produce "Raw Output" for real-time use (when the "new character" probability crossed a threshold at time *t*, the most likely character at time *t*+0.3s was emitted from the decoder and shown on the screen).

The authors nicely isolate the effect of the RNN from the more discriminable neural activity (supplemental table S4). Though I think they somewhat overstate the importance of the RNN compared to HMM in the main manuscript methods, since the RNN's main advantage is it's robustness against noise (by the authors design with noise-training for the RNN). It's actually quite noteworthy that the neural activity differences alone still lead to solid performance in a 31-way classification task with a linear HMM.

Thank you, we too agree that it is very encouraging that a simpler decoding algorithm was able to achieve good performance. Nevertheless, we do think that the numbers in table S4 (0.23% error rate with an RNN and 2.96% with an HMM, when a language model was applied to both) actually do show a large improvement for the RNN, when one considers that a 0.23% error rate means *ten times* fewer errors.

We made the following change to the Methods section where this issue is discussed to make our statement more quantitatively precise:

> We found that a recurrent neural network decoder ~~strongly~~ outperformed a simple hidden Markov model decoder (Table S4, 0.23% error rate vs. 2.93% error rate under the most favorable conditions for the HMM, and 0.70% vs. 80.1% under the least favorable), but note that quite-discriminable neural activity enabled even the HMM decoder to perform reasonably well.

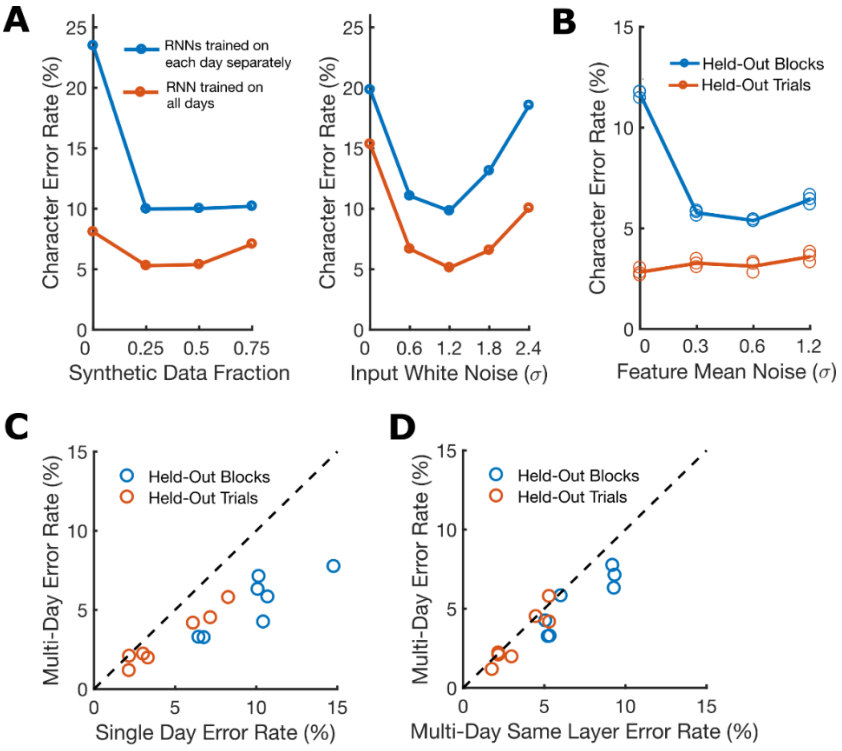
The "character stretch factor" is not well explained in the supplements. What does this factor represent?

Thank you for pointing this out. We added the following sentence of explanation to the supplement:

> The stretch factor determines how the character template is contracted or dilated in time (using linear interpolation) to be longer or shorter than its average duration.

Figure S3C and D -- are these differences statistically significant? More quantification rather than just "substantially improved" would be useful.

Thank you for this suggestion. We added a table to Figure S3 which summarizes the error rate improvement generated by each RNN parameter/technique, with a 95% confidence interval so that statistical significance can be assessed. The new figure is reproduced below (the new table is shown in panel E):



E

Parameter	Panel	Comparison	Error Rate Improvement [95% CI]
Synthetic data fraction	A (left)	Single-day RNN (0 vs. 0.25)	12.9 [11.9, 14.0]
		Multi-day RNN (0 vs. 0.25)	2.7 [2.2, 3.2]
Input white noise	A (right)	Single-day RNN (0 vs. 1.2)	9.8 [8.7, 11.0]
		Multi-day RNN (0 vs. 1.2)	10.1 [9.2, 11.1]
Feature mean noise	B	Held-out blocks (0 vs. 0.3)	5.7 [4.9, 6.5]
		Held-out trials (0 vs. 0.3)	-0.7 [-1.3, -0.1]
Single-day vs. multi-day	C	Held-out blocks	4.7 [4.1, 5.3]
		Held-out trials	1.5 [0.6, 2.4]
Day-specific layers vs. same layer	D	Held-out blocks	1.6 [1.1, 2.7]
		Held-out trials	0.1 [-0.7, 0.7]

Additionally, we updated the Methods text to include quantifications of the performance improvement.

> Including multiple days of data, and fitting separate input layers for each day, ~~substantially~~ significantly improved performance (decreased the error rate percentage by 4.7 and 1.6, respectively; ~~(~~SFig. 3C-D).

I'm left with an impression that many design choices in the machine learning algorithms were hand tailored. This is fine, especially for initial proof of concept. But the discussion might benefit from mentioning that methods for more automated algorithm development/training will be needed for wider utility.

Indeed, many of the parameters were hand-tuned (as we think is typical in machine learning applications). For later days, some hyperparameters were tuned via a random search over possible parameter values. We added the following text to the Methods section to highlight this issue:

> Hyperparameter values were largely hand-tuned; for later sessions, some parameters were tuned via small random searches over possible parameter values (see Supplemental Methods for values). Ultimately, automated parameter tuning may be required (and would certainly be useful) when applying these techniques to new participants in future clinical applications.

Point-by-point responses to referee #3

A. Summary of the key results
The work reports a single subject's performance using an intracortical BCI that can decode imagined handwriting movements from neural activity in motor cortex and map it to text in real-time. Overall the work fits within the growing body of literature intended to demonstrate faster and more accurate BCIs with improved understanding of movement encoding and more sophisticated decoding methods.
Outstanding features of the work are:
• Typing speeds of on-screen prompt at 90 characters per minute at 99% error rate with the use of a general-purpose autocorrect and 73.8 characters at 8.54% error rate for self-generated sentences (2.25% with a language model) are significant advances over the highest reported point and-click typing with a 2D computer cursor, peaking at 40 characters per minute. Results open a new approach for BCIs and demonstrate the feasibility of accurately decoding imagined handwriting movements years after losing ability to move and speak.
• The combination of probabilistic and modeling frameworks forming a hierarchical decoding approach with multiple time scales to combat neural signal variability.
• An interesting theoretical principle is proposed in which point-to-point movements may be harder to decode from one another compared to handwritten letters. Authors attribute this to the idea that temporally complex movements, such as handwriting, may be fundamentally easier to decode than point-to-point movements.

We are gratified that the reviewer expresses that this is a new approach that makes significant gains in BCI performance. We thank the reviewer for their thorough read of the manuscript and their insightful and helpful suggestions.

B. Originality and significance:

The paper draws upon handwriting or touch typing as a faster means to communicate by a specific population of neurologically impaired subjects. The work is an extension to this group's past contributions on BCIs for communications to the 'locked-in' population. Results presented here would be of interest to people in the BCI community who are working on restoring communication to these people who cannot move or speak.

Overall, the work is significant and original but can be better articulated.

We thank the reviewer for noting the originality and significance of the work, and for their detailed suggestions below on how to improve its presentation. We have made every attempt to follow these helpful recommendations, and we believe that the manuscript is much stronger as a result. Again, thank you.

First, authors should cite the prevalence of such conditions to put this contribution in the right context.

Thank you for this suggestion. We added the following to the Discussion:

> Locked-in syndrome (paralysis of nearly all voluntary muscles) severely impairs or prevents communication, and is most frequently caused by brainstem stroke or late-stage ALS (estimated prevalence of locked-in syndrome: 1 in 100,000 [25]).

[25]Pels, Elmar G.M., Erik J. Aarnoutse, Nick F. Ramsey, and Mariska J. Vansteensel. "Estimated Prevalence of the Target Population for Brain-Computer Interface Neurotechnology in the Netherlands." *Neurorehabilitation and Neural Repair* 31, no. 7 (July 2017): 677–85. https://doi.org/10.1177/1545968317714577.

Second, the primary performance metric is typing speed. However, on numerous occasions, the authors attempt to give the impression that this is the primary metric that could be the sole determinant for adopting the technology. While this metric is undoubtedly critical, I think the authors should reframe this argument differently, in that it is the combination of a number of factors —one of which is typing speed—

that would ultimately make the technology a first choice for the intended population. For example, the recalibration of decoders is another such factor, and while it is acknowledged by the authors that their approach is quite extensive, it is unclear how much time and resources the recalibration process takes (see detailed comments below). Another factor is the integrity of the signals over the longevity of the implant, which is a prime issue with all invasive technology (see detailed comments below).

Thank you for the important recommendation to reframe the argument differently, including the suggestion to address the decoder calibration process and electrode array longevity, which has led to new analyses and discussion points (described below) which we believe have significantly improved the manuscript. Since these are important themes that recur in this (R3's) review, we take some space here to outline our overall philosophy and approach, as well as highlight each major addition to the paper.

Ultimately, we see this study as being primarily focused on demonstrating the feasibility of decoding handwriting movements well enough to substantially increase BCI communication rates. This opens the door to a promising new approach for iBCIs, which we believe is an important and exciting advance. To our knowledge, this is the first demonstration that rapid sequences of dexterous movements can be decoded in a person who has been paralyzed for several years. However, by no means does our BCI yet constitute a 'complete product' that would be appropriate for immediate clinical adoption.

First, as the reviewer has noted here and below, array longevity is a critical issue for any intracortical BCI. Before a product is taken to market, a systematic study must be conducted which demonstrates longevity across many subjects. While no such study has yet been published, preliminary results from several studies indicate that arrays retain their functionality for several years in people, with multiple examples of retained functionality for 1000+ days (Bullard et al., 2020; Simeral et al., 2011). In this current study, high performance was obtained 1200+ days post-implant, and these arrays continue to record high-quality spiking activity (see below). We are currently preparing a separate manuscript summarizing array safety and longevity data from all 14 participants in the BrainGate pilot clinical trials (collected over a span of 15 years), which will be the first systematic study of its kind in people. Given the complex and multiple factors contributing to array longevity, we believe this important fundamental question is outside the scope of the current work (beyond simply noting that the results were obtained 3+ years post-implant and that the arrays continue to record high-quality signals). We now clearly highlight this issue in the Discussion and have added a new supplementary figure demonstrating that the arrays continue to record high-quality spiking activity.

Second, as the reviewer notes, minimizing decoder recalibration time is also an important problem for intracortical BCIs (as well as many non-invasive BCIs). This issue must also be addressed before a viable product can be taken to market. However, we see this as another deep topic in and of itself, which has been the sole focus of several recent studies (Jarosiewicz et al., 2015; Dyer et al., 2017; Degenhart et al., 2020). For example, one new method uses an unsupervised approach to track a stable subspace of neural activity over time (Degenhart et al., 2020); the evaluation and design of this method was the subject of an entire paper. Additionally, to our knowledge, daily decoder recalibration is still standard practice in the intracortical BCI field and many important papers have used this method [e.g. (Hochberg et al., 2006, 2012; Collinger et al., 2013; Bouton et al., 2016; Ajiboye et al., 2017; Pandarinath et al. 2017; Nuyujukian et al. 2018]). We think it is therefore reasonable to leave this aspect of handwriting decoding to be more fully investigated in future work. Only now that we have shown that handwriting decoding can achieve higher performance than any other communication BCI, is it properly motivated to begin searching for algorithms that can minimize the calibration data needed to retrain a handwriting decoder.

Nevertheless, we wholeheartedly agree that it is important to both (1) clearly highlight the issue in the Results and Discussion, and (2) preliminarily address whatever key issues we can while remaining within the scope of this work (which we have done via new data analyses, presented below and in the paper). We have taken the following four actions to address the longevity and recalibration issues.

(1) We added new Discussion paragraphs which more clearly address the limitations of the current work, including limitations with intracortical array technology in general (e.g., that more studies are needed to show array longevity). We also give some broader context as to why we believe intracortical technology is

a promising route forward for restoring rapid communication, despite not necessarily being ready for widespread clinical adoption at the current time. For convenience, we reproduce these new paragraphs here:

> Finally, it is important to recognize that ~~our~~ the current system is a proof-of-concept that a high-performance handwriting BCI is possible (in a single participant capable of handwriting prior to his injury); it is not yet a complete, clinically viable system. More work is needed to demonstrate high performance in additional people, expand the character set (e.g. capital letters), enable text editing and deletion, and maintain robustness to changes in neural activity without interrupting the user for decoder retraining. More broadly, intracortical microelectrode array technology is still maturing, and requires further demonstrations of longevity, safety, and efficacy before widespread clinical adoption[33,34]. Variability in performance across participants is also a potential concern that may require improvements in intracortical recording technology to increase consistency (in a prior study, T5 achieved the highest performance of 3 participants[7]).
>
> Nevertheless, we believe the future of intracortical BCIs is bright. Current microelectrode array technology has been shown to retain functionality for 1000+ days post implant[35,36] (including in this work - see SFig 6 for examples of high-quality spiking activity), and has enabled the highest BCI performance to date compared to other recording technologies (EEG, ECoG) for restoring communication[7], arm control[2,5], and general-purpose computer use[37]. New developments are underway for implant designs that increase the electrode count by an order of magnitude, which will further improve performance and longevity[33,34,38,39]. Finally, we envision that a combination of algorithmic innovations and improvements to device stability will continue to increase the robustness of intracortical BCIs, which have so far typically required daily decoder retraining to account for changes in neural recordings that accrue over time (although see e.g. [40,41]). In this study, offline analyses showed that large amounts of daily calibration data are not needed to achieve good performance, and that an unsupervised approach is promising for enabling behind-the-scenes decoder retraining without interrupting the user. Other recent work has also advanced new algorithms for unsupervised decoder retraining[42,43], making important steps towards robust, easy-to-use intracortical BCI systems.

(2) We added a new figure (Fig. 3) to the main text focused solely on decoder recalibration. This figure reports results from new offline analyses that quantify how much calibration data is needed to achieve high performance. As noted by the reviewer, our original study design used a large amount of calibration data to retrain the decoder each day (50 sentences). However, this was not because it was *necessary* to use that many sentences to achieve good performance. Our new analysis demonstrates that high performance could have be obtained with much less data (10 sentences). We also assess whether the amount of time that passes between sessions affects how much calibration data is needed. We show that, when 7 days or less pass between sessions, it is possible to achieve good performance even with *no* decoder calibration. Moreover, we demonstrate that an unsupervised decoder recalibration method can achieve high performance without requiring any explicit data labels. This is promising from a clinical viability standpoint, as it suggests that a decoder recalibration routine may be able to run in the background without interrupting the user. We believe that these new results improve the paper significantly by (a) highlighting this important issue and (b) showing initial promise that it is possible to achieve high performance with modest amounts of calibration data.

We reproduce the new figure (Fig. 3) and accompanying Results text below for convenience:

> Following standard practice for BCIs (e.g. [1,2,19,4,5]), we retrained our handwriting decoder each day before evaluating it, with the help of "calibration" data collected at the beginning of each day. Retraining helps account for changes in neural recordings that accrue over time. Ideally, to reduce the burden on the user, little or no calibration data would be required. In a retrospective analysis of the copy typing data reported above in Fig. 2, we assessed whether high performance could still have been achieved using less than the original 50 calibration sentences per day (Fig. 3A). We found that 10 sentences (8.7 minutes) were enough

to achieve a raw error rate of 8.5% (1.7% with a language model), although 30 sentences (26.1 minutes) were needed to match the raw online error rate of 5.9%.

However, our copy typing data were collected over a 28-day time span, possibly allowing larger changes in neural activity to accumulate. We therefore tested whether more closely-spaced sessions reduce the need for calibration data (Fig. 3B), using an offline analysis of copy typing data across 8 sessions. We found that when only 2-7 days passed between sessions, performance was reasonable with *no* decoder retraining (11.1% raw error rate, 1.5% with a language model). Finally, we tested whether decoders could be retrained in an unsupervised manner by using a language model to error-correct and retrain the decoder, thus bypassing the need to interrupt the user for calibration (i.e. by recalibrating automatically during normal use). Encouragingly, unsupervised retraining achieved a 7.3% raw error rate (0.84% with a language model) when sessions were separated by 7 days or less (see Methods & Supplemental Methods for details). Ultimately, whether decoders can be successfully retrained with minimal recalibration data depends on how quickly the neural activity changes over time. We assessed the stability of the neural patterns associated with each character and found high short-term stability (mean correlation = 0.85 when 7 days apart or less), and neural changes that seemed to accumulate at a steady and predictable rate (SFig. 4 provides a quantitative visualization). The above results are promising for clinical viability, as they suggest that unsupervised decoder retraining, combined with more limited supervised retraining after longer periods of inactivity, may be sufficient to achieve high performance.

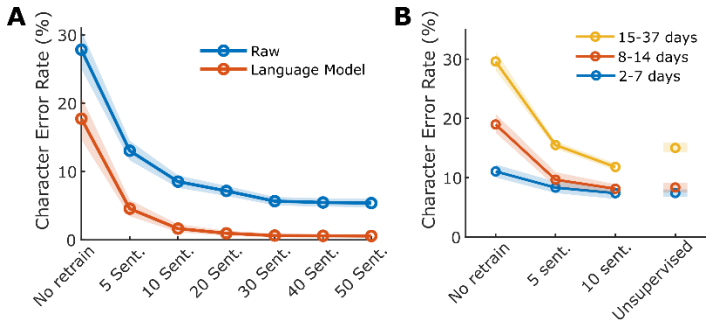
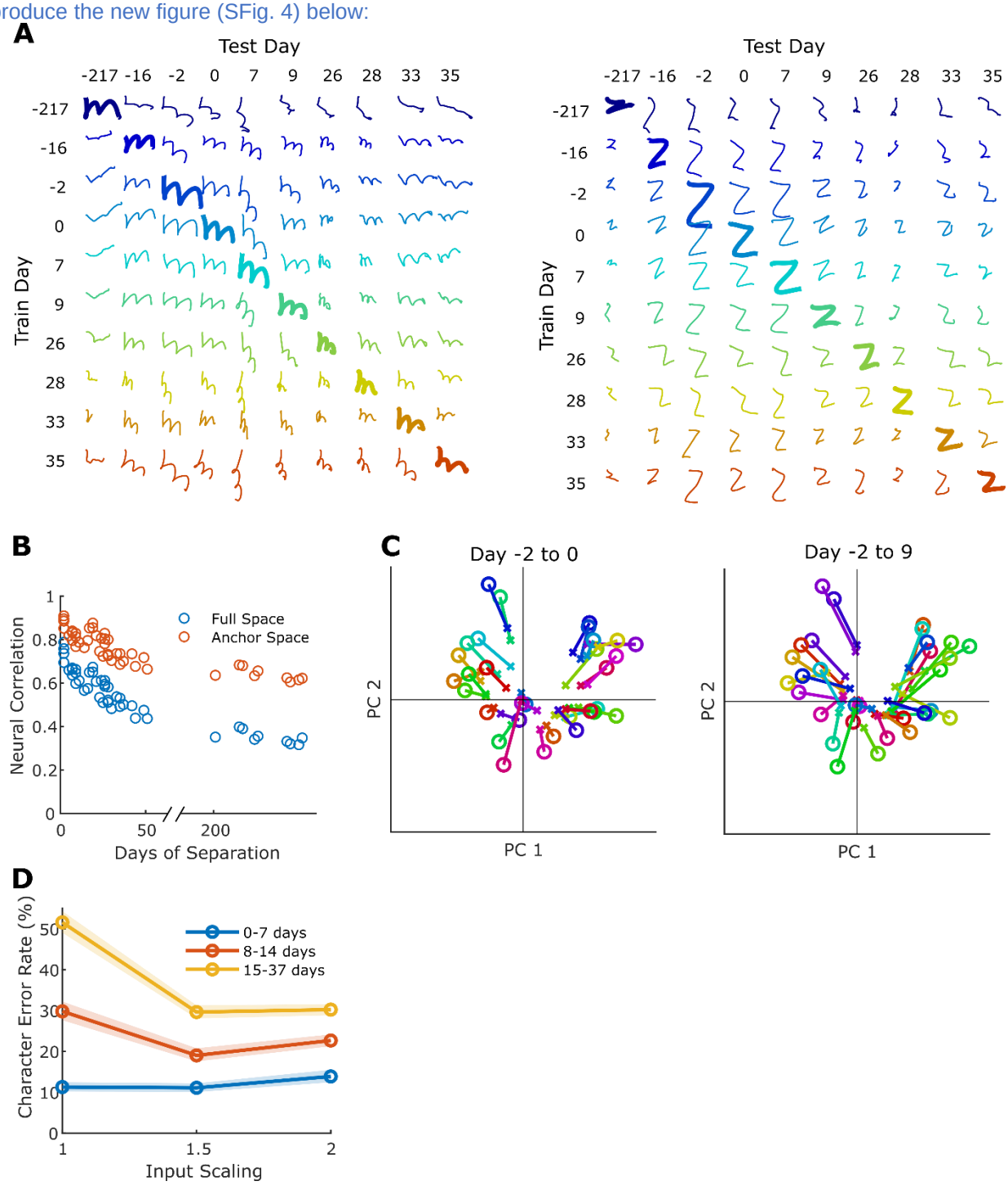


Figure 3. Performance remains high when decoder retraining is limited or omitted. (A) To account for neural activity changes that accrue over time, we retrained our handwriting decoder each day before evaluating it. Here, we simulate offline what the decoding performance shown in Fig. 2 would have been if less than 50 calibration sentences were used. Lines show the mean error rate across all data and shaded regions indicate 95% CIs (computed via bootstrap resampling of single trials, N=10,000). **(B)** Copy typing data from eight sessions were used to assess whether less calibration data are required if sessions occur closer in time. All session pairs (X, Y) were considered. Decoders were first initialized using training data from session X and earlier, and then evaluated on session Y under different retraining methods (no retraining, retraining with limited calibration data, or unsupervised retraining). The average raw character error rate is plotted for each category of time elapsed between sessions X and Y, and for each retraining method. Shaded regions indicate 95% CIs.

(3) We added a new supplemental figure (now SFig. 4) that assess the stability of the neural patterns associated with each character over time, since this is a critical issue that ultimately determines how much data is needed for daily decoder recalibration. We found high short-term stability (mean correlation = 0.85 when 7 days apart or less), and neural changes that seemed to accumulate at a steady and predictable rate. Again, this is promising for the possibility of recalibrating decoders with limited amounts of data (or even in an unsupervised manner without interrupting the user). We also found that as neural activity slowly rotates into new neural subspaces over time, it tends to shrink towards the origin in the original neural subspace, but otherwise retains a very similar structure there. This suggests the following idea: if we scale up the inputs to the decoder when transferring it to a new day, this might prevent the decoder from perceiving smaller-than-expected modulation in the original subspace. We found that input re-scaling does indeed improve performance, and we include this result as part of the supplemental figure. We think this is a useful principle that could benefit other types of BCIs as well.

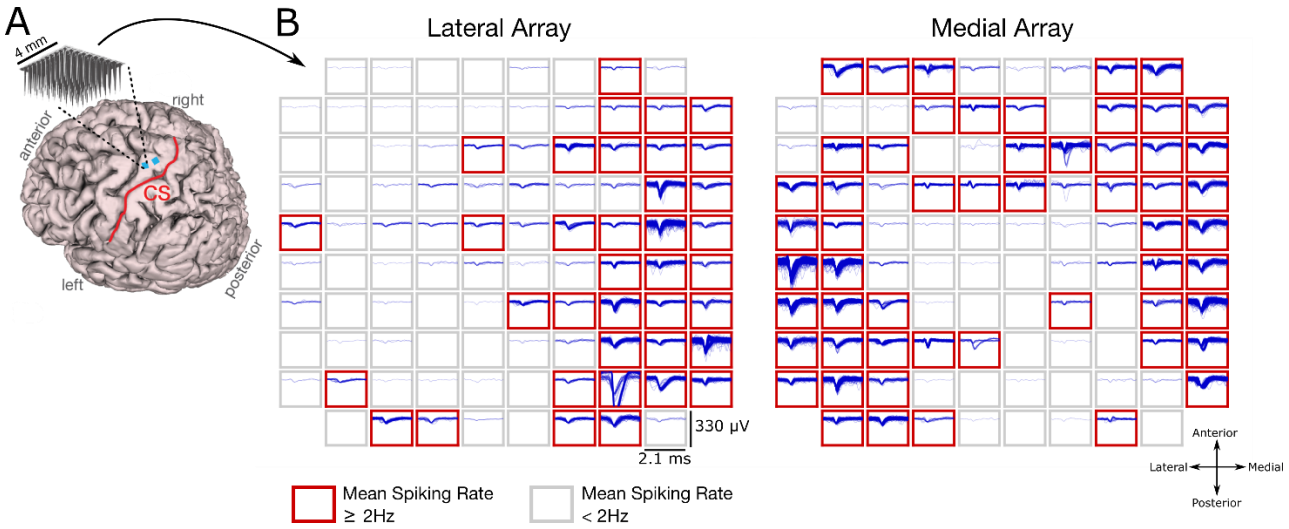
We reproduce the new figure (SFig. 4) below:



Supplemental Figure 4. Changes in neural recordings across days. (A) To visualize how much the neural recordings changed across time, decoded pen tip trajectories were plotted for two example letters ("m" and "z") for all ten days of data (columns), using decoders trained on all other days (rows). Each session is labeled according to the number of days passed relative to Dec. 9, 2019 (day #4). Results show that although neural activity patterns clearly change over time, their essential structure is largely conserved (since decoders trained on past days transfer readily to future days). (B) The correlation (Pearson's *r*)

between each session's neural activity patterns was computed for each pair of sessions and plotted as a function of the number of days separating each pair. The *r* values were computed by correlating the spatiotemporal neural patterns of average firing rates associated with each character (see Supplemental Methods for more detail). Blue circles show the correlation computed in the full neural space (all 192 electrodes) while red circles show the correlation in the "anchor" space (top 10 principal components of the earlier session). High values indicate a high similarity in how characters are neurally encoded across days. The fact that correlations are higher in the anchor space suggests that the structure of the neural patterns stays largely the same as it slowly rotates into a new space, causing shrinkage in the original space but little change in structure. (C) A visualization of how each character's neural representation changes over time, as viewed through the top two PCs of the original "anchor" space. Each "o" represents the neural activity pattern for a single character, and each "x" shows that same character on a later day (lines connect matching characters). The left panel shows a pair of sessions with only two days between them ("Day -2 to 0"), while the right panel shows a pair of sessions with 11 days between them ("Day -2 to 9"). The relative positioning of the neural patterns remains similar across days, but most conditions shrink noticeably towards the origin. This is consistent with the neural representations slowly rotating out of the original space into a new space, and suggests that scaling-up the input features may help a decoder to transfer more accurately to a future session (by counteracting this shrinkage effect). (D) Similar to Fig. 3B, copy typing data from eight sessions was used to assess offline whether scaling-up the decoder inputs improves performance when evaluating the decoder on a future session (when *no* decoder retraining is employed). All session pairs (X, Y) were considered. Decoders were first initialized using all data from session X and earlier, then evaluated on session Y under different input scaling factors (e.g., an input scale of 1.5 means that input features were scaled up by 50%). The average raw character error rate is plotted for each category of time elapsed (between sessions X and Y) and each retraining method. Shaded regions indicate 95% CIs. Results show that when long periods of time pass between sessions, input-scaling improves performance. We therefore used an input scaling factor of 1.5 when assessing decoder performance in the "no retraining" conditions of Fig. 3.

(4) We added a new supplemental figure (now SFig. 6) to demonstrate that high quality spiking activity can still be recorded on many of the microelectrodes 3+ years post-implant. This demonstrates that intracortical microelectrode arrays have the potential to last for several years in people (although as stated above, additional evidence from more subjects will ultimately be required to systematically demonstrate longevity). We also quantified how many of the total 192 electrodes could still record high-quality spiking activity and now report this number in the Methods (81.9 ± 5.6), which we believe gives the reader useful additional context. We used a simple, conservative metric to estimate if an electrode still recorded spike-like activity that could have arisen from single neurons. Specifically, if the voltage crossed a -4.5 RMS threshold more than 2 times per second on average, the electrode was considered to record spiking activity. Note that a -4.5 RMS threshold excludes almost all background noise (and many electrodes therefore record almost no spiking events at this threshold). Although we could have also spike-sorted these waveforms, spike-sorting is a subjective and somewhat arbitrary process since it is not always clear whether a cluster of waveforms truly belongs to one (and only one) neuron. Thus, this metric is a lower bound on the number of spike clusters (since the activity on each spiking electrode could be sorted into *at least* one cluster). This new figure is reproduced below:



Supplemental Figure 6. Example spiking activity recorded from each microelectrode array. (A) Participant T5's MRI-derived brain anatomy. Microelectrode array locations (blue squares) were determined by co-registration of postoperative CT images with preoperative MRI images. (B) Example spike waveforms detected during a ten second time window are plotted for each electrode (data were recorded on post-implant day 1218). Each rectangular panel corresponds to a single electrode and each blue trace is a single spike waveform (2.1 millisecond duration). Spiking events were detected using a -4.5 RMS threshold, thereby excluding almost all background activity. Electrodes with a mean threshold crossing rate ≥ 2 Hz were considered to have 'spiking activity' and are outlined in red (note that this is a conservative estimate that is meant to include only spiking activity that could be from single neurons, as opposed to multiunit 'hash'). Results show that many electrodes still record large spiking waveforms that are well above the noise floor (the y-axis of each panel spans 330 µV, while the background activity has an average RMS value of only 6.4 µV). On this day, 92 electrodes out of 192 had a threshold crossing rate ≥ 2 Hz.

Taken together, we believe that these new results and discussion points improve the manuscript considerably by providing more perspective about the limitations and potential benefits of intracortical BCIs, while also offering new evidence that minimizing (or in some cases, even eliminating) supervised decoder recalibration appears feasible.

References for this section:

Ajiboye, A.B., Willett, F.R., Young, D.R., Memberg, W.D., Murphy, B.A., Miller, J.P., Walter, B.L., Sweet, J.A., Hoyen, H.A., Keith, M.W., Peckham, P.H., Simeral, J.D., Donoghue, J.P., Hochberg, L.R., Kirsch, R.F., 2017. Restoration of reaching and grasping movements through brain-controlled muscle stimulation in a person with tetraplegia: a proof-of-concept demonstration. The Lancet 389, 1821–1830. https://doi.org/10.1016/S0140-6736(17)30601-3

Bouton, C.E., Shaikhouni, A., Annetta, N.V., Bockbrader, M.A., Friedenberg, D.A., Nielson, D.M., Sharma, G., Sederberg, P.B., Glenn, B.C., Mysiw, W.J., Morgan, A.G., Deogaonkar, M., Rezai, A.R., 2016. Restoring cortical control of functional movement in a human with quadriplegia. Nature 533, 247–250. https://doi.org/10.1038/nature17435

Bullard, A.J., Hutchison, B.C., Lee, J., Chestek, C.A., Patil, P.G., 2020. Estimating Risk for Future Intracranial, Fully Implanted, Modular Neuroprosthetic Systems: A Systematic Review of Hardware Complications in Clinical Deep Brain Stimulation and Experimental Human Intracortical Arrays. Neuromodulation Technol. Neural Interface 23, 411–426. https://doi.org/10.1111/ner.13069

Collinger, J.L., Wodlinger, B., Downey, J.E., Wang, W., Tyler-Kabara, E.C., Weber, D.J., McMorland, A.J., Velliste, M., Boninger, M.L., Schwartz, A.B., 2013. High-performance neuroprosthetic control by an individual with tetraplegia. The Lancet 381, 557–564. https://doi.org/10.1016/S0140-6736(12)61816-9

Degenhart, A.D., Bishop, W.E., Oby, E.R., Tyler-Kabara, E.C., Chase, S.M., Batista, A.P., Yu, B.M., 2020. Stabilization of a brain–computer interface via the alignment of low-dimensional spaces of neural activity. Nat. Biomed. Eng. 1–14. https://doi.org/10.1038/s41551-020-0542-9

Dyer, E.L., Gheshlaghi Azar, M., Perich, M.G., Fernandes, H.L., Naufel, S., Miller, L.E., Körding, K.P., 2017. A cryptography-based approach for movement decoding. Nat. Biomed. Eng. 1, 967–976. https://doi.org/10.1038/s41551-017-0169-7

Hochberg, L.R., Bacher, D., Jarosiewicz, B., Masse, N.Y., Simeral, J.D., Vogel, J., Haddadin, S., Liu, J., Cash, S.S., Smagt, P. van der, Donoghue, J.P., 2012. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. Nature 485, 372–375. https://doi.org/10.1038/nature11076

Hochberg, L.R., Serruya, M.D., Friehs, G.M., Mukand, J.A., Saleh, M., Caplan, A.H., Branner, A., Chen, D., Penn, R.D., Donoghue, J.P., 2006. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. Nature 442, 164–171. https://doi.org/10.1038/nature04970

Jarosiewicz, B., Sarma, A.A., Bacher, D., Masse, N.Y., Simeral, J.D., Sorice, B., Oakley, E.M., Blabe, C., Pandarinath, C., Gilja, V., Cash, S.S., Eskandar, E.N., Friehs, G., Henderson, J.M., Shenoy, K.V., Donoghue, J.P., Hochberg, L.R., 2015. Virtual typing by people with tetraplegia using a self-calibrating intracortical brain-computer interface. Sci. Transl. Med. 7, 313ra179-313ra179. https://doi.org/10.1126/scitranslmed.aac7328

Simeral, J.D., Kim, S.-P., Black, M.J., Donoghue, J.P., Hochberg, L.R., 2011. Neural control of cursor trajectory and click by a human with tetraplegia 1000 days after implant of an intracortical microelectrode array. J. Neural Eng. 8, 025027. https://doi.org/10.1088/1741-2560/8/2/025027

Third, given the paper's emphasis on how the character and word decoding rates surpass existing state of the art, the data may actually have much more information about the nature of neural representation of attempted handwriting that could benefit a broader audience (particularly the neurobiology and neurophysiology communities), but this is not emphasized in the current version of the paper. As such, it is unclear if the work will be of immediate interest to many people from several disciplines.

Thank you for this suggestion. We appreciate the desire to understand how handwriting is neurally represented and what this might mean for the cortical motor system in general. We are currently working on a separate manuscript to accomplish this goal. Since there are already numerous BCI-related results and methods that we must cover in this manuscript, we believe that it is best to retain the current focus on the BCI aspects.

A BCI-centered focus keeps within the tradition of previous "first-of" BCI papers (examples referenced below), which have all achieved wide interest and impact by focusing largely on their BCI achievement. Additionally, we believe that the computational richness of the problem of neurally decoding sequences of handwriting movements, combined with a public release of this unique dataset, should attract broad interest across the machine learning community as well.

Ajiboye, A.B., Willett, F.R., Young, D.R., Memberg, W.D., Murphy, B.A., Miller, J.P., Walter, B.L., Sweet, J.A., Hoyen, H.A., Keith, M.W., Peckham, P.H., Simeral, J.D., Donoghue, J.P., Hochberg, L.R., Kirsch, R.F., 2017. Restoration of reaching and grasping movements through brain-controlled muscle stimulation in a person with tetraplegia: a proof-of-concept demonstration. The Lancet 389, 1821–1830. https://doi.org/10.1016/S0140-6736(17)30601-3

Anumanchipalli, G.K., Chartier, J., Chang, E.F., 2019. Speech synthesis from neural decoding of spoken sentences. Nature 568, 493–498. https://doi.org/10.1038/s41586-019-1119-1

Bouton, C.E., Shaikhouni, A., Annetta, N.V., Bockbrader, M.A., Friedenberg, D.A., Nielson, D.M., Sharma, G., Sederberg, P.B., Glenn, B.C., Mysiw, W.J., Morgan, A.G., Deogaonkar, M., Rezai, A.R., 2016. Restoring cortical control of functional movement in a human with quadriplegia. Nature 533, 247–250. https://doi.org/10.1038/nature17435

Collinger, J.L., Wodlinger, B., Downey, J.E., Wang, W., Tyler-Kabara, E.C., Weber, D.J., McMorland, A.J., Velliste, M., Boninger, M.L., Schwartz, A.B., 2013. High-performance neuroprosthetic control by an individual with tetraplegia. The Lancet 381, 557–564. https://doi.org/10.1016/S0140-6736(12)61816-9

Hochberg, L.R., Bacher, D., Jarosiewicz, B., Masse, N.Y., Simeral, J.D., Vogel, J., Haddadin, S., Liu, J., Cash, S.S., Smagt, P. van der, Donoghue, J.P., 2012. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. Nature 485, 372–375. https://doi.org/10.1038/nature11076

Hochberg, L.R., Serruya, M.D., Friehs, G.M., Mukand, J.A., Saleh, M., Caplan, A.H., Branner, A., Chen, D., Penn, R.D., Donoghue, J.P., 2006. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. Nature 442, 164–171. https://doi.org/10.1038/nature04970

Fourth, direct comparison to behaviors requiring dexterous movements such as typing at speeds of 120 characters per minute for intact subjects is somewhat irrelevant since the ability to modulate brain signals to become a reliable source of control of these assistive devices vary considerably among human subjects who cannot move or speak. For example, it is unclear that the achieved speed/error rates will generalize to other subjects with similar impairment. In other occasions, they draw comparison to speech-decoding BCIs for restoring verbal communication, but this technology is at a very early stage to be compared to the current approach.

Thank you for highlighting this important point. Indeed, subject-to-subject variability is an important issue in BCI research, especially for a single-subject study. In our Discussion section, we now more explicitly mention that this is a limitation of the current work (reproduced below for convenience):

> Finally, it is important to recognize that ~~our~~ the current system is a proof-of-concept that a high-performance handwriting BCI is possible (in a single participant capable of handwriting prior to his injury); it is not yet a complete, clinically viable system. More work is needed to demonstrate high performance in additional people, expand the character set (e.g. capital letters), enable text editing and deletion, and maintain robustness to changes in neural activity without interrupting the user for decoder retraining. More broadly, intracortical microelectrode array technology is still maturing, and requires further demonstrations of longevity, safety, and efficacy before widespread clinical adoption[33,34]. Variability in performance across participants is also a potential concern that may require improvements in intracortical recording technology to increase consistency (in a prior study, T5 achieved the highest performance of 3 participants[7]).

Again, we agree that subject-to-subject variability is an important issue to highlight (as per above), and we also believe that it is helpful to readers to place these BCI typing rates into a broader context by comparing them to able-bodied typing rates. Comparing to able-bodied typing rates can help the reader better appreciate how fast the current BCI typing rates are, and how much of a gap between BCI performance and able-bodied typing remains. We think that BCI research should seek to achieve communication rates that are as close to able-bodied communication rates as possible, as presumably this gives the most benefit to the user (although it may not always be possible to do so).

Regarding speech-decoding BCIs, we thought that it would offer the reader valuable context to briefly review other types of communication BCIs and how they compare with the handwriting BCI. For example, readers might wonder whether there is value in a handwriting BCI if a speech BCI can restore communication at much faster speeds. We think it is therefore appropriate to let the reader know that although speech is faster than handwriting, no speech BCI has yet demonstrated accuracies high enough to restore general-purpose communication. We briefly mention speech BCIs only once, in the following sentence in the Discussion (which we have re-worded in a more positive way):

> Recently, speech-decoding BCIs have shown exciting promise for restoring rapid communication (e.g. [32,17,18]), but their accuracies and vocabulary sizes require further improvement to support~~ies are currently too limited for~~ general-purpose use.

Taken together, the authors should present their findings within the broader context in which the population of potential beneficiaries need to opt for a brain surgery with unknown longevity of the implanted device and a relatively long calibration process to gain additional typing speeds (extra 33 characters/min as I consider the self-paced performance reported here to be the real use case of such communication technology).

Thank you for this important suggestion to address the broader context. We have done our best to place this work into the broader context of intracortical BCI technology. The Discussion now mentions both the current limitations of intracortical BCIs and the reasons to be optimistic about how the technology may continue to evolve. In addition, we have addressed the calibration issue directly, as described above, and

now highlight it more extensively in the Discussion and Results. We believe that our new analyses demonstrate that a long calibration process is likely not necessary.

Regarding brain surgery and array longevity, we believe that a product should be brought to market only after safety and efficacy clinical trial studies systematically demonstrate array safety and longevity. We do not advocate that the general patient population opt for a medical product of unknown longevity/efficacy, and of course FDA approval would be required before this is even possible. As is standard practice with clinical trials, only participants who clearly understand that there is no benefit assured if they elect to be a part of an early clinical trial (e.g., BrainGate) should consider providing informed consent and participating in the clinical trial. To help better assess array longevity, we are currently preparing a manuscript that summarizes longevity and efficacy data systematically across all 14 participants in the BrainGate pilot clinical trials. Similarly, we envision that neurotechnology companies (e.g. Synchron, Neuralink, Paradromics) will (and must) conduct systematic trials to evaluate the safety and longevity of any new electrode device before a product is released. We have thus chosen to structure the Discussion with an eye towards the fact that more safety and longevity data will be collected in the future, as opposed to weighing the current lack of such data as a disadvantage for the handwriting BCI. We view our work as providing additional motivation to collect such data and for research groups to pursue this line of research (and for companies to pursue such a product).

Finally, we would like to briefly clarify that, to our knowledge, the current BCI typing record for free-typing (as opposed to copy-typing) is 24 characters per minute. This record was set by an intracortical point-and-click BCI (Pandarinath et al. *eLife* 2017). Thus, our current free-typing rate of 73 characters per minute is an extra 49 characters per minute (three-fold increase).

C. Data & methodology:

General comments:

The presentation is clear, logical and readable to general audience. The reporting of data and methodology is sufficiently detailed to enable reproducing the results. They state that they will share the data and code to enable reproducibility.

Thank you.

Major Comments:

The authors state that they ‘linearly decoded pen tip velocity from neural activity’. Arguably, this variable varies considerably among different people depending on their handwriting style, accuracy, appearance, readability, etc. Did the authors have a sample handwriting from the subject before injury so they can be compared to the ones they decoded? If so, could they analyze such data to infer the pen tip speed profiles the subject likely used to better understand if the observed neural activity correlated with the character shapes? it would be more helpful if the work attempts to provide some understanding of the extent to which the dynamics of the ensemble neural activity do actually reflect this critical behavioral parameter.

Thank you for this interesting idea. Unfortunately, we did not have any handwriting samples readily available to us, as T5’s injury occurred 9 years prior to this study (otherwise we would have proceeded as you describe). Instead, we describe below what we did do, but it in greater detail so that it is clearer (and we have also added detail to the manuscript to make it clearer as well).

To understand T5’s writing style, we interviewed T5 about how exactly he wrote each letter. Then, we used a computer mouse to trace the trajectory of each letter in the same way that T5 reported doing so (while recording the X & Y velocity of the mouse pointer). These trajectory templates, *which are time series of velocity vectors* (not spatial drawings), were then used to train a linear decoder to decode the pen tip velocity. Although these trajectories cannot be expected to match T5’s trajectories in a precise way, they should nevertheless capture the general features of each letter trajectory. Figure 1D,

reproduced below for convenience, shows the output of linear decoders trained to decode pen tip velocity using these templates:

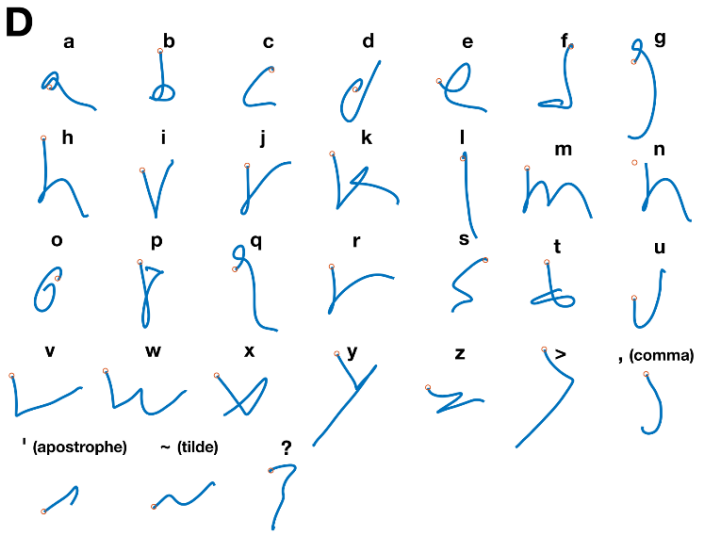


Fig. 1D. Decoded pen trajectories are shown for all 31 tested characters: 26 lower-case letters, commas, apostrophes, question marks, tildes (~) and greater-than signs (>). Intended 2D pen tip velocity was linearly decoded from the neural activity using cross-validation (each character was held out). The decoded velocity was then averaged across trials and integrated to compute the pen trajectory (orange circles denote the start of the trajectory).

Importantly, these letter reconstructions were *held-out* reconstructions. In other words, each letter shape was reconstructed using a decoder that was trained only on *other* letters. The output of each velocity decoder was then cumulatively integrated to compute a pen tip position trajectory, which was drawn as the character reconstruction in Fig.1D. The fact that recognizable letter shapes were decoded demonstrates that there was a consistent neural encoding of pen tip velocity. Otherwise, the decoders might have been able to overfit to the training data but would not have been able to reconstruct pen tip velocity correctly for held-out characters, resulting in unrecognizable shapes. It is worth noting that the reconstructed pen trajectories are well correlated with the letter templates (r = 0.74 across all held-out reconstructions).

In the original manuscript, this important detail about decoder training was mentioned only in the figure legend and Methods. We now clarify in the Results text that reconstructions were only made with decoders *not* trained on that character:

> Readily recognizable letter shapes confirm that pen tip velocity is robustly encoded (each character's reconstruction was made using a decoder trained only on *other* characters).

We also amended the Methods section to add more detail:

> To train the decoder, we used hand-made templates that describe each character's pen trajectory. The character templates were made by drawing each character with a computer mouse in the same way as T5 described writing the character. As each character was drawn, the X and Y velocity trajectories of the mouse pointer were recorded.~~;~~ T~~t~~hese templates (which are time series of velocity vectors) then defined the target velocity vector for the decoder on each time step of each trial. We used ordinary least squares regression to train the decoder to minimize the error between the template velocities and the decoded velocities (see Supplemental Methods for more details). The reconstructed pen tip velocities that were decoded in Fig. 1D were well correlated with the mouse templates (r = 0.74 across all characters).

Next, to understand how large the neural encoding of pen tip velocity was compared to other elements of the neural activity, we used a linear encoding model to fit neural activity as a function of the reconstructed pen tip velocity. This quantifies how much of the neural activity is captured by the neural dimensions that encode pen tip velocity. We found that 30% of the variance was accounted for by pen tip velocity. This is a sizeable portion but still leaves much of the neural activity unaccounted for, which is consistent with recent studies that have highlighted non-kinematic aspects of motor cortical activity [1-2]. We now report this in the Results:

The neural dimensions that represented pen tip velocity accounted for 30% of the total neural variance.

[1] Kaufman, Matthew T., Jeffrey S. Seely, David Sussillo, Stephen I. Ryu, Krishna V. Shenoy, and Mark M. Churchland. "The Largest Response Component in the Motor Cortex Reflects Movement Timing but Not Movement Type." *ENeuro* 3, no. 4 (July 1, 2016): ENEURO.0085-16.2016. https://doi.org/10.1523/ENEURO.0085-16.2016.

[2] Churchland, Mark M., John P. Cunningham, Matthew T. Kaufman, Justin D. Foster, Paul Nuyujukian, Stephen I. Ryu, and Krishna V. Shenoy. "Neural Population Dynamics during Reaching." *Nature* 487, no. 7405 (July 5, 2012): 51–56. https://doi.org/10.1038/nature11129.

Also, the authors should demonstrate the extent to which character encoding might have changed as a function of trials/sentences/sessions, particularly during times when the subject was observing the prompted text, the decoded text, and when the subject was asked to write from memory. This characterization is also needed to provide credence for the claim made in the conclusion that this is a BCI without visual feedback.

Thank you for this suggestion. Due to this suggestion and others below, we have now removed the claim that our BCI can function without visual feedback. While we did collect some data with his eyes closed, it is not a major point and we believe that it is better to remove this to help the manuscript stay focused.

We think that analyzing differences in neural tuning across contexts is a valuable direction for future work, but one that lies outside the scope of the current study, as it does not directly bear on the central claims of this manuscript. Analyzing how characters are neurally encoded during sentence writing is also a difficult task, as it requires accurate segmentation of unlabeled data. Although we have solved this problem well enough for BCI decoding, it is unclear whether small errors in data segmentation could cause artifactual differences in neural encoding to appear.

Given the data we have shown, we would propose that the neural encoding must be at least broadly similar across contexts, since decoders trained on open-loop data (where T5 is copying sentence prompts but no BCI is active) can transfer accurately to the closed-loop context (where T5 is using the BCI and observing the decoded text appear on the screen). Nevertheless, we do agree that differences in neural coding across contexts may have been the cause of some decoding errors, and that it would be worthwhile and interesting to pursue this possibility in future work focused on addressing this question.

It is unclear if the authors have characterized the performance long enough (beyond the stated 10 sessions) to report how nonstationarity in the neural signals can potentially deteriorate the performance reported. In fact, with the exception of the first couple of sessions that were spaced almost a month apart, the remaining 9 sessions took place almost 6 months afterwards and were closely spaced, happening within the span of 7-8 weeks. From the extensive calibration protocol described, there seems to be substantial variability in these signals.

Thank you again for this helpful suggestion. As we laid out above, we believe the new analyses on nonstationarity and decoder calibration provide useful insight into these questions. We think that the 10 sessions we reported, which comprise the entirety of our data, are sufficient for preliminary estimates of nonstationarity and the amount of calibration data required for decoder training.

More specifically, closely-spaced sessions are the most relevant to this question, as we imagine that this kind of BCI will be used at least once every few days or possibly weeks. Given the size of neural

nonstationarities that accrue on intracortical electrode arrays over long time spans (e.g. several months) [1], we don't expect decoders to be able to retain high performance after months of time have passed with no recalibration (at least with the current state of electrode array technology). Our new supplemental figure confirms this. It shows that for sessions 6 months apart, changes in neural activity are substantial (SFig. 4A-B). However, these new figures also demonstrate that for more closely-spaced sessions, good performance can be achieved with no recalibration at all, or unsupervised recalibration that need not consume any user time (as it can run in parallel during normal use). We envision a usage scenario involving light recalibration (or unsupervised recalibration) during periods of regular use, combined with larger calibration datasets when users return from months of inactivity.

We understand, however, that there is a desire (and a need) to characterize nonstationarities systematically across many more subjects and larger datasets. We believe that the best way to do this is with a comprehensive study of many participants spanning many years. We are currently in the process of quantifying how neural signals recorded on intracortical electrode arrays change over time using data from all 14 BrainGate participants over a time span of 15 years, which we plan to report in a future publication. In this way we believe that we will be able to most meaningfully, and rigorously, contribute new insight on this important question.

[1] Downey, John E., Nathaniel Schwed, Steven M. Chase, Andrew B. Schwartz, and Jennifer L. Collinger. "Intracortical Recording Stability in Human Brain–Computer Interface Users." *Journal of Neural Engineering* 15, no. 4 (May 2018): 046016. https://doi.org/10.1088/1741-2552/aab7a0.

Specific comments:

Line 93: Why did the subject write 'periods as '~' and spaces as '>'?

Thank you, we should have clarified. We instructed T5 to write periods with a '~' symbol because we thought that '~' would be easier to detect than just a single dot. Similarly, we wanted to associate spaces with a symbol instead of just the absence of writing. The '~' and '>' symbols were chosen with an eye towards being easy to write and classify, but were not the result of a systematic study of which symbols would be the best. We added the following sentence of explanation to the Results section:

> The '~' and '>' symbols were chosen to make periods and spaces easier to detect.

Line 100: Clarify if the statement 'After each new day of decoder evaluation,' refers to offline or online decoding.

Thank you. Decoders were calibrated each day *before* they were evaluated online, using data collected from that day combined with all prior days.

We re-worded that section to now state the following:

> Prior to the first day of real-time use described here, we collected a total of 242 sentences across 3 days that were combined to train the RNN (sentences were selected from the British National Corpus). On each day of real-time use, additional training data was collected to retrain the RNN prior to real-time evaluation, yielding a combined total of 572 training sentences by the last day (comprising 7.3 hours and 30.4k characters). ~~After each new day of decoder evaluation, that day's data was cumulatively added to the training dataset for the next day (yielding a total of 572 sentences by the last day).~~

Line 112: How did the authors know the exact timing of completion of each letter by the subject in real time to be able to display it after it was completed? It is stated that visual feedback about the decoder output was 'estimated to be between 0.4-0.7'. The supplementary material explains how they arrived at these estimates, but this inherently assumes that the character was 'completed' when the start of a new one was detected. One can argue that natural handwriting of a word does not entail separating in time the representation of characters — they are all 'connected'.

Thank you, this is indeed an important and somewhat complex aspect that we should have explained more clearly. During real-time use, a simple thresholding scheme was used to decide when to decode and display each letter to the screen. Specifically, the RNN's "new character" output (see Fig 2A, reproduced below) was thresholded (threshold = 0.3). Whenever it crossed the threshold at time *t*, the most likely character at time *t*+0.3s was emitted. The most likely character was determined by examining the RNN's 'character' output.

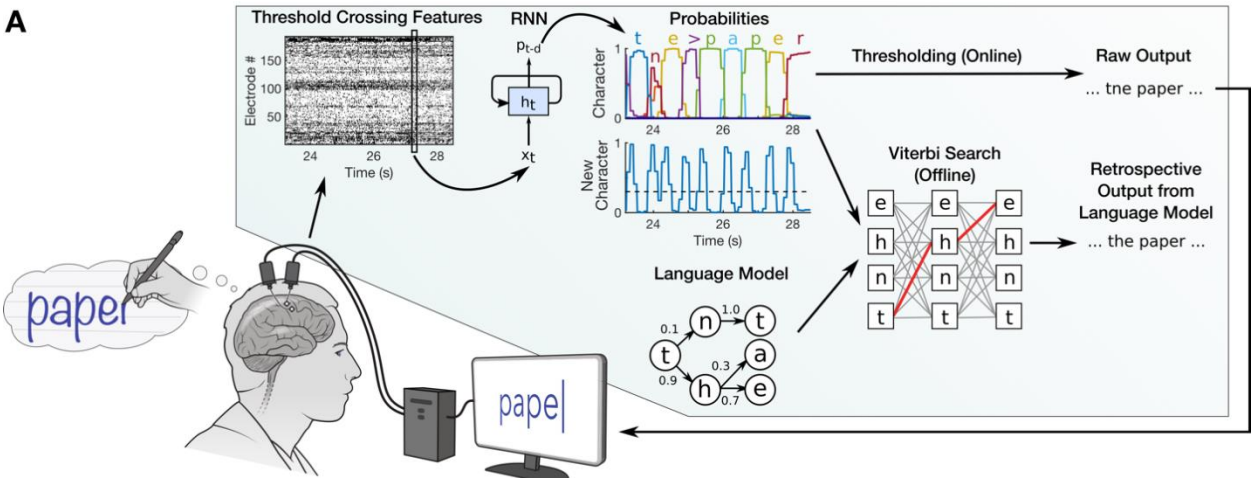


Figure 2A. Diagram of our decoding algorithm. First, the neural activity (multiunit threshold crossings) is temporally binned (20 ms bins) and smoothed on each electrode. Then, a recurrent neural network (RNN) converts this neural population time series (x_t) into a probability time series (p_{t-d}) describing the likelihood of each character and the probability of any new character beginning. The RNN has a one second output delay (d) so that it has time to observe the full character before deciding its identity. Finally, the character probabilities were thresholded to produce "Raw Output" for real-time use (when the "new character" probability crossed a threshold at time *t*, the most likely character at time *t*+0.3s was emitted and shown on the screen). In an offline retrospective analysis, the character probabilities were combined with a large-vocabulary language model to decode the most likely text that the participant wrote (we used a custom 50,000-word bigram model).

Given the absence of ground truth data about T5's attempted pen movements, we can only offer a "best guess" of the visual latency. In some sense, answering this question with certainty would require a complete solution to the original decoding problem posed here: segmentation and classification of characters from an unlabeled data stream. Given that our RNN decoder is not perfect, the RNN outputs can only offer a rough estimate of the latency.

Regarding the possibility of 'connected' characters, we do appreciate that there is some ambiguity and arbitrariness in defining exactly when a character ends and another begins, and that there is likely some 'transition time' which occurs between any two characters. Mitigating this issue somewhat is the fact that T5 reported writing each character in a print (not cursive) font, with each letter printed directly on top of the previous one as if writing on a PalmPilot. We added the following clarification to the Results section:

> T5 attempted to write each character in print (not cursive), with each character printed on top of the previous one.

Finally, it is unclear to us how exactly we could modify the estimated latency to account for the potential time spent transitioning between letters (since this transition time is unknown); as such, and after much discussion, we decided it would be best to keep the estimate as-is, with the understanding that it is only an estimate.

One can also argue that their approach (delaying the decoder output by 1 sec and adding the filter kernel widths to the total interval) prevents visual feedback about the state of neural activity until a complete character is encoded by the subject, but the reality is that the subject can 'covertly' infer information from the structure of the word being typed (self-generated case) and visual feedback from the screen (on-prompt case).

Thank you for pointing this out, we see now that visual feedback of the prompt and/or previously decoded letters could be used by T5. Due to this suggestion and others, we have removed any claim that our BCI can operate without visual feedback.

Line 115: How did neural activity look like when an error was made? and when the subject was provided visual feedback about the language model's autocorrection of that error? Did the subject stop modulating, eventually relying on the model to autocorrect, or did he continue to modulate neural activity to correct the typo? Was the decoder 'disengaged' in those instances? did the neural activity occupy different regions of the state space relative to the intended character or the corrected character?

Thank you for these interesting questions and suggestions. First, we want to clarify that the language model was only applied *offline* in a retrospective analysis and was never used online (i.e., T5 never saw the results of the language model).

Since there was no 'backspace' implemented, T5 was simply instructed to ignore errors and continue uninterrupted. T5 reported spending most of his time looking at the prompt during the copy-typing task, instead of watching the decoded letters appear on the screen and scanning for errors. We confirmed this using an eye tracker. Analyzing eye position data, we found that during copy-typing T5 spent 93% of the time looking at the prompt. T5 looked at the decoded text mostly at the end of each trial after all characters had been typed (but before he triggered the beginning of the next trial). During this "end-of-trial" period, T5 spent 82% of the time looking at the decoded text.

We added the following details to the Results section:

> Since there was no 'backspace' function implemented, T5 was instructed to continue writing if any decoding errors occurred. T5 reported spending most of his time looking at the prompt instead of watching the decoded letters appear on the screen (eye tracking data confirmed that T5 spent 93% of the time looking at the prompt; Tobii 4C eye tracker).

As suggested by the reviewer, it is interesting to ask how the perception of errors affect the neural activity. Recent reports from our group suggest that errors cause a distinct neural signature in motor cortex that can even be detected with a BCI and used to 'undo' errors [1-2]. We think this is an interesting line of future research for handwriting decoders.

[1] Even-Chen, Nir, Sergey D. Stavisky, Jonathan C. Kao, Stephen I. Ryu, and Krishna V. Shenoy. "Augmenting Intracortical Brain-Machine Interface with Neurally Driven Error Detectors." *Journal of Neural Engineering* 14, no. 6 (November 2017): 066007. https://doi.org/10.1088/1741-2552/aa8dc1.

[2] Even-Chen, N., S. D. Stavisky, C. Pandarinath, P. Nuyujukian, C. H. Blabe, L. R. Hochberg, J. M. Henderson, and K. V. Shenoy. "Feasibility of Automatic Error Detect-and-Undo System in Human Intracortical Brain–Computer Interfaces." *IEEE Transactions on Biomedical Engineering* 65, no. 8 (August 2018): 1771–84. https://doi.org/10.1109/TBME.2017.2776204.

Line 118: It is stated that the raw decoder output plateaued at 90 characters per minute with a 5.4% character error rate. But the comparison drawn in the sentence that followed argues that the 'word error rate' decreased to 3.4% average across all days. The authors should provide the reduction in 'character error rate' not 'word error rate' with the use of the language model to make this comparison objective. Arguably, many words share the same characters and understanding of words depends on the sentence context.

Thank you, we now mention both character error rate and word error rate in the Results text:

> Importantly, typing rates were high, plateauing at 90 characters per minute with a 5.4% character error rate (Fig. 2C, average of red circles). When a language model was used to autocorrect errors, error rates decreased considerably (Fig. 2C, open squares below filled circles; Table 1). The character error rate fell to 0.89% and the word error rate fell to 3.4% averaged across all days, which is comparable to state-of-the-art speech recognition systems (e.g. word error rates of 4-5% [15,16]) …

Line 120: it is stated that 'a new RNN was trained using all available sentences to process an entire sentence'. This means that offline decoding of an entire sentence achieved the stated 0.17% character error rate. As stated this decoder has not been used by the subject in real time to see if this newly trained decoder will be able to display an entire sentence at the end of a neural activity modulation epoch by the subject in the absence of the delayed character-by-character feedback as in the online case. As such, what is the significance of this result?

Thank you, we should have been clearer. In this analysis, we trained a bidirectional, acausal RNN to use *all* of the neural data in a sentence in order to decode that sentence (as opposed to using, for each time point *t*, only data that occurred prior to *t* (i.e., causal)). We see the significance of this result as two-fold: (1) providing a point of comparison to other work in the BCI and machine learning fields that process neural activity, handwriting or speech in an acausal manner, and (2) demonstrating a high ceiling for accurate performance, meaning that the trial-to-trial neural variability is not too great to prevent very high decoder performance.

As we see it, this result is mainly to provide more context and insight into the data, not necessarily to suggest that such a decoder be used in real-time as part of the BCI (which, as the reviewer points out, would not give the user character-by-character feedback). Nevertheless, it is possible to combine the causal decoder with the bidirectional decoder. One could use the causal decoder to give character-by-character feedback, and then run the bidirectional decoder at the end of each sentence to further clean up any decoding errors.

We added the following additional explanation to the Results section:

> Finally, to probe the limits of possible decoding performance, we retrospectively trained a new RNN using all available sentences to process the entire sentence in a non-causal way (comparable to other BCI studies [17,18]). In this regime, accuracy was extremely high (0.17% character error rate averaged across all sentences), indicating a high potential ceiling of performance. Although such an acausal decoder would not be able to provide letter-by-letter feedback to the user, it could be used to correct errors after the user finishes typing a sentence.

Table 1: Can the authors explain why the word error rate is so high (25.1%) in the raw online output case despite a character error rate of 5.9%?

Under the standard definition of word error rate, a word is incorrect if *any* character in that word is incorrect. On average, English words have five characters in them. Thus, with a character error rate of 5.9%, if we assume that each character independently has a 94.1% chance of being accurate, we might expect a word error rate of $1-(0.941)^5 = 26.2\%$. We added the following explanation to the table caption:

> Word error rates are high for "online output" because a word is considered incorrect if *any* character in that word is wrong.

Supplementary material:

Line 427: it is stated that "some micromotions of the right hand were visible during attempted handwriting (see 10 for neurologic exam results and SVideo 4 for hand micromotions). Have the authors quantified the extent of variance in the neural data that could be explained by this potential confound?

Thank you for raising this interesting question, which we have considered by did not clarify in the original manuscript. In our view, the potential leakage of motor commands into small amounts of muscle activity is not a confound here. First, we have added additional text to clarify the extent of T5's injury and paralysis, which is severe.

The description in the Results now reads:

> T5 has a high-level spinal cord injury (C4 AIS C) and was paralyzed from the neck down; his hand movements were entirely non-functional and limited to twitching and micromotion.

In the Methods section, we have added neurological exam data:

> Below the injury, T5 retained some very limited voluntary motion of the arms and legs that was largely restricted to the left elbow; however, some micromotions of the right hand were visible during attempted handwriting (see [12] for full neurologic exam results and SVideo 4 for hand micromotions). T5's neurologic exam findings were as follows for muscle groups controlling the motion of his right hand: Wrist Flexion=0, Wrist Extension=2, Finger Flexion=0, Finger Extension=2 (MRC Scale: 0=Nothing, 1=Muscle Twitch but no Joint Movement, 2=Some Joint Movement, 3=Overcomes Gravity, 4=Overcomes Some Resistance, 5=Overcomes Full Resistance).

Thus, we believe that T5 is a good / reasonable model of someone who could benefit from a communication BCI – i.e., someone who might be able to generate some hand micromotions but retains essentially no hand function. In our experience, severe paralysis is rarely fully complete. This is supported by a recent study of potential BCI users [1] that found that "incomplete" locked-in syndrome, which still prevented normal communication due to severe paralysis, was significantly more common than complete locked-in syndrome.

[1] Pels, Elmar G.M., Erik J. Aarnoutse, Nick F. Ramsey, and Mariska J. Vansteensel. "Estimated Prevalence of the Target Population for Brain-Computer Interface Neurotechnology in the Netherlands." *Neurorehabilitation and Neural Repair* 31, no. 7 (July 2017): 677–85. https://doi.org/10.1177/1545968317714577.

Second, we believe that the neural activity is generated primarily by the *intention* to move, and not overt motion itself. This is supported by a recent study from our group which included data from participant T5; in that work, we found that body parts which T5 still had control over (e.g. head, shoulder) did not have a stronger representation than body parts which were fully or almost fully paralyzed [2]. This view is also supported by a previous study on point-and-click BCIs from our group that included a participant (T6) who still retained hand function (and used thumb/index finger motor imagery to control the cursor). We found that we could achieve high performance whether or not the participant made overt finger motions, suggesting that the neural activity was primarily driven by motor intent and not, for example, sensory feedback generated by overt motion [3].

[2] Willett, Francis R., Darrel R. Deo, Donald T. Avansino, Paymon Rezaii, Leigh R. Hochberg, Jaimie M. Henderson, and Krishna V. Shenoy. "Hand Knob Area of Premotor Cortex Represents the Whole Body in a Compositional Way." *Cell*, March 26, 2020. https://doi.org/10.1016/j.cell.2020.02.043.

[3] Pandarinath, Chethan, Paul Nuyujukian, Christine H. Blabe, Brittany L. Sorice, Jad Saab, Francis R. Willett, Leigh R. Hochberg, Krishna V. Shenoy, and Jaimie M. Henderson. "High Performance Communication by People with Paralysis Using an Intracortical Brain-Computer Interface." *ELife* 6 (February 21, 2017): e18554. https://doi.org/10.7554/eLife.18554.

We added the following explanation to the *Methods* section where T5's injury is described in detail:

> In a recent study from our group which included data from participant T5, we found that body parts which T5 still had control over (e.g. head, shoulder) did *not* have a stronger representation than body parts which were fully or almost fully paralyzed[1]; thus, T5's limited hand motion likely did not have a large effect on the neural activity, which seems to be generated primarily by the *intention* to move and not overt motion itself.

Line 491: It would be informative for the authors to comment on how did the neural activity differ between repetitions of each character individually and when they are within a word or a sentence.

Thank you for this suggestion. While very interesting, as we articulated above, we think that examining the neural encoding of characters within words and sentences is not trivial, since it may introduce artifacts due to imperfect segmentation of words and sentences. Also, as it does not directly bear on the claims made in this manuscript, we prefer to leave this analysis for future work. Nevertheless, we do appreciate that examining how the context in which characters are written affects neural encoding is a useful and important direction that may yield decoder performance improvements and basic neuroscience insight.

D. Appropriate use of statistics and treatment of uncertainties:
Figures are well illustrated. Probability values and error bars are explained. There were no statistical significance tests performed.

Thank you.

Line 178: Authors should provide more explanation for "the participation ratio (PR), which quantifies approximately how many spatial or temporal axes are required to explain 80% of the variance in the neural activity patterns" in this section. Readers have to refer to the supplementary methods section to understand this metric.

Thank you for this suggestion. We do appreciate the desire to have every metric clearly explained as it is introduced in the Results. However, we think that referring to the Methods section, at least some of the time, is unavoidable in a Results-first format that is highly space-constrained like *Nature*. In our mind, to understand the result readers only need to know that this is a continuous quantification of dimensionality. However, to understand how the metric is computed seems to require an equation and a paragraph-sized description, which doesn't fit in the Results. Note that the metric is explained in the *Methods* section, not the *Supplementary Methods* (which are in an entirely separate document that contains much more detailed protocols).

We now offer the following additional clarification and refer the reader to the Methods section:

> Spatial and temporal dimensionality were quantified using the participation ratio (PR), which ~~quantifies~~ is a continuous quantification of approximately how many spatial or temporal axes are required to explain 80% of the variance in the neural activity patterns[21] (see Methods for details).

Line 192 Figure 3: The authors find that increased temporal complexity in neural state space trajectories could make movements easier to decode compared to trajectories that do not have such complexity, or have only spatial complexity. They then present a toy example in Figure 3 to make this point. I would partly disagree with their assessment and argument for the following reasons:

i) In the toy example in (Figure 3F) they increased variations of neural trajectories over time to illustrate that this increases separability (measured by nearest neighbor distance) compared to the case where the neurons' activity is constrained to a single spatial dimension, the unity diagonal). But the example lacks inclusion of noise, the temporal characteristics of which can easily 'fool' the classifier, making it think there is more temporal complexity in the trajectories than really is.

ii) The nearest neighbor distance and consequently classifier performance should be characterized when noise is present in this toy example, with a parameter that controls the amount of temporal complexity in noisy neural trajectories. Directions of fluctuations around these trajectories are likely to influence the conclusion made, both in the straight line as well as the handwritten characters cases.

Thank you for these important suggestions. Below we expand on our approach and rationale, and while in principle we are very open to adding this entire treatment to the manuscript's supplemental materials, we are facing space limitations such that we would need to seek guidance on how to be able to do this and if

it is possible at all. Thus, we thought that we would provide this explanation here and potentially go from there if there is still a need to do so.

First, we would like to clarify that this toy example does include noise. The three panels (F, G, H) from Fig. 3 (now Fig. 4) are reproduced below for convenience. Panel H shows how classification accuracy varies as a function of the amount of neural noise present (simulated as white noise). When there is no noise present, it is trivially easy to classify between the four conditions because there is no chance that one could be confused for another. Panel H shows that in the no-noise case (σ=0), there is no difference between the classification performance of "simple" trajectories (shown in F) and "complex" trajectories (shown in G) because classification performance is 100% for both. However, as the amount of noise increases, complex trajectories become easier to classify because their nearest neighbor distances are larger (and thus nearby trajectories are less likely to be confused with each other).

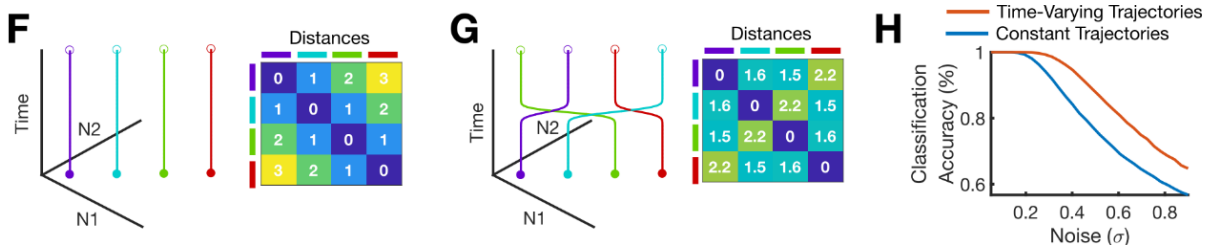


Fig. 3F-H. (Now Fig. 4).

Note that the temporal complexity (dimensionality) of the noise is much higher than that of the trajectories themselves. By definition, white noise occupies all possible temporal dimensions. In this toy example, we discretized the trajectories into 100 time steps; thus, the temporal dimensionality of the white noise was 100. The temporal dimensionality of the underlying neural trajectories themselves was much lower (1 for the simple trajectories, 2 for the complex trajectories). Why is the temporal dimensionality of the noise so high? Because white noise is independent for each time step, it requires one dimension for each time step in order to fully describe it. The underlying neural trajectories, on the other hand, are much smoother across time.

We added the following clarifications to the Methods section:

> … Thus, we performed the simulated classification ~~on~~ using the true neural patterns themselves (but still in the presence of observation noise). The simulated trajectories were discretized into 100 time steps and white noise was added to each time step independently.

Why, then, is the classifier not confused by the temporal complexity in the noise? To understand this, it may help to define some terms. Let f_a be a vector that describes the underlying neural trajectory for movement *a* (i.e., the mean neural firing rates across time for movement *a*). Each entry in the vector f_a is the mean firing rate for a given time step (to describe multiple neurons, the activity profile of each neuron can be stacked one on top of the other in the vector). Let ε be a neural noise vector for an example trial of movement *a*. The observed neural activity on that trial is then f_a+ ε. A classifier confusion will only happen for this trial if f_a+ ε looks more like the mean firing rates for a different movement (e.g. f_b) than it looks like f_a. We can formalize this notion of "looking like" with Euclidean distance; in other words, a confusion will occur if:

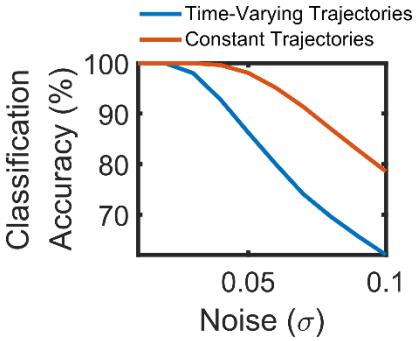
$$\|(f_a + \varepsilon) - f_b\| < \|(f_a + \varepsilon) - f_a\|$$

These confusions can be reduced if f_a and f_b look more different from each other, thereby reducing the chance that ε will corrupt f_a into looking like f_b. In other words, classification performance is improved when nearest neighbor distances are increased. Temporal variety is just one way to increase this distance (spatial variety is another). In our view, the temporal complexity of the noise ε is thus not

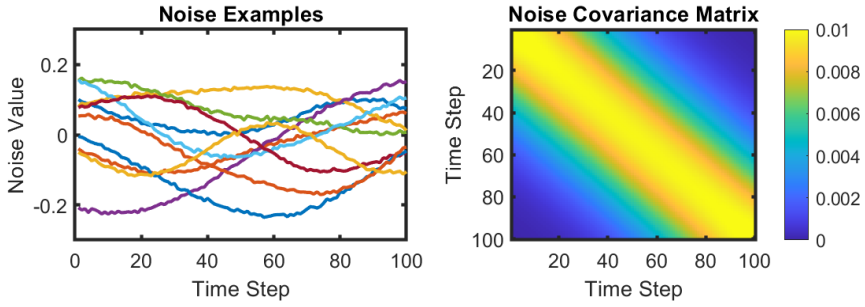
necessarily important here (although its size is – the larger ε is, the greater the chance that it can cause f_a to look like f_b).

One noise property that can end up making a big difference for performance is the "shape" of the noise cloud, or in other words the directions along which ε is particularly concentrated (as suggested by the reviewer). White noise extends equally in all directions, but the most relevant directions for classification are those directions that connect nearby classes (here, this direction would be $f_b - f_a$). This is because, for f_a to be corrupted into looking like f_b, ε must be similar to $f_b - f_a$ (since in this case $f_a + \varepsilon = f_a + f_b - f_a = f_b$). If anything, we think this is actually another reason to prefer movements with higher temporal dimensionality. Larger temporal dimensionality will cause the directions between nearby classes to be more diverse, and less likely to align with directions that contain large amounts of noise. In our experience, large-noise directions typically describe correlated increases and decreases in firing rates across time. Thus, having movements which are more complex in time will make them more robust to correlated noise fluctuations.

To confirm this, we simulated classification performance using "colored" noise with concentrated power in lower frequencies (i.e. correlated noise). The results obtained were the same as in panel H, except with an even greater difference between the time-varying trajectories and constant trajectories:



The plot below shows examples of the noise vectors (to show how they are correlated in time) and the covariance matrix used to generate this noise (by drawing random samples from a multivariate normal distribution). The diagonal band causes nearby time steps to have correlated noise.



Line 244: Authors state that "One unique advantage of our handwriting BCI is that, in theory, it does not require vision (since no feedback of the imagined pen trajectory is given to the participant, and letters appear only after they are completed)." I would argue against that, partially because this claim is contingent on: 1) exact knowledge of the length of time interval where each decoded character is fully known and, 2) the instructed text was always present on the screen in the on-prompt case. To my understanding this was estimated (see my comment on Line 112 above) based on approximations made by the delayed decoder training and time warping algorithm (1.4 sec delay), which was used offline to build spatiotemporal neural "templates" of the characters.

Thank you, as stated above we have removed this claim about visual feedback.

Line 534: Please clarify what is a 'single movement condition'. Is it a character, a word or a sentence? From line 801 it seems it corresponds to character but the earlier sentence needs clarification.

Thank you for pointing this out. Indeed, we had meant to refer to a character. We have rephrased the sentence as follows:

> Next, we used time-warped PCA (https://github.com/ganguli-lab/twpca)[8,9] to find continuous, regularized time-warping functions that align ~~the~~ all trials ~~within a single movement condition~~belonging to the same character together.

Line 553: Authors used character templates drawn by a computer mouse in the same way as T5 described writing the character. This description provides a shape of the character but it is unclear how this information was translated into pen velocity to train the decoder.

Thank you, we should have been clearer about this. As each character was drawn, we recorded the X and Y velocity of the mouse pointer. We have clarified this by adding the following sentence:

> As each character was drawn, the X and Y velocity trajectories of the mouse pointer were recorded.

Line 577: "the criteria for excluding data points from display in Figure 1E is not clear. It is stated that these data labeled as "outliers in each class" were excluded "To make the t-SNE plot clearer". While it is stated that this resulted in removing 3% of data points, the explanation that these "were likely caused by lapsed attention by T5" is not convincing. How did the authors ascertain that this was the case?

Thank you for raising this interesting and important point. T5 reported that he would occasionally fail to complete a trial due to a lapse in attention; however, it is true that there is no easy way to know whether any particular outlier was due to a lapse in attention or some other cause. We have therefore remade the plot with *all* trials included; the result is very similar, save for some outliers that may be distracting but don't change the core result. We reproduce the new Fig. 1E below, and this now appears in the main paper:

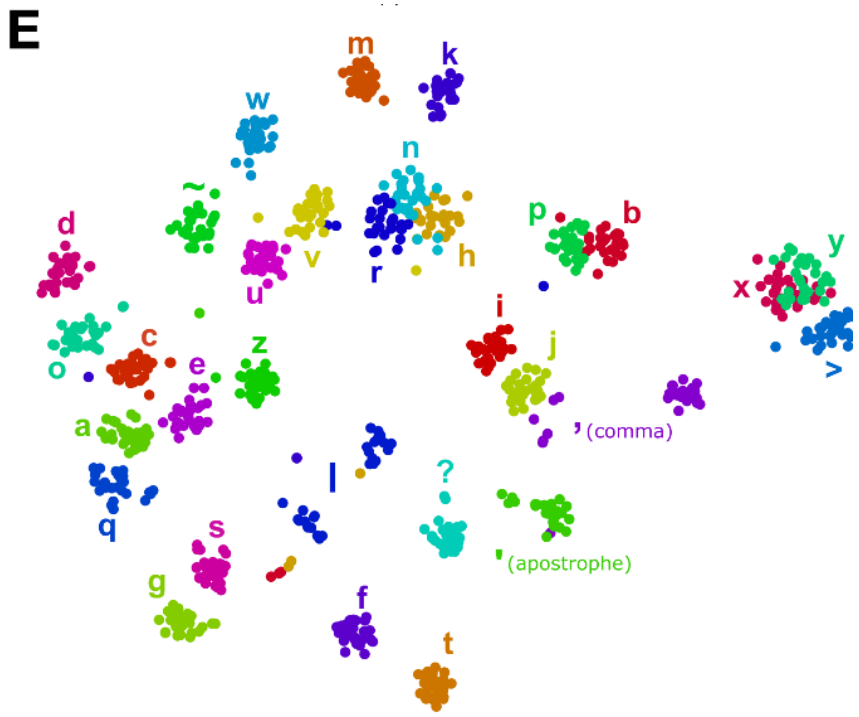


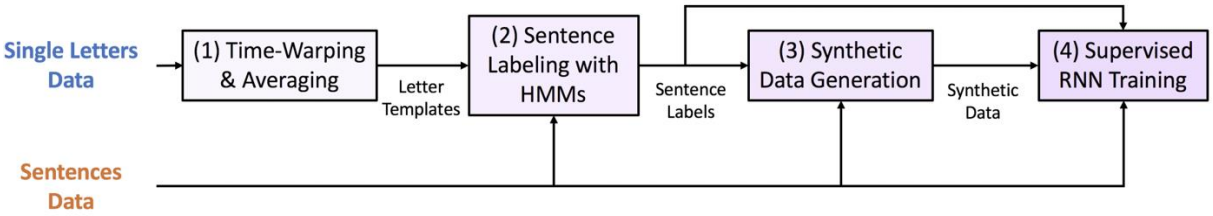
Fig. 1E

Supp Fig 2 and lines 642-667: The authors use a technique from automatic speech recognition literature called forced alignment labeling with HMMs in which they augmented the data via synthetic sentence generation to cope with the limited data size. This section needs improvement regarding how the method works. For example, creating snippets to make synthetic sentences assumes the neural data correspouding to each snippet is independent of the others. How it is then integrated into a new synthetic sentence that is then labeled by the HMM? How 'one-hot representation' is defined based on the heatmaps generated in SF-2D?

Thank you for these questions. Note that there is a detailed explanation of each step in the "Supplemental Methods" document, which is a separate document from the Methods section and contains much more detail about each step. We now make this clearer in the "RNN Training" section in the Methods:

> See SFig. 2B for a diagram of the RNN training flow and Supplemental Methods for a detailed protocol (the Supplemental Methods are contained in a separate online document).

The four steps of RNN training are illustrated in SFig 2B (reproduced below). In step (2), the training data is segmented & labeled with HMMs using a "forced alignment" technique. This step determines, for each time step of data, what character was being written during that time step. In step (3), synthetic sentences are created to 'augment' the data (i.e., these synthetic sentences are added to the original data as additional examples for the RNN to train with). Synthetic sentences are created 'cutting out' the characters from the training data (with the help of the labels from step 2) and placing them into a library of character snippets. These snippets are then re-arranged randomly into new sentences.



SFig. 2B.

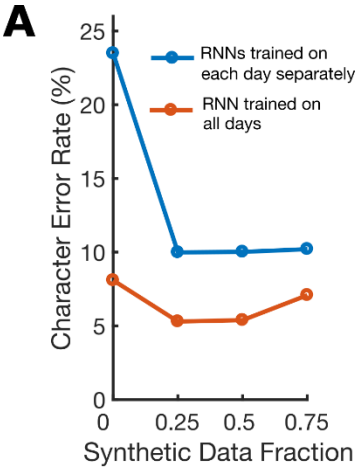
The synthetic sentence creation step does indeed assume that the characters are independent from each other, with one exception: it attempts to choose character examples such that each adjacent pair of characters has matching transition characteristics. This is explained in the Supplemental Methods as follows (in the "Synthesizing the Neural Activity" section):

> For each character, a snippet was chosen from the library at random in a way that attempted to respect pen transition movements between letters. For example, when transitioning from 'e' to 't', the pen must traverse upwards before beginning the downstroke for 't'. However, when transitioning from 'd' to 't', no such pen re-positioning is needed (when written in the way shown in Figure 1). To do this, we discretized the starting heights for each character to the following values: 0, 0.25, 0.5, 1. The assignment of each letter to each category is depicted in the table below.

Start Height	0	0.25	0.5	1.0
Character	comma	a, o, e, g, q	c, d, m, j, i, n, p, r, s, u, v, w, x, y, z, space (>), period (~)	b, t, f, h, k, l, apostrophe, question mark

> When choosing a snippet from the library, we selected at random from all snippets whose next character in the training data began at the same height as the next character in the synthetic sentence. When this wasn't possible, we selected uniformly at random from all snippets.

While these assumptions are simplistic, the main point is that the synthetic data are good enough to significantly improve decoder performance. Supplemental Figure 3A, reproduced below, shows results from an offline analysis that assesses how adding synthetic data reduces the character error rate.



SFig. 3A.

For RNNs trained on a single day, adding synthetic data reduced the character error rate percentage by 12.9 (95% CI = [11.9, 14.0]). For RNNs trained on all the days, adding synthetic data reduced the

character error rate percentage by 2.7 (95% CI = [2.2, 3.3]). A greater performance improvement for single-day RNNs makes sense, as data augmentation is likely to help more when the data is scarcer.

We added the following clarification to the main Methods section:

> Although this method is simplistic in that it assumes that the neural representation of a character is independent of past and future characters, it was nevertheless important ~~This data augmentation step was critical~~ for achieving high performance (decreased the error rate percentage by 12.9 when training on single days and 2.7 when training on all days; SFig. 3A).

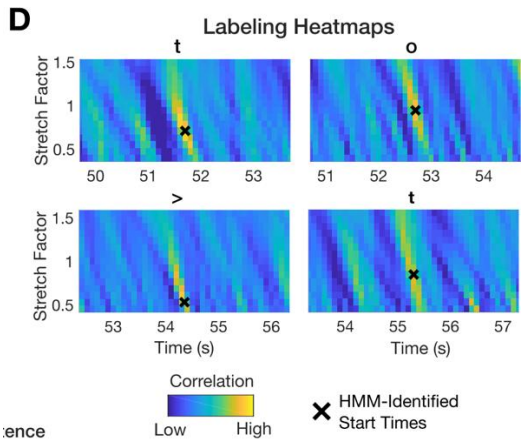
Note that there is no need to label the synthetic sentence with the HMM, since the character identities at each time step are already known. All that is needed is to straight-forwardly construct a time series of probability "targets" that the RNN is trained to output when the synthetic data is given as input. These targets consisted of (1) a one-hot representation of the active character at each time step and (2) a binary "new character" signal which went high whenever a new character started (and remained high for 200 ms). A "one-hot" representation simply means a vector whose entries are all equal to zero except for the entry corresponding to the currently active character (which is equal to one). The following text from the Supplemental Methods defines these target variables in detail:

> Step 6: Construct RNN Targets Finally, target variables for supervised RNN training were generated using the letter start times found above. Two target time series were created: a series of one-hot character vectors (y_t), where each vector is a one-hot representation of the most recently started character, and a scalar time series (z_t) that indicates whether *any* new character has recently been started. The z_t signal allows repeated characters to be distinguished (these would otherwise appear identical to a longer, single character as seen through y_t).
>
> Intuitively, y_t is a 'sample and hold' signal that stores whatever the most recently started character was indefinitely. For example, even if T5 pauses for several seconds after writing the character "a", y_t will still continue to reflect "a" indefinitely until a new character is started. The z_t signal is a complementary binary signal that goes high for a brief time whenever *any* new character begins. z_t can be thresholded to detect the presence of new letters and type them on the screen, which we did online. More formally, y_t and z_t were defined as follows:
>
> $$y_{t,i} = \begin{cases} 0, & \text{the most recently started character was not i} \\ 1, & \text{the most recently started character was i} \end{cases}$$
>
> $$z_t = \begin{cases} 0, & \text{the most recent character was started} > 200 \text{ ms ago} \\ 1, & \text{the most recent character was started} \leq 200 \text{ ms ago} \end{cases}$$

We added extra text to the Methods section to clarify the definition of a "one-hot" representation:

> The vector of target character probabilities (denoted as y_t above) was constructed by setting the probability values at each time step to be a one-hot representation of the most recently started character (i.e., the most recently started character's entry in the vector is equal to 1 while all other entries are 0).

Note that the heatmaps shown in Supplemental Figure 2D (and reproduced below for convenience) were only used to qualitatively assess whether the HMM labeling process succeeded in a reasonable way. The heatmaps themselves were not directly used to construct the RNN targets.



SFig. 2D.

The only thing that was used to construct the RNN targets were the "HMM-identified Start Times", i.e. the time steps when each character began to be written in the training data (as determined by the HMM). Since the heatmaps show hotspots around these HMM-identified start times, we can infer that the labeling process was reasonably accurate (this is just a useful method for sanity checking the labeling). The true proof of the labeling process is the high performance of the RNN decoder that results from using those labels. We added the following disclaimer to the supplemental figure legend:

> Note that these heatmaps are depicted only to qualitatively show label quality and aren't used for training (only the character start times are needed to generate the targets for RNN training).

E. Conclusions

The conclusions are generally based on findings in the work performed in One subject. At times though there are some overstatements about the far reaching ability of the technology which should be scaled down. For example, I did not find the conclusion that this is a BCI without visual feedback to be convincing. If it were, then how can the authors explain the difference in performance between the on-prompt typing and self-paced typing? It is unclear whether there was any type of eye tracking to determine the type of visual feedback the subject was receiving at each moment. For example, was the subject always staring at the text prompt, or was the subject always looking to the decoded characters? Or a combination of both? unless they have an objective measure of visual feedback, it is unclear whether the BCI was truly operating without vision as claimed.

Thank you for these points about visual feedback. As explained above, we have eliminated the claim that the BCI can operate without visual feedback.

F. Suggested improvements:

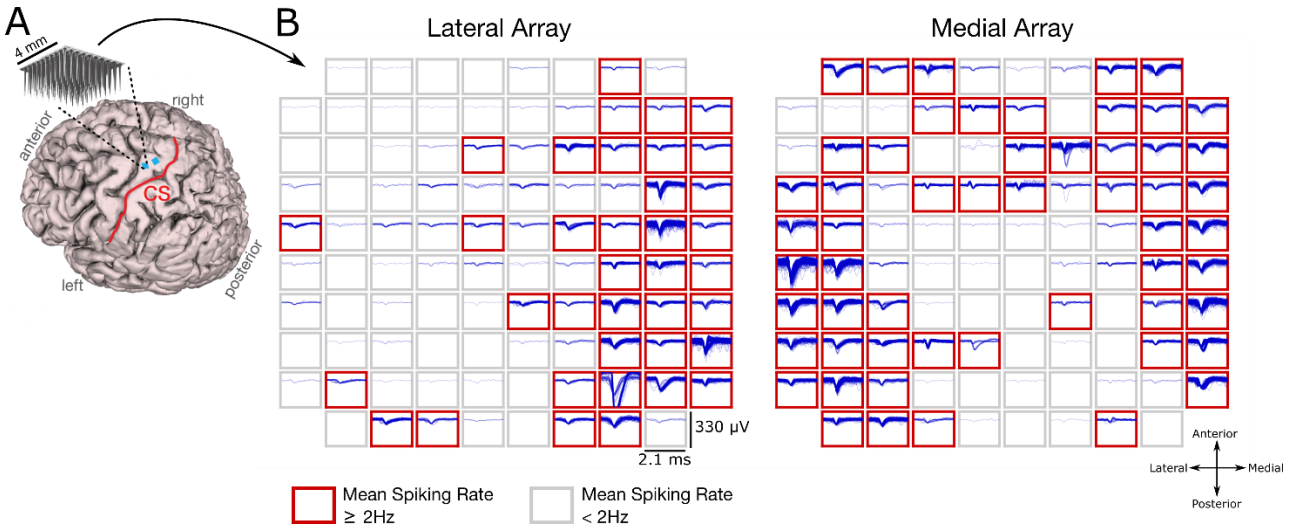
In addition to the above, I think a critical experiment/analysis to be performed is one in which the authors characterize the longevity and stability of representation of neural signals of the decoded variable(s). The extensive calibration process indicates that the data is highly nonstationary but none of this is characterized.

We thank the reviewer again for these insightful and important suggestions, which we believe have significantly improved the paper by leading us to perform new analyses that demonstrate the feasibility of training decoders with a more limited calibration process. Again, this is all described in detail above.

Based on a few published studies, it is reasonably expected that the implanted device can leverage single cell resolution of neural spiking signals within the first year of implant. However, authors used multiunit activity (binned threshold crossing), implying the activity could not be spike sorted to reveal individual

neuronal activity encoding of the pen tip velocity. More explanation should be provided on how the nonuniform distribution of session dates affected the data quality. Authors explain in the supplementary material that this approach allowed them to “leverage information from more electrodes, since many electrodes recorded activity from multiple neurons that could not be precisely spike-sorted into single units.” Although they cite a paper from their group that demonstrated that neural population structure can be accurately estimated from threshold crossing rates alone, it is unclear if sorting spikes from a lower number of electrodes (which they did not state) on which single units could be identified would provide similar results.

Thank you for these questions. First, we would like to clarify that our use of threshold crossings was not motivated by an inability to spike-sort single neuron activity. As stated above, we added a new supplemental figure to demonstrate that high-quality spiking activity can still be recorded on these arrays 1200 days post-implant (reproduced again below for convenience).



Supplemental Figure 6. Example spiking activity recorded from each microelectrode array. (A) Participant T5’s MRI-derived brain anatomy. Microelectrode array locations (blue squares) were determined by co-registration of postoperative CT images with preoperative MRI images. (B) Example spike waveforms detected during a ten second time window are plotted for each electrode (data were recorded on post-implant day 1218). Each rectangular panel corresponds to a single electrode and each blue trace is a single spike waveform (2.1 millisecond duration). Spiking events were detected using a -4.5 RMS threshold, thereby excluding almost all background activity. Electrodes with a mean threshold crossing rate ≥ 2 Hz were considered to have ‘spiking activity’ and are outlined in red (note that this is a conservative estimate that is meant to include only spiking activity that could be from single neurons, as opposed to multiunit ‘hash’). Results show that many electrodes still record large spiking waveforms that are well above the noise floor (the y-axis of each panel spans 330 µV, while the background activity has an average RMS value of only 6.4 µV). On this day, 92 electrodes out of 192 had a threshold crossing rate ≥ 2 Hz.

We added the following sentence to the Methods section to give the reader a better understanding of the data quality:

> Note that both arrays still recorded high-quality spiking activity from many electrodes; on average, 81.9 ± 5.6 (mn ± sd) out of 192 electrodes recorded spike waveforms each day at a rate of at least 2 Hz when using a spike-detection threshold of -4.5 RMS (see SFig 6).

Additionally, we now clarify that our decision to use multiunit threshold crossings was not because spike waveforms could no longer be recorded on the arrays:

We used multiunit threshold crossing rates as neural features for analysis and neural decoding (as opposed to spike-sorted single units). This was not because spike waveforms could not be recorded (see SFig 6 for examples); rather, u~~U~~sing multiunit threshold crossings allowed us to leverage information from more electrodes, since many electrodes recorded activity from multiple neurons that could not be precisely spike-sorted into single units.

In our experience, using threshold crossings is simpler, can lead to higher performance (for BCIs), higher signal-to-noise ratios (for neural encoding analyses), and greater stability since action potential waveforms are able to grow/shrink some without affecting threshold crossing detection. Partly though, this depends on how spike-sorted neurons are defined. If one includes *only* well-isolated single neurons, then this excludes a lot of potential data and decreases BCI performance [1]. Good performance can be achieved by spike-sorting more liberally and including multiunit clusters, but this approach does not seem to have clear advantages over multiunit threshold crossings alone [1]. Because threshold crossings have been demonstrated to perform just as well (or within 5% at most) as spike-sorted clusters for BCI applications [1] and for analyzing neural population structure [2], we have chosen to use multiunit threshold crossings throughout our paper. Since these ideas have already been demonstrated in prior work from several nonhuman primate groups and clinical trial groups, we believe it is not necessary to revisit this issue by comparing our multiunit results to spike-sorted results.

[1] Christie, Breanne P., Derek M. Tat, Zachary T. Irwin, Vikash Gilja, Paul Nuyujukian, Justin D. Foster, Stephen I. Ryu, Krishna V. Shenoy, David E. Thompson, and Cynthia A. Chestek. "Comparison of Spike Sorting and Thresholding of Voltage Waveforms for Intracortical Brain–Machine Interface Performance." *Journal of Neural Engineering* 12, no. 1 (December 2014): 016009. https://doi.org/10.1088/1741-2560/12/1/016009.

[2] Trautmann, Eric M., Sergey D. Stavisky, Subhaneil Lahiri, Katherine C. Ames, Matthew T. Kaufman, Daniel J. O'Shea, Saurabh Vyas, et al. "Accurate Estimation of Neural Population Dynamics without Spike Sorting." *Neuron* 103, no. 2 (July 17, 2019): 292-308.e4. https://doi.org/10.1016/j.neuron.2019.05.003.

Finally, the reviewer writes that "More explanation should be provided on how the nonuniform distribution of session dates affected the data quality". Since good BCI performance and/or neural encoding results were achieved on all reported dates, we would propose that data quality is reasonably high throughout. Beyond this, we are unsure what particular question the reviewer might be raising related to the nonuniform distribution of dates, but we believe that we have addressed it above when we provided analyses of how much recalibration is needed depending on the time between sessions. Session dates were nonuniform due to (1) variability in the time needed to analyze data, develop decoding techniques, and prepare experiments and (2) fundamental constraints of the clinical trial, which sometimes preclude regular data collection due to outside demands on the participant and/or unrelated experiments taking priority.

G. References: appropriate credit to previous work?
Mostly relevant and appropriate. The work could benefit from a few more citations that documented the idea of training decoders from 'desired' behavioral templates when overt movements could not be performed.

Thank you for this suggestion. We have added the following references to the Methods section where training velocity decoders to reconstruct pen trajectories is discussed:

As each character was drawn, the X and Y velocity trajectories of the mouse pointer were recorded.~~;~~ T~~t~~hese templates then defined the target velocity vector for the decoder on each time step of each trial, much like prior work has trained decoders to predict the user's 'intended' velocity for continuous movement BCIs [10,11].

[10]Collinger, Jennifer L, Brian Wodlinger, John E Downey, Wei Wang, Elizabeth C Tyler-Kabara, Douglas J Weber, Angus JC McMorland, Meel Velliste, Michael L Boninger, and Andrew B Schwartz. "High-Performance Neuroprosthetic Control by an Individual with Tetraplegia." *The Lancet* 381, no. 9866 (February 2013): 557–64. https://doi.org/10.1016/S0140-6736(12)61816-9.

[11]Gilja, Vikash, Chethan Pandarinath, Christine H. Blabe, Paul Nuyujukian, John D. Simeral, Anish A. Sarma, Brittany L. Sorice, et al. “Clinical Translation of a High-Performance Neural Prosthesis.” *Nature Medicine* 21, no. 10 (October 2015): 1142–45. https://doi.org/10.1038/nm.3953.

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

No issues.

Thank you.

Again, we deeply appreciate all of these helpful questions and recommendations!

Reviewer Reports on the First Revision:

Referee #2 (Remarks to the Author):

The authors have adequately addressed my concerns and I feel it is suitable for publication. I congratulate them on an impressive work.

Referee #3 (Remarks to the Author):

The revised manuscript has substantially improved in a number of aspects. First, the authors have considerably scaled down some unsubstantiated claims (such as the visual feedback). They also clarified the difference between imagined and attempted movements in their explanation of the results. Second, the authors present new results to address some of my comments. In particular, they clarified that the work is primarily a classification approach of discrete neural activity states as opposed to continuous decoding. They also propose an unsupervised decoder recalibration method using language models that can achieve high performance without interrupting the user. This benefits from the existence of language models to streamline the recalibration process which has been a major element to combat neural signal variability that undoubtedly has an effect on their primary outcome measure: typing speed.

The authors, however, suggested that some of my proposed improvements should be part of future manuscript(s), particularly comments related to longevity of signals affecting decoding reliability and robustness. I think the authors understood my argument in the wrong context -- that their approach should be ready for prime time deployment in clinical applications which was not what I intended. My issue has to do with the level of explanations provided given the results they observed, which I feel is not at the level of the findings and can still be improved. Let me take some space below to clarify what I mean.

The work is primarily a multi-hierarchical classification of neural population dynamic states that starts with non-linear transformation of raw, thresholded activity to create the spatiotemporal templates to be used later for classification (example illustrated nicely in Fig 1E). In the context of online decoding, any features extracted from neural activity will be affected by variability resulting from multiple factors (e.g. array longevity, plasticity in neuronal tuning, attentional levels, etc). The variability can be quantified through two elements: 1) signaling – which has to do with the quality of spikes and robustness of sorting to permit reliable extraction of firing rates from as many single units. This was not quantified in this work because they did not do spike sorting (see comment above). And 2) information coding – which has to do with the actual representation of characters being encoded in the neural activity. Again, this was not characterized in this work because the authors stated that this is out of the manuscript focus and should be the topic of a future manuscript.

Interestingly, the authors demonstrate in new Supp Fig 4 how much of this variability resulted in variations in the decoded spatiotemporal templates. In particular, the shrinkage effect that the new figure shows highlights the main issue that I have raised. Somehow this information is embedded in the 'new' knowledge that the RNN learns with continued additions of new templates to the training dataset. However, there should be more in depth explanation or discussion on how this observation (as well as the 7-day stability result they also found, see related comment below) could enhance our understanding of handwriting movement representation in the brain. At the least, more discussion should be included regarding how decoders should be engineered to account for movements that have similar structured temporal variations (which could be very useful for other types of sequential movements). While it is not a major flaw and the modified text helps, I think the authors need to improve the discussion related to these two points in particular.

The new result in Figure 3 shows offline decoding performance when less than 50 calibration

sentences were used. While the result is interesting, the authors need to put it into perspective given that online decoding performance does diminish considerably compared to offline simulations, as their own results in Table 1 have shown. How would this result carry over to the online decoding case? how this performance is a function of character probability in these sentences, as well as the particular choice of the 10 sentences?
Another related issue is the explanations given to the difference between online and offline decoding performance. For example, it is well established that well isolated unit spiking does provide more information compared to local field potentials for BCI decoding but comes at the cost of increased computational complexity and variability over time, both within session and across sessions. I did not find their argument about not using spike sorting to be particularly compelling. Even though it is a subjective process as they state – but so is the thresholding process they've used, it has the potential to increase their information rate and consequently typing speed which is their main outcome measure. Unless it was observed that this process does somehow affect the spatiotemporal templates they use in the classification, the reasons for not using putative single or multi-unit clusters of waveforms in building the firing rate templates are not entirely clear.

It is also important to explain why 7 days or less seem to maintain uncalibrated decoder accuracy. It is important to cite prior published work in which it was demonstrated that the same duration tends to also be associated with stability of single unit spiking1,2. Is this a coincidence? I think the same is happening here, that decoders need to be calibrated because unit spiking and character representation seems to shift over intervals > 7 days.
1. Dickey, Adam S., et al. "Single-unit stability using chronically implanted multielectrode arrays." Journal of neurophysiology 102.2 (2009): 1331-1339.
2. Eleryan, Ahmed, et al. "Tracking single units in chronic, large scale, neural recordings for brain machine interface applications." Frontiers in neuroengineering 7 (2014): 23.

Specific comments:
Authors state that "The character templates were made by drawing each character with a computer mouse in the same way as T5 described writing the character. As each character was drawn, the X and Y velocity trajectories of the mouse pointer were recorded.;" I can understand how T5 can describe how the final shape would look like, but it's unclear how can T5 describe the velocity by which he attempted to write different parts of the characters from which the target velocity vector for the decoder was defined.
Line 115: How did neural activity look like when an error was made? and when the subject was provided visual feedback about the language model's autocorrection of that error? Did the subject stop modulating, eventually relying on the model to autocorrect, or did he continue to modulate neural activity to correct the typo? Was the decoder 'disengaged' in those instances? did the neural activity occupy different regions of the state space relative to the intended character or the corrected character?
Authors report that T5 spent 93% of the time looking at the prompt in the copy-typing task. They should also state what the eye tracking statistics were in the free typing trials in which there was no prompt. They also did not respond to the question if the perception of errors had any influence on the decoded patterns, particularly given that they cite their own work showing how errors result in distinct signature in motor cortex. They should clarify if and how these signatures, particularly when an error was made in the free typing trials, was handled by the decoder.

Authors have responded to my comment about the toy example by stating that the example included noise. My reference was to Figure 3F in which noise is absent in the trajectories shown. While I agree with the authors that "as the amount of noise increases, complex trajectories (may) become easier to classify because their nearest neighbor distances are larger (and thus nearby trajectories are less likely to be ", this is the case only under two assumptions: 1) The noise is white (as they have simulated already) and 2) the signal (i.e. the trajectory being classified) is uncorrelated with that noise. In the brain, however, there is ample evidence to suggest that the noise is not white, and is strongly correlated with the signal, and if not, it would at least be correlated among adjacent electrodes.

As they show in their inequality, when the distance between noisy fa and fb is less than the noise norm, misclassification will happen. However, if fa is more different than fb but the noise is more similar (i.e. correlated) to fb, adding that noise to fa will bring the noisy fa closer to fb, thus increasing the probability of misclassification. This is the case when the noise cluster extends along specific directions closer to those spanned by the signals. The examples simulated in the rebuttal use temporally colored noise, but the extent to which it is correlated with the actual signals being classified is unclear. I suggest that the authors bring more realism into their toy example regarding the noise characteristics (especially its temporal correlation with the signal while keeping its variance small) to make the explanation they are offering more compelling.

Overall, I think the work is very valuable and would be an important contribution to the field, provided the authors address some of the remaining issues above.

Author Rebuttals to First Revision:

Reply to Reviewers – Round 2

Note: reviewers' comments appear in **black text**. Our replies appear in **blue text**, and revised manuscript text appears indented (with old text shown in **black** and new edits in **red**).

Referee #3 (Remarks to the Author):

The revised manuscript has substantially improved in a number of aspects. First, the authors have considerably scaled down some unsubstantiated claims (such as the visual feedback). They also clarified the difference between imagined and attempted movements in their explanation of the results. Second, the authors present new results to address some of my comments. In particular, they clarified that the work is primarily a classification approach of discrete neural activity states as opposed to continuous decoding. They also propose an unsupervised decoder recalibration method using language models that can achieve high performance without interrupting the user. This benefits from the existence of language models to streamline the recalibration process which has been a major element to combat neural signal variability that undoubtedly has an effect on their primary outcome measure: typing speed.

Thank you for this kind summary and recognition of the effort we devoted to address your helpful questions and improve the manuscript. We are also grateful for the additional suggestions detailed below. We have done our best to address them within the strict space-constraints of a Nature article, which restricts the main text to 6.0 pages (manuscript length before this final revision was 6.7 pages, due to the inclusion of many excellent requests by all three reviewers).

The authors, however, suggested that some of my proposed improvements should be part of future manuscript(s), particularly comments related to longevity of signals affecting decoding reliability and robustness. I think the authors understood my argument in the wrong context -- that their approach should be ready for prime time deployment in clinical applications which was not what I intended. My issue has to do with the level of explanations provided given the results they observed, which I feel is not at the level of the findings and can still be improved. Let me take some space below to clarify what I mean.

Thank you for these helpful clarifications.

The work is primarily a multi-hierarchical classification of neural population dynamic states that starts with non-linear transformation of raw, thresholded activity to create the spatiotemporal templates to be used later for classification (example illustrated nicely in Fig 1E). In the context of online decoding, any features extracted from neural activity will be affected by variability resulting from multiple factors (e.g. array longevity, plasticity in neuronal tuning, attentional levels, etc). The variability can be quantified through two elements: 1) signaling – which has to do with the quality of spikes and robustness of sorting to permit reliable extraction of firing rates from as many single units. This was not quantified in this work because they did not do spike sorting (see comment above). And 2) information coding – which has to do with the actual representation of characters being encoded in the neural activity. Again, this was not characterized in this work because the authors stated that this is out of the manuscript focus and should be the topic of a future manuscript.

Thank you for this clarification. It is true that characterizing single neuron spiking quality and/or the neural representation of handwriting are not the focus of this work (although we do think Fig. 1 provides some important characterization of the neural representation of handwriting, by showing that pen-tip velocity can be decoded and by providing a low-dimensional visualization of the neural population structure via t-SNE).

Interestingly, the authors demonstrate in new Supp Fig 4 how much of this variability resulted in variations in the decoded spatiotemporal templates. In particular, the shrinkage effect that the new figure shows highlights the main issue that I have raised. Somehow this information is embedded in the 'new' knowledge that the RNN learns with continued additions of new templates to the training dataset. However, there should be more in depth explanation or discussion on how this observation (as well as the 7-day stability result they also found, see related comment below) could enhance our understanding of handwriting movement representation in the brain.

Thank you for this suggestion. As the reviewer mentions below, one likely reason for the changes in multiunit neural activity we observed over time (including the shrinkage effect) is the instability of spiking activity as observed through microelectrode arrays, an unknown fraction of which is caused by device micromotion. Therefore, we think that any changes in multiunit neural activity over time do not necessarily provide insight into neural plasticity or neural representations, as an unknown portion of that change is due to array micromotion. In the main text, when discussing decoder retraining, we now clarify for readers that the source of the neural changes may be due to plasticity or device micromotion:

> Retraining helps account for changes in neural recordings that accrue over time (which might be caused by neural plasticity or electrode array micromotion).

At the least, more discussion should be included regarding how decoders should be engineered to account for movements that have similar structured temporal variations (which could be very useful for other types of sequential movements). While it is not a major flaw and the modified text helps, I think the authors need to improve the discussion related to these two points in particular.

We appreciate this suggestion. The new Results section added in the previous revision highlights extensively the fact that neural decoders are negatively affected by changes in neural activity over time and typically require frequent retraining to combat this (either with explicit calibration data or via unsupervised retraining). After reporting our new analyses on this point, we offer the following interpretation for how decoders might be designed to be robust to temporal variations that have the medium-length time scale shown in our analyses:

> The above results are promising for clinical viability, as they suggest that unsupervised decoder retraining, combined with more limited supervised retraining after longer periods of inactivity, may be sufficient to achieve high performance.

While we too see the value in discussing this issue more thoroughly, particularly with regards to the interesting temporal shrinkage effect now shown in Extended Data Figure 4, the strict space

constraints of a Nature article prevent us from doing so (without removing other central results or discussion points). We have discussed this, and other space limitation restraints with the Editor to be sure that we are balancing this appropriately.

The new result in Figure 3 shows offline decoding performance when less than 50 calibration sentences were used. While the result is interesting, the authors need to put it into perspective given that online decoding performance does diminish considerably compared to offline simulations, as their own results in Table 1 have shown. How would this result carry over to the online decoding case? how this performance is a function of character probability in these sentences, as well as the particular choice of the 10 sentences?

Thank you for these suggestions. We now more explicitly highlight that these new analyses were performed offline and thus require future work to confirm online:

> … unsupervised decoder retraining, combined with more limited supervised retraining after longer periods of inactivity, may be sufficient to achieve high performance. Nevertheless, future work must confirm this online, as offline simulations are not always predictive of online performance.

While we do appreciate that decoders can perform worse online than they do offline, we do not think this is always the case (nor necessarily to be expected). Since in this work the user only receives delayed feedback of the decoded characters after they have been completed/detected by the RNN, we think offline simulations are more likely to transfer to the online case as compared to continuous motion BCIs which rely heavily on moment-to-moment visual feedback corrections.

To clarify, note that the results in Table 1 do not show a failure of decoding results to transfer to the online domain. Although the last row of the table reports best performance with an offline decoder, this offline decoder was *acausal* (a bidirectional RNN) and was not tested online or shown to be worse online. Its high performance was likely due to its acausal nature. Note that this acausal decoder was tested to provide a point of comparison to prior BCI work which has also used acausal methods.

Finally, we would like to clarify that the 10 sentences selected for calibration in our offline simulation were subsampled at even intervals from the 50 possible sentences (thus ensuring that the 10 sentences are distributed evenly in time). To understand the effect of this choice on performance, we re-ran the analysis 10 more times, each time with sentences chosen at random (i.e., uniformly at random instead of deterministically at even intervals). Results show a tight clustering near the originally reported result, suggesting that the choice of sentences does not have a strong effect on decoder performance. The originally reported error rate was 8.5%; the mean of these new random runs was 9.2% with a standard deviation of 0.6%. In the Methods, we now clarify that, in the offline simulations shown in Figure 3, sentences were subsampled from the original set of sentences at even intervals and that this choice does not affect the conclusions:

> When reducing the amount of calibration data, we subsampled from the original 50 sentences at even intervals (thus ensuring that the subsampled data contained sentences spaced evenly in

time). Note that results are similar when choosing sentences uniformly at random. To test this, we re-ran the analysis 10 more times using 10 sentences chosen randomly instead of evenly. The reported error rate in Fig. 3a was 8.5% for 10 sentences; the mean of these 10 random runs was 9.2% with a standard deviation of 0.6%.

Another related issue is the explanations given to the difference between online and offline decoding performance. For example, it is well established that well isolated unit spiking does provide more information compared to local field potentials for BCI decoding but comes at the cost of increased computational complexity and variability over time, both within session and across sessions. I did not find their argument about not using spike sorting to be particularly compelling. Even though it is a subjective process as they state – but so is the thresholding process they've used, it has the potential to increase their information rate and consequently typing speed which is their main outcome measure. Unless it was observed that this process does somehow affect the spatiotemporal templates they use in the classification, the reasons for not using putative single or multi-unit clusters of waveforms in building the firing rate templates are not entirely clear.

Thank you for these considerations. Ultimately, since we did not compare multiunit threshold crossings to spike-sorted clusters in this work, it is unknown whether spike-sorting could have improved our system's performance. It does seem plausible that at least some small performance benefit could have been gained by using spike-sorting. The only place in the manuscript where this issue is addressed is a paragraph in the Methods that motivates our choice of multiunit threshold crossings. We have revised this paragraph as follows:

We used multiunit threshold crossing rates as neural features for analysis and neural decoding (as opposed to spike-sorted single units). We made this choice to simplify the methods, not ~~This was not~~ because spike waveforms could not be recorded (see ~~SFig 6~~Extended Data Fig. 7 for examples~~);~~). ~~rather, using multiunit threshold crossings allowed us to leverage information from more electrodes, since many electrodes recorded activity from multiple neurons that could not be precisely spike-sorted into single units.~~ Recent results ~~indicate~~ suggest that neural population structure can be accurately estimated from threshold crossing rates alone [45](Trautmann et al., 2019), and that neural decoding performance is ~~similar~~ comparable (within 5%) to using sorted units (Chestek et al., 2011; Christie et al., 2014) – although see also (Todorova et al., 2014)[46].

It is also important to explain why 7 days or less seem to maintain uncalibrated decoder accuracy. It is important to cite prior published work in which it was demonstrated that the same duration tends to also be associated with stability of single unit spiking1,2. Is this a coincidence? I think the same is happening here, that decoders need to be calibrated because unit spiking and character representation seems to shift over intervals > 7 days.

1. Dickey, Adam S., et al. "Single-unit stability using chronically implanted multielectrode arrays." Journal of neurophysiology 102.2 (2009): 1331-1339.
2. Eleryan, Ahmed, et al. "Tracking single units in chronic, large scale, neural recordings for brain machine interface applications." Frontiers in neuroengineering 7 (2014): 23.

Thank you for this suggestion. Indeed, a relatively high stability within a 7-day window is consistent with this prior work, which we now cite in the main text:

We found that when only 2-7 days passed between sessions, performance was reasonable with *no* decoder retraining (11.1% raw error rate, 1.5% with a language model), as might be expected from prior work indicating short-term stability of neural recordings[19–21].

19. Dickey, A. S., Suminski, A., Amit, Y. & Hatsopoulos, N. G. Single-Unit Stability Using Chronically Implanted Multielectrode Arrays. *J Neurophysiol* **102**, 1331–1339 (2009).
20. Eleryan, A. *et al.* Tracking single units in chronic, large scale, neural recordings for brain machine interface applications. *Front. Neuroeng.* **7**, (2014).
21. Downey, J. E., Schwed, N., Chase, S. M., Schwartz, A. B. & Collinger, J. L. Intracortical recording stability in human brain–computer interface users. *J. Neural Eng.* **15**, 046016 (2018).

Specific comments:
Authors state that "The character templates were made by drawing each character with a computer mouse in the same way as T5 described writing the character. As each character was drawn, the X and Y velocity trajectories of the mouse pointer were recorded.;" I can understand how T5 can describe how the final shape would look like, but it's unclear how can T5 describe the velocity by which he attempted to write different parts of the characters from which the target velocity vector for the decoder was defined.

To clarify, the target velocity vectors were only meant to be a rough approximation of T5's intended movement, based on the assumption that another person drawing the same character shape with a computer mouse would naturally follow a similar velocity trajectory. T5 never described the velocity of each character, and no attempt was made to check it against T5's understanding.

The algorithm we used to decode the pen tip velocity is invariant to linear scaling in the overall writing speed (since it stretches/shrinks each target template to best match the neural activity), but is not invariant to more subtle differences in the velocity of each individual stroke. Nevertheless, the fact that recognizable character trajectories could be decoded with this method shows that, even though the computer mouse trajectories are (necessarily) only rough approximations of T5's intended velocities, they are close enough to enable successful decoder training.

We now further clarify this point in the Methods section:

To train the decoder, we used hand-made templates that describe each character's pen trajectory. The character templates were made by drawing each character with a computer mouse in the same way as T5 described writing the character. As each character was drawn, the X and Y velocity trajectories of the mouse pointer were recorded. These templates then defined the target velocity vector for the decoder on each time step of each trial, much like prior work has trained decoders to predict the user's "intended" velocity for continuous movement tasks [2,49]. These templates were only intended to be a rough approximation of T5's intended pen tip velocities, based on the assumption that another person drawing the same character shape with a computer mouse would naturally follow a similar velocity trajectory (up to some time-scaling factor, to account for differences in overall writing speed).

Line 115: How did neural activity look like when an error was made? and when the subject was provided visual feedback about the language model's autocorrection of that error? Did the subject stop modulating, eventually relying on the model to autocorrect, or did he continue to modulate neural activity to correct the typo? Was the decoder 'disengaged' in those instances? did the neural activity occupy different regions of the state space relative to the intended character or the corrected character?

Thank you for these interesting suggestions. We would like to clarify that the language model was applied *offline* only, in a retrospective analysis to simulate an autocorrect feature. Thus, the participant never saw the autocorrections. To make this clearer, we added the word "Offline" to the system diagram in Figure 2 that depicts the language model (previously it was described only as "Retrospective" in the figure diagram).

Authors report that T5 spent 93% of the time looking at the prompt in the copy-typing task. They should also state what the eye tracking statistics were in the free typing trials in which there was no prompt. They also did not respond to the question if the perception of errors had any influence on the decoded patterns, particularly given that they cite their own work showing how errors result in distinct signature in motor cortex. They should clarify if and how these signatures, particularly when an error was made in the free typing trials, was handled by the decoder.

Thank you again for these suggestions. As the space limitations of a Nature article are strict, we feel we must retain the manuscript's focus on the central results – (1) demonstrating that handwriting movements are neurally encoded even years after paralysis, (2) that complete handwritten sentences can be neurally decoded at high speeds and accuracies using a novel decoding approach, and (3) that theoretical considerations suggest that temporally complex movements are easier to decode than point-to-point movements, making high-performance handwriting decoding possible.

While the effect of errors on neural activity (and, relatedly, the pattern of gaze in BCI use) is an important and interesting topic, which we too are deeply curious about, the results of such analyses would not directly bear on these central claims. Even if they did, there would unfortunately be no space to include or discuss them in the main text without removing other central results.

To clarify, our decoder was not designed to detect neural signatures of error – it was trained only to maximize classification accuracy. Thus, errors were not handled in a special way by the decoder. If it made an error, it simply continued unaware as if it had not.

Finally, we would like to also clarify that our prior studies on the encoding of errors are not cited in this manuscript – we cited them only in our response to reviewers in the previous round.

Authors have responded to my comment about the toy example by stating that the example included noise. My reference was to Figure 3F in which noise is absent in the trajectories shown. While I agree with the authors that "as the amount of noise increases, complex trajectories (may) become easier to classify because their nearest neighbor distances are larger (and thus nearby trajectories are less likely to be ", this is the case only under two assumptions:

1) The noise is white (as they have simulated already) and 2) the signal (i.e. the trajectory being classified) is uncorrelated with that noise. In the brain, however, there is ample evidence to suggest that the noise is not white, and is strongly correlated with the signal, and if not, it would at least be correlated among adjacent electrodes.

As they show in their inequality, when the distance between noisy fa and fb is less than the noise norm, misclassification will happen. However, if fa is more different than fb but the noise is more similar (i.e. correlated) to fb, adding that noise to fa will bring the noisy fa closer to fb, thus increasing the probability of misclassification. This is the case when the noise cluster extends along specific directions closer to those spanned by the signals. The examples simulated in the rebuttal use temporally colored noise, but the extent to which it is correlated with the actual signals being classified is unclear. I suggest that the authors bring more realism into their toy example regarding the noise characteristics (especially its temporal correlation with the signal while keeping its variance small) to make the explanation they are offering more compelling.

Thank you for this helpful suggestion. In addition to temporally correlated noise, we now also simulated signal-correlated noise (i.e., noise that spans the dimensions which connect the class means). The results continue to hold in the case of signal-spanning noise as well. This can be explained by the fact that signal-spanning noise acts like white noise in dimensions that span the class means, but is zero elsewhere. Since noise in dimensions that *don't* align with the class means are not as relevant for classification performance, it makes sense that their absence does not change the main result.

We now summarize these new analyses in a new Extended Data Figure (now Extended Data Fig. 5) and Supplementary Note, which we reproduce below at the end of this document for convenience. Additionally, we now call out this Supplementary Note and the assumption of uncorrelated white noise in the main text:

> Although neural noise in the toy model was assumed to be independent white noise, we also found that these results hold for noise that is correlated across time and neurons (Extended Data Fig. 5, Supplementary Note 1).

Thank you again, as we too feel that exploring these additional noise types helps to strengthen the argument.

Overall, I think the work is very valuable and would be an important contribution to the field, provided the authors address some of the remaining issues above.

Thank you again for all of these helpful comments and suggestions. We did our best to incorporate them, given the page limit constraints of a Nature article.

Supplementary Note 1 – Effect of Noise Correlations on the Toy Model of Classifiability

In the toy example presented in Fig. 4F-H, we showed that additional temporal dimensions can be used to improve the classifiability of a set of neural patterns in the presence of Gaussian *white noise* that is uncorrelated across time points and neurons. Under these assumptions, the Euclidean distance between each pair of neural patterns is the relevant factor determining classification accuracy, and it therefore follows that greater temporal dimensionality will improve classification performance if it helps to spread out those patterns more evenly. Here, we examine how *correlated noise* might affect this result.

First, it is helpful to define some terms. Let f_x be a vector that describes the underlying neural trajectory for movement x (i.e., the mean neural firing rates across time for movement x). Each entry in the vector f_x is the mean firing rate for a single time step. To describe multiple neurons, the activity profile of each neuron can be stacked one on top of the other in the vector. Let ε be a neural noise vector of the same length that has a multivariate normal distribution with zero mean and covariance matrix Σ. If Σ is non-diagonal, the noise is said to be correlated.

Given a vector of noisy observed firing rates $r = f_x + \varepsilon$, a maximum likelihood classifier will choose to classify r into the class that has the minimum Mahalanobis distance to r (assuming uniform class priors). In other words:

$$\underset{x}{\operatorname{argmin}}(r - f_x)^T \Sigma^{-1} (r - f_x)$$

In the case of white noise, Σ is a diagonal matrix with all diagonal entries equal to σ. In this case, the classifier will simply choose the class whose mean has the smallest Euclidean distance to r. This justifies the idea that nearest neighbor distances should be increased to reduce classifier confusions (potentially via spreading the neural patterns out into additional temporal dimensions).

If Σ is non-diagonal, this means that the noise cloud will extend more in some directions and less in others. The directions that are most harmful for classification are those that connect nearby class means (e.g., the direction f_x-f_y, as this would make noise more likely to 'corrupt' class x to look more like y). In the general case where Σ can take any arbitrary shape, it is not always true that classification accuracy can be improved by using extra temporal dimensions to increase Euclidean distances. For example, it could be the case that these extra temporal dimensions are particularly noisy, cancelling out the benefit of increased distance between the class means. Nevertheless, under reasonable constructions of Σ that we test below, we show that the toy model in Fig. 4 still holds in the presence of correlated noise.

Temporally Correlated Noise

First, we tested noise with *temporal* correlations (meaning that the noise associated with each neuron was positively correlated in time). This noise can describe slow (but random) fluctuations in neural firing rates over time, and in this sense is more realistic than white noise. Temporal correlations would generally cause the noise to be more concentrated along dimensions that span the class means, since the underlying neural patterns are also smooth across time (as is the case in this toy example). Extended Data Fig. 5a shows examples of temporally correlated noise vectors and the covariance matrix used to generate them. The wide diagonal band in the covariance matrix causes nearby time steps to have correlated noise.

In Extended Data Fig. 5b, we compared the classification accuracy between time-varying trajectories and constant trajectories in the presence of temporally correlated noise, finding an even more pronounced improvement for time-varying trajectories. This is because neural patterns that vary more quickly in time are less aligned with slow-varying noise directions, enabling greater robustness to this type of noise. Here, classification was performed with a maximum likelihood classifier (under the assumption that the means of each class and the covariance matrix of the noise are known). However, results also hold using

a simpler “Euclidean distance” classifier that assumes the noise is white by choosing the class whose mean has the smallest Euclidean distance to *r*.

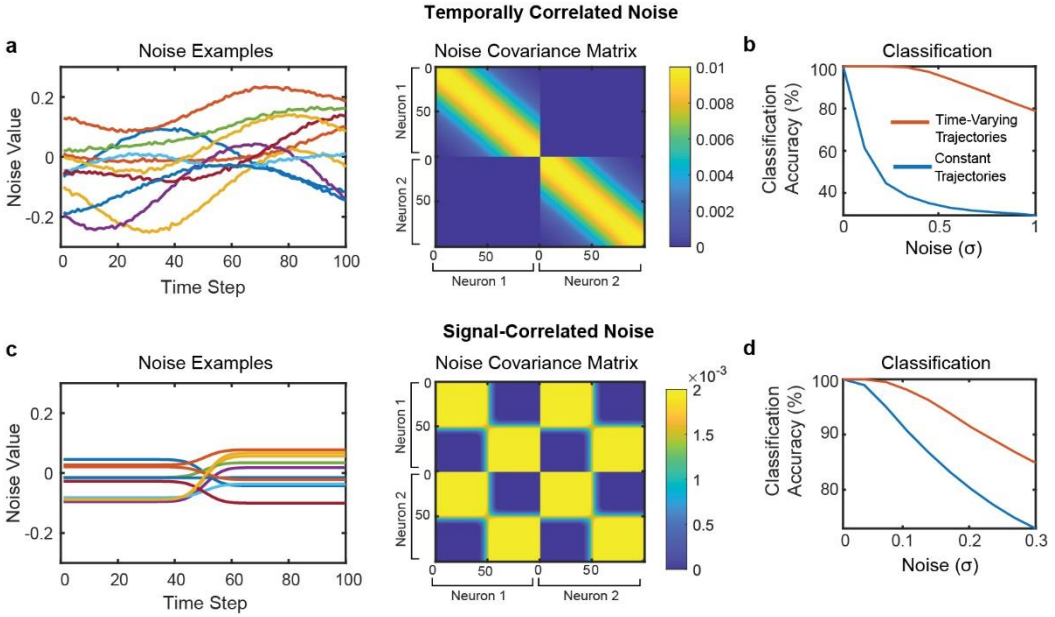
Signal-Correlated Noise

Finally, we tested noise vectors that were directly correlated with the underlying neural signal (that is, noise vectors that contained variance *only* in signal-spanning dimensions that connect the class means, such as f_x - f_y). This type of noise is more realistic than white noise in the sense that neural variability is often larger in neural dimensions that carry the signal. To find these dimensions, PCA was applied separately to the constant and time-varying trajectories to find the one (constant) or two (time-varying) spatiotemporal axes containing the neural signal. The covariance matrix was then designed to place noise in these axes only (with equal variance for each axis):

$$\Sigma = \sigma \mathrm{A}\mathrm{A}^T$$

Here, A is a matrix whose columns are the PCA axes and σ scales the overall size of the noise. Extended Data Fig. 5c shows what these noise vectors look like for the time-varying trajectories. Because the time-varying trajectories have only two temporal dimensions, the noise vectors also have this structure (where the first 50 time points are highly correlated with each other and the last 50 time points are highly correlated with each other).

Again, even in the presence of noise that is correlated with the signal, we found that it is still easier to classify time-varying trajectories than constant trajectories (Extended Data Fig. 5d). This result can be explained by the fact that signal-spanning noise acts like white noise in dimensions that span the class means, but is zero elsewhere. Since noise in dimensions that *don’t* align with the class means are not as relevant for classification performance, it makes sense that their absence does not change the main result.



Extended Data Fig. 5: Effect of correlated noise on the toy model of temporal dimensionality. a, Example noise vectors and covariance matrix for temporally correlated noise. On the left, example noise vectors are plotted (each line depicts a single example). Noise vectors are shown for all 100 time steps of neuron 1. On the right, the covariance matrix used to generate temporally correlated noise is plotted (dimensions = 200 x 200). The first 100 time steps describe neuron 1’s noise and the last 100 time steps describe neuron 2’s noise. The diagonal band creates noise that is temporally correlated within each

simulated neuron (but the two neurons are uncorrelated with each other). **b,** Classification accuracy when using a maximum likelihood classifier to classify between all four possible trajectories in the presence of temporally correlated noise. Even in the presence of temporally correlated noise, the time-varying trajectories are still much easier to classify. **c,** Example noise vectors and noise covariance matrix for noise that is correlated with the signal (i.e., noise that is concentrated only in spatiotemporal dimensions that span the class means). Unlike the temporally correlated noise, this covariance matrix generates *spatiotemporal* noise that has correlations between time steps *and* neurons. **d,** Classification accuracy in the presence of signal-correlated noise. Again, time-varying trajectories are easier to classify than constant trajectories.