A neural network for online spike classification that improves decoding accuracy

Deepa Issar[1,2], Ryan C. Williamson[2,3,4], Sanjeev B. Khanna[1], and Matthew A. Smith[1,3,5]

[1]Department of Bioengineering, University of Pittsburgh, Pittsburgh, PA, USA

[2]School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA

[3]Center for the Neural Basis of Cognition, Pittsburgh, PA, USA

[4]Department of Machine Learning, Carnegie Mellon University, Pittsburgh, PA, USA

[5]Department of Ophthalmology, Eye and Ear Institute, University of Pittsburgh, Pittsburgh, PA, USA

Abstract

Objective. Separating neural signals from noise can help to improve brain-computer interface performance and stability. However, "spike-sorting" suffers from a lack of a ground truth for classifying waveforms, and most algorithms for separating neural action potentials from noise are not suitable for use in real time and have shown mixed effects on decoding performance. Our goals were to automate the intuition of human spike-sorters to operate in real time with an easily tunable parameter governing the stringency with which spike waveforms are classified. *Approach*. We trained an artificial neural network with one hidden layer on neural waveforms that were hand-labeled as either spikes or noise. The network output was a likelihood metric for each waveform it classified, and we tuned the network's stringency by varying the minimum likelihood value for a waveform to be considered a spike. Using the network's labels to exclude noise waveforms, we decoded remembered target location from electrode arrays implanted in prefrontal cortex (PFC) during a memory-guided saccade task in two rhesus macaque monkeys. We assessed our network's performance by measuring how it influenced decoding accuracy. *Main results*. The network classified waveforms in real time, and its classifications were qualitatively similar to those of a human spike-sorter. Compared to decoding with all waveforms (i.e. threshold crossings), we found that we could improve decoding performance, or at worst cause only a small change, if we removed waveforms with low spike likelihood values. Furthermore, decoding with our network's classifications became more beneficial as time since the array was implanted increased. *Significance*. Our

classifier serves as a feasible preprocessing step that could be applied to both offline neural data analyses and online decoding. Additionally, we demonstrated that using our network's classifications for noise removal could improve decoding performance with little risk of harm.

Introduction

Brain computer interfaces (BCIs) have been used as both research tools to understand neural phenomena (1-4) and devices to improve patient control of prosthetics (5-7). BCIs interpret neural signals arising from electrodes implanted in the cortex using real-time decoding algorithms; however, their performance is limited by the difficulty of segregating individual neurons from extracellular voltage signals ("spike-sorting"), the characteristics of the information present in individual neurons and multi-unit activity (their "tuning"), as well as the nuances of interactions among neurons (8). Refining the raw data is necessary to improve BCI performance, but identifying waveforms that contain relevant information is challenging.

One common noise removal method is to use all waveforms crossing a minimum voltage threshold for decoding, where all waveforms on a particular channel are considered to be from a single neural "unit" (9, 10). Using threshold crossings also serves to reduce the amount of stored data. However, threshold crossings often capture noise and the summed activity of multiple nearby neurons, called multi-unit activity, on each channel (11, 12). Alternatively, since waveforms with the characteristic shape of an action potential are the focus of most offline neural data analyses, a different approach is to isolate well-defined spike waveforms from the threshold crossings via spike-sorting, and use only those waveforms for decoding (9, 12). Unfortunately, most spike-sorting techniques are unsuitable for BCI use because they are time-intensive and require manual refinement.

Furthermore, even with the development of real-time, automated spike-sorting algorithms (13, 14), there is no clear ground truth for what constitutes a spike waveform based solely on extracellular recordings (12, 15-17). While some studies have found spike-sorted units improved decoding performance (18-21), others have found that threshold crossings resulted in comparable or better decoding

that was more stable over time (10, 22-25). These conflicting results highlight the gap in our understanding of what information in neural recordings is most valuable for BCIs. Even within studies, there are inconsistent effects of spike-sorting on decoding, both between subjects and over time (10, 22). It is also difficult to compare the effects of spike-sorting on decoding performance across studies due to the inherent variability in spike-sorting techniques (12, 15-17) and differences in brain regions, parameters being studied, and decoding algorithms that may influence which waveforms are the most decodable (9, 22, 23, 26, 27).

Given the subjective nature of spike-sorting and its history of variable effects on decoding performance, we sought to avoid explicitly sorting our data and instead developed an easily tunable spike classifier that could objectively and efficiently assign waveforms to one of two classes: spike or noise. To achieve this goal, we sought a classification approach that would (1) take advantage of the spike-sorted waveforms from previous analyses that formed a readily available, labeled training data set; (2) output a likelihood of being a spike for each waveform so we could simply adjust the minimum cutoff for binary spike classification; (3) once trained could operate in real time. We built a neural network classifier that satisfied these three criteria. To evaluate our classifier's performance, we decoded the activity of electrode arrays implanted in prefrontal cortex (PFC), and compared decoding accuracy of task condition with all threshold crossings to decoding accuracy with the waveforms selected by our classifier. We also assessed how decoding performance changed as a function of the stringency of our classifier (i.e. the minimum cutoff for spike classification) and explored how decoding accuracy with our classifier changed as a function of time since the array implantation (array implant age).

By removing waveforms that the network identified as unlikely to be spikes, we improved decoding performance for most sessions relative to decoding with all threshold crossings. Moreover, in the remaining sessions there was no substantial detriment to decoding accuracy, meaning we could apply our method with little risk of harming decoding performance. Our network classifications continued to improve decoding accuracy in sessions recorded long after array implantation, even though the overall decoding accuracy decreased and the array recordings became noisier. Thus, our real-time, tunable

waveform classifier demonstrates promise for long-term BCI applications and our findings highlight the value of spike-sorted data as a training set to inform waveform classification.

Materials and Methods

We trained a neural network, using a database of spike-sorted waveforms, to assign waveforms to a spike or noise class. We then tested the network's classifications with another data set that was independent from the training set. To evaluate the network's classifications, we decoded task location using only waveforms labeled as signal by the network and compared the accuracy to decoding accuracy using all recorded waveforms. All experimental procedures were approved by the Institutional Animal Care and Use Committee of the University of Pittsburgh and were performed in accordance with the United States National Institutes of Health's *Guidelines for the Care and Use of Laboratory Animals*.

Neural recordings

We analyzed neural recordings from six adult male rhesus macaques (*Macaca mulatta*) which had previously been spike-sorted for ongoing experiments in the laboratory. Data from four subjects were used to train our spike-classifying neural network, and data from the remaining two subjects (Monkey Pe and Monkey Wa) were used to assess decoding accuracy using the network's classifications. Raw recordings from both 96-electrode 'Utah' arrays (Blackrock Microsystems, Salt Lake City, UT) and 16-channel linear microelectrode arrays (U-Probe, Plexon, Dallas, TX) were band-pass filtered from 0.3 to 7,500 Hz, digitized at 30 kHz, and amplified by a Grapevine system (Ripple, Salt Lake City, UT). For each recording session, a threshold was defined for each channel independently based on the root-mean-squared voltage of the waveforms (V_{RMS}) recorded on that channel at the beginning of the session (i.e. V_{RMS}*X, where X was a multiplier set by the experimenter). Each time the signal crossed that threshold, a 52-sample waveform segment was captured (i.e. a threshold crossing), with 15 samples prior to and 36 samples following the sample in which the threshold excursion occurred.

Offline spike-sorting

Waveform segments were classified as spike or noise waveforms using a custom, offline MATLAB spike-sorting algorithm that used an automated competitive mixture decomposition method (28). These automated classifications for each recording session were subsequently refined manually by a researcher. While only a single researcher spike-sorted the waveforms from any particular session, the data in the training set (7 sessions total) collectively consisted of data spike-sorted by four different researchers. If multiple unique spike waveform shapes were present on a channel, the sorter would label those as different units.

Training data set

We sought a diverse neural waveform training set that captured a variety of subjects, brain regions, implanted array ages (i.e. number of days since the array was implanted) and recording devices (Table 1). The four training data set subjects each had a 96-electrode 'Utah' array implanted in right or left hemisphere visual area V4. Two of those subjects also had 16-channel U-probe recordings from the frontal eye field (FEF), located in the anterior bank of the arcuate sulcus. Data from two sessions were used for three of the subjects: one session recorded shortly after the array implant, and one session recorded at least four months later. For the fourth subject, only a single recording session recorded shortly after the array implant was used because there were no later recordings from that array. These recordings were collected and sorted as part of previous studies which describe the experimental preparations in detail (29-32). Data in the training set were assigned a label of 0 or 1. For each session, all waveforms labeled as noise during spike-sorting were included in the training set (labeled as 0). For the spike waveforms in the training set, only waveforms that were both labeled as spikes during spike-sorting and from a channel with very distinct spike waveforms, defined as a channel with a signal-to-noise ratio greater than 2.5 (33), were included in the training set (labeled as 1). The aim of these selections was to emphasize relatively well-isolated single unit action-potential shapes while also exposing the network to a variety of noise waveforms. If there were multiple single-units identified by the sorter on a particular channel, they were not distinguished in the training set and all were treated as spike waveforms (labeled as 1).

Overall, the network training set consisted of 24,810,795 waveforms from four monkeys, two brain regions (V4 and FEF), two recording devices (Utah array and U-probe), and different array implant ages. These waveforms were classified as spikes or noise using the offline spike-sorting technique described above (Fig. 1a).

Table 1. Training data set information, including time since implant (array age), brain area (RH-right hemisphere, LH-left hemisphere), and number of recording sessions. Utah arrays contained 96 channels and U-Probes had 16 channels.

	Recording device	*Brain area*	*Time since implant*	*# of sessions*
Monkey Wi	Utah array	RH V4	~1 month	1
	Utah array	RH V4	~11 months	1
	U-Probe	RH FEF	-	1
Monkey Ro	Utah array	LH V4	<1 month	1
	Utah array	LH V4	~8 months	1
	U-Probe	LH FEF	-	1
Monkey Bo	Utah array	RH V4	<1 month	1
	Utah array	RH V4	~4 months	1
Monkey Bu	Utah array	RH V4	~1 month	1

Testing data set

We used recordings from two additional subjects, not included in the training set, for testing to ensure that the trained network was generalizable to subjects not included in its training. The two testing data set subjects (Monkey Pe and Monkey Wa) each had two 96-electrode 'Utah' arrays implanted, one in visual area V4 and one in dorsolateral PFC (on the prearcuate gyrus just medial to the principal sulcus, area 8Ar). Both arrays were implanted in the right hemisphere for Monkey Pe and in the left hemisphere for Monkey Wa. Only data recorded from PFC were used for analyses (Table 2) because this provided a matched data set from two monkeys for the purpose of assessing the impact of our method on decoding and because it allowed us to test our network's generalizability in situations where the data were recorded from a different region than the training set. However, we repeated some of the analyses in held out data from V4 and included these results in the supplement (Supplementary Figure 4). Recordings within 50 days of the array implant were considered early sessions and recordings from over 50 days after the implant were considered late sessions (see Table 2 for session counts by subject). For Monkey Pe, the thresholds across channels were similar for all recording sessions (90% of channels had thresholds

between -39 and -20 μV, median: -29 μV). For Monkey Wa, the thresholds were more permissive for the later recording sessions (> 6 months after the array implant) as the experiments sought to extract the maximum remaining signal in arrays that were decreasing in recording quality. The median of Monkey Wa's early sessions was -32 μV (90% between -43 and -23 μV), and the median of the later sessions was -15 μV (90% between -29 and -12 μV). The waveforms in the testing data set were also spike-sorted offline via the technique described above for a comparative analysis between our network classifications and offline spike-sorting.

Table 2. Testing data set information, including time since implant (array age), brain area (RH-right hemisphere, LH-left hemisphere), and number of recording sessions.

	Recording device	*Brain area*	*Time since implant*	*# of sessions*
Monkey Pe	Utah array	RH PFC	11-43 days	14
		RH PFC	80-147 days	22
Monkey Wa	Utah array	LH PFC	27-40 days	5
		LH PFC	180-224 days	11

Neural Network

Using TensorFlow (34) and Keras (35), we developed a neural network to classify data segments as spike or noise waveforms (Fig. 1b). The network accepts a waveform segment with '*s*' samples (*s*=52), which passes through a hidden layer with '*n*' units (*n*=50) that uses a rectified linear unit (ReLU) activation function. The product of the hidden layer passes through one output unit and then the network applies a sigmoid function to its output.

With the aforementioned training waveforms, the network was trained to maximize accuracy based on binary labels (0 for noise, 1 for spikes) using an Adam optimization algorithm (36) and a binary cross-entropy loss function in batch sizes of 100.

For a single waveform input run through the trained network, the output was a value between 0 (likely not a spike) and 1 (likely a spike), which represented the network's assessment of the likelihood that the input waveform was a spike. While this output value was not a conventional probability, we refer to it as the probability of being a spike or P(spike) because it was a likelihood metric scaled between 0 and 1. Supplementary Figure 1 provides more intuition regarding the hidden layer units and how the

network assessed waveforms. In order to classify the waveforms using the network's output probability, we set a minimum P(spike) for a waveform to be classified as a spike. We referred to this minimum probability as the γ threshold. A γ threshold of 0 classified all of the waveforms that were captured for the session as spikes; this is often referred to as 'threshold crossings' in the literature. Increasing γ resulted in fewer waveforms classified as spikes because the P(spike) cutoff was higher.

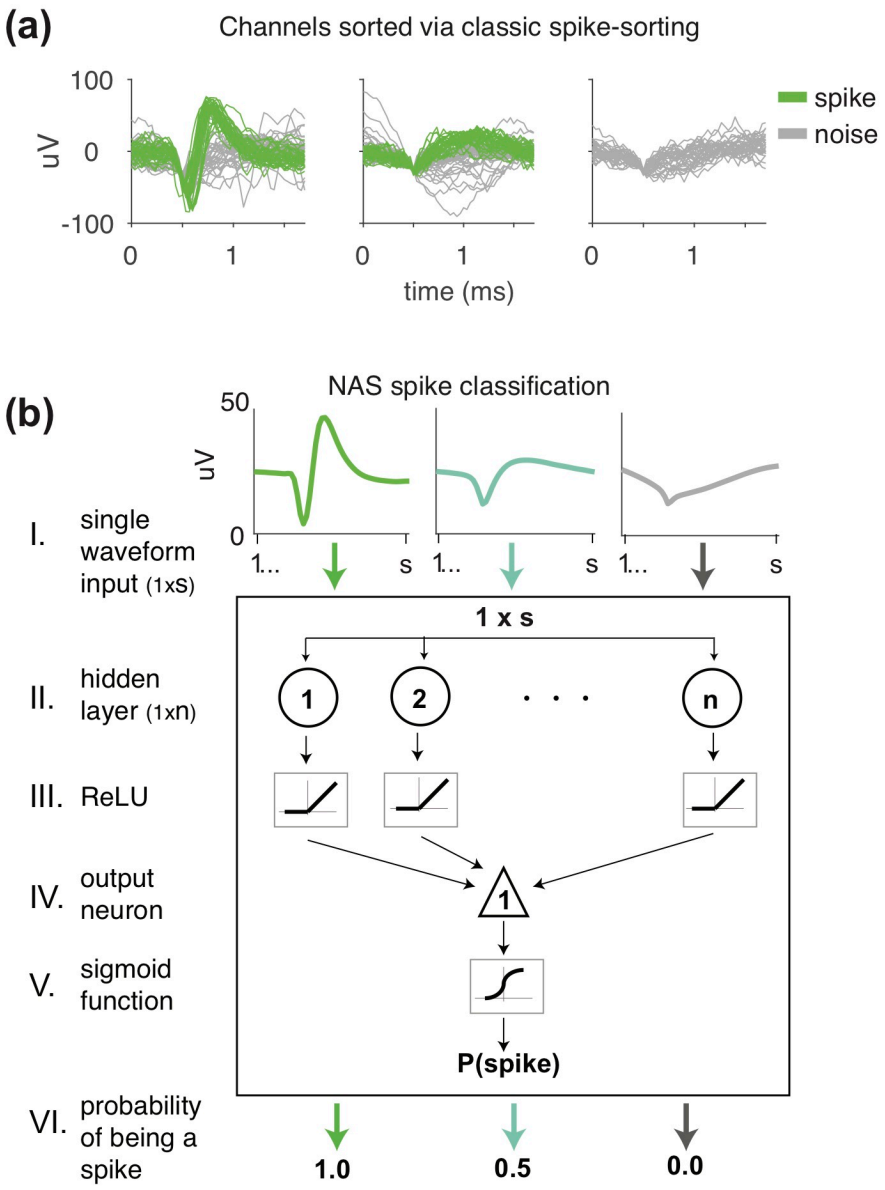


Figure 1. Methods of classifying waveforms: classic (manual) spike-sorting and our neural network spike classifier, Not A Sorter (NAS). (a) Spike-sorted channels with waveforms that were manually classified as spikes (green) and noise (grey). The network was trained on 24,810,795 spike-sorted neural waveforms from four monkeys. The waveforms were recorded on arrays implanted in V4 as well as U-probes placed in FEF. (b) Neural network structure and output for three sample waveforms. Each s=52 waveform input (I) was

passed through a hidden layer (II) with n=50 units. The resultant linear weighting of the waveform voltages was passed through a rectified linear unit (ReLU) non-linearity (III). The output was again passed through a weighted sum (IV) followed by a sigmoid non-linearity (V). The resulting value was the network's assessment of the likelihood that the input waveform was a spike waveform (VI). We referred to this value as P(s) or the "probability" of being a spike.

Custom Python and MATLAB scripts used to train the neural network and classify waveforms, as well as sample training and testing data, will be available on GitHub at the time of publication. We refer to our neural network as Not A Sorter (NAS), because it classifies waveforms as spikes or noise but does not attempt to sort the spikes into separable single units.

Memory-guided saccade task

The two testing data set subjects (Monkey Pe and Monkey Wa) performed a memory guided saccade task (Fig. 2a). Each subject fixated on a central point for 200 ms. Then, a target stimulus flashed at a set amplitude and direction for 50 ms, followed by a 500 ms delay. There were a total of 40 possible conditions: the radius could be one of five eccentricities and the target direction could be one of eight (spaced in 45° steps). After the delay, the fixation point disappeared instructing the subject to make a saccade to the location where the target flashed. For some sessions, during the saccade the target reappeared to help the subject; however, for all analyses we only used data from the beginning of the delay period (i.e. prior to the saccade). Subjects were rewarded with water or juice for making a saccade to the correctly remembered target location. For all decoding analyses, the decoded condition was the target location. Trials at a single eccentricity (Monkey Pe: 9.9°, Monkey Wa: ~7.6°) were used to maximize the number of recording sessions that could be compared. Thus, there were eight unique target locations to decode.

Only rewarded trials were analyzed. The number of trial repeats per condition varied over sessions. Monkey Pe had an average of 43 repeats per session (range: 25-75, sd: 17) and Monkey Wa had an average of 62 repeats per session (range: 51-83, sd: 11) for each target condition.

Decoding

To classify target direction prior to movement from PFC waveforms, we used a Poisson Naïve Bayesian decoder (Fig. 2b). All decoding analyses were performed offline. Waveforms classified as

spikes from 300 ms to 500 ms after fixation (i.e. 50 ms after stimulus offset) were used for decoding, while waveforms classified as noise were discarded. Each channel was considered a single decodable unit with one exception: to calculate decoding accuracy using manual spike-sorting, if there were multiple single-units identified by the manual spike-sorting method on a channel, each single-unit on that channel remained separate for the purposes of decoding. Units with an average firing rate across trials below 0.25 spikes per second were discarded. We allocated training and test trials using two different methods. For one analysis, we assigned 80% of the trials from each session to the training set and the remaining 20% to the test set (Fig. 4a,b). For the remaining analyses, we wanted two test sets, so we used 60% of the trials for training and 20% for each of the two test sets. The two test sets allowed us to search for an optimal γ threshold using the first test set and cross validate this value with the second test set. Training and test data were rotated such that each trial was used for a test set only once.

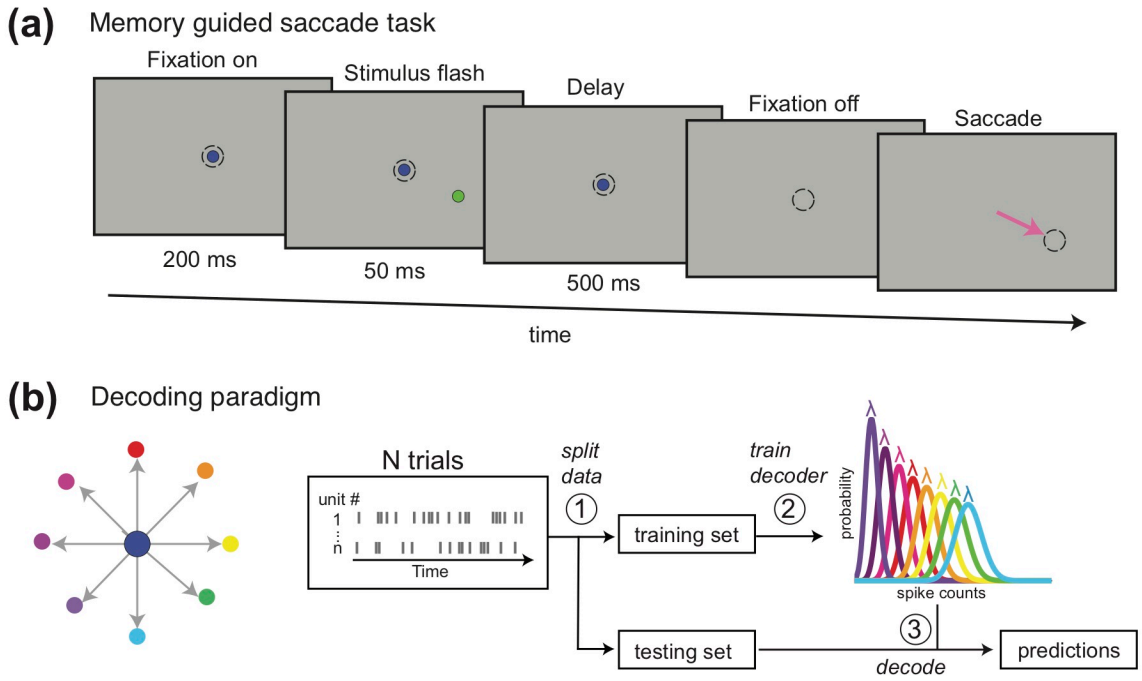


Figure 2. Offline decoding of planned direction from a memory guided saccade task. (a) Memory guided saccade task. The monkey fixated on a central point. A brief stimulus flash occurred followed by a delay period. When the fixation point turned off, the subject was required to make a saccade to the location of the previously flashed stimulus. (b) Decoding paradigm. The stimulus flash could occur at 8 different angles around the fixation point. A Poisson Naïve Bayes classifier was used to decode stimulus direction offline using spikes recorded from a 96-electrode Utah array in PFC 50 ms after target offset. For each recording session, we used 5-fold cross validation, where for each fold the data were split (step 1) into a training set to create model distributions for each direction (step 2) and an independent testing set to test the accuracy of the

model's predictions (step 3). Note that the curves in step (2) do not depict actual distributions and simply represent how a Poisson decoder could use spike counts from trials in the training set to distinguish between different target conditions.

Our decoding algorithm created a Poisson distribution model for each target location (θ) using the average spike count for each unit (n_{spike}) in the training set. For each test trial, the target location with the maximum prediction probability, $P(\theta|n_{spike})$, was the predicted label. In equation (1), $P(n_{spike}|\theta)$ was calculated using the Poisson model developed with the training trials:

$$P(\theta|n_{spike}) = \frac{P(n_{spike}|\theta) * P(\theta)}{P(n_{spike})}$$

Decoding accuracy was calculated as the ratio of the number of correct prediction labels to the total number of predictions. We used 5-fold cross validation and computed the average decoding accuracy across folds.

Results

We trained a one-layer neural network called Not a Sorter (NAS) to evaluate the likelihood that a neural waveform was a spike. We used a diverse set of waveforms in the training set from different brain regions, subjects, and array implant ages (time since array implant) in order to expose the network to a variety of waveform types. Each waveform in the training set was assigned a binary label of noise (0) or spike (1) via offline spike-sorting with manual refinement by researchers. The network was also exposed to variability in spike classification due to manual sorting because different researchers sorted different sets of waveforms in the training data. The network learned to assess how spike-like a waveform was based on binary labels, but itself output a continuous value between 0 and 1 for each waveform allowing for tunable classification. We referred to this output as the probability of being a spike or P(spike). The network classified 1000 waveforms in less than one millisecond on average (computed on a 2011 iMac with a 2.8 GHz Intel Core processor). In a 10 ms bin, 1000 waveforms would be the expected output of 1000 channels each with a firing rate of 100 spikes/s. Thus, our network could be easily integrated to operate within real-time computing constraints.

Qualitative assessment of NAS classifications

To assign each NAS-classified waveform a binary spike or noise label, we set a parameter called the γ threshold, which was the minimum P(spike) a waveform needed to be considered a spike waveform. We found that even a low γ threshold, such as γ=0.20 (Fig 3a), assigned most waveforms to classes that a human spike-sorter would deem appropriate. Most spike-sorters and spike-sorting algorithms search for waveforms with a canonical action potential shape – an initial voltage decrease followed by a sharp increase, a narrow peak, and a return to baseline. Increasing the γ threshold to 0.70 (i.e. only waveforms assigned a P(spike) > 0.70 were considered spikes) mimicked the effect of a more selective spike-sorter (Fig 3b), where the percentage of spike waveforms on the channel decreased since more waveforms were placed in the noise class and the remaining spike waveforms had a clearer single-unit shape. Considering all of the waveforms from this same channel, the range of P(spike) values coincided well with our subjective impression of the match of individual waves with a canonical action potential shape (Fig 3c), and this was also true across all channels from this array (Fig 3d). Thus, tuning the γ threshold from low values (near 0) to high values (near 1) resulted in a shift from a more permissive to a more restrictive regime.

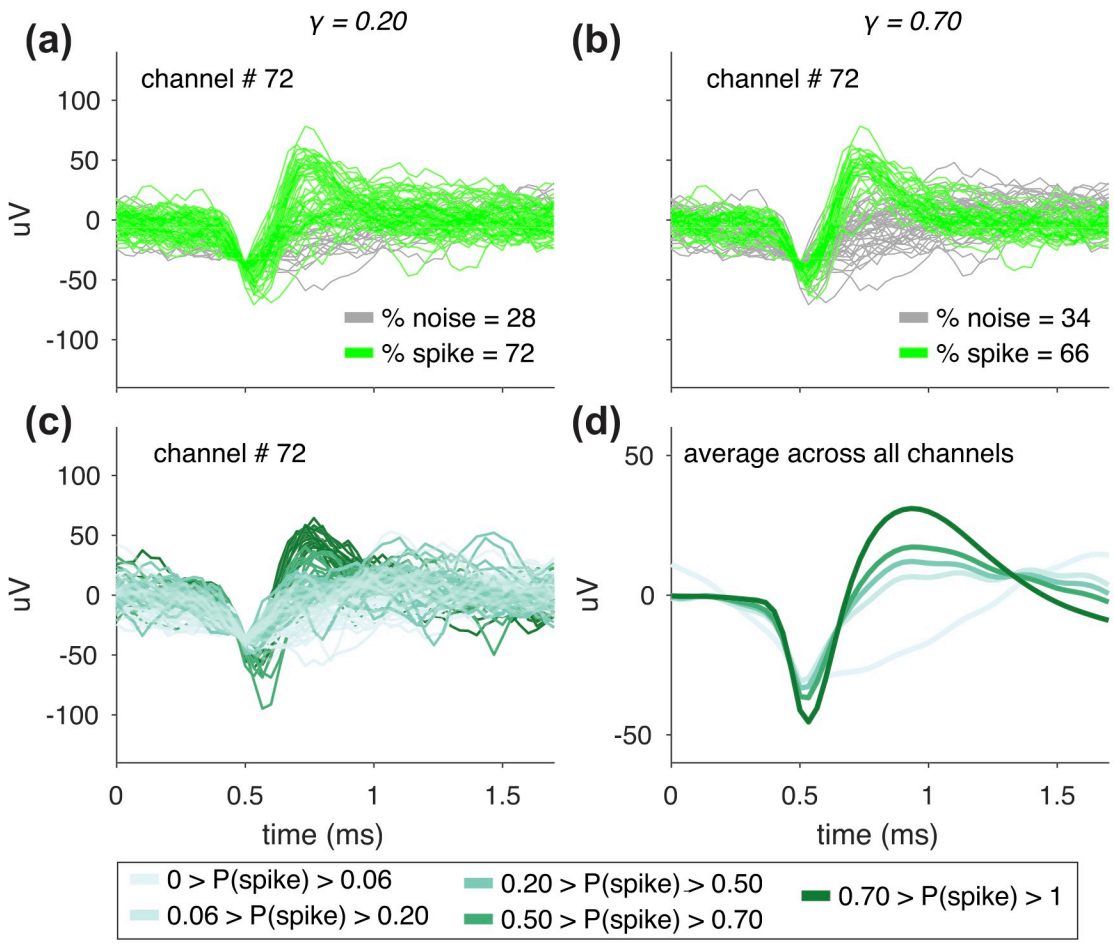


Figure 3. Classifying waveforms based on their probability of being a spike, P(spike), by setting γ, a tunable parameter. (a-b) Waveforms classified as spikes (green) or noise (grey) for a sample channel in Monkey Pe. The network outputs a value between 0 and 1, referred to as P(spike), for each waveform, where a value close to 1 means the network identifies that waveform as very likely to be a spike waveform. After running waveforms through the network, we set the minimum P(spike) to classify waveforms as spikes and referred to this value as the γ threshold. In (a) only waveforms assigned a spike probability greater than 0.20 (i.e. γ threshold = 0.20) and in (b) greater than 0.70 (i.e. γ threshold = 0.70) were classified as spikes. Increasing the γ threshold by definition results in a smaller percentage of waveforms captured in the spike class. (c) Waveforms from the same channel in (a) and (b) colored based on their network assigned P(spike) value. For a waveform where γ_1 < P(spike) < γ_2, the waveform would be classified as a spike when the threshold is γ_1, but would be labeled as noise for γ_2. (d) Similar to (c) except the average of waveforms across all channels within the indicated P(spike) ranges. Modifying the γ threshold tuned the stringency of the spike classifications.

Quantitative assessment of NAS classifications with decoding

Although by eye the network appeared to classify waveforms reasonably well, there was no available ground truth data with which to assess its performance. Instead, we sought to objectively assess its performance by measuring its effect on offline decoding accuracy. Specifically, we hypothesized that decoding using spikes classified by our network would be better than using threshold crossings. We

281 applied our network to a set of data recorded from PFC on which the network had not been trained. First,
282 we set the γ threshold and discarded any waveforms with a P(spike) below the threshold. Next, we split
283 the data into training and testing sets. We trained the decoder using the training set to decode the task
284 condition (the remembered location, out of 8 possibilities) during the delay period of a memory-guided
285 saccade task on each trial (see Methods). Then, we used the trained decoder to assess decoding accuracy
286 in the test set(s). We repeated this process for the same data set using multiple γ thresholds between 0
287 (i.e., all threshold crossings were considered spikes) and 0.95 (i.e., only waveforms assigned a 0.95 or
288 greater probability by the network were considered spikes). Since only waveforms classified as spikes
289 were used for decoding, as the γ threshold increased, fewer waveforms remained for decoding.

290 Initially, we used 80% of trials to train the decoder and the remaining to test it. We analyzed
291 cross-validated decoding accuracy as a function of γ (sample recording sessions in Fig. 4a,b). The
292 decoding accuracy using threshold crossings ($\gamma = 0$) for the sample recording sessions in Fig. 4a and 4b
293 was close to chance levels (1 out of 8, 12.5%) for Monkey Pe (12.8%) and was also low for Monkey Wa
294 (20.1%). Using the network's classifications improved decoding accuracy relative to using threshold
295 crossings. The peak decoding benefit of using the network's spike classifications occurred at a low γ
296 threshold (Monkey Pe: $\gamma = 0.16$, Monkey Wa: $\gamma = 0.16$). Being very selective about the spikes used for
297 decoding (by setting a high γ threshold) did not have a large impact on decoding accuracy, which
298 appeared to plateau after the peak. We found that increments of the γ threshold did not remove the same
299 number of waveforms. A substantial proportion of waveforms were removed at the lowest γ threshold
300 tested (especially in Monkey Pe), and small increments of the γ threshold as it approached a value of 1
301 could also result in large increases in the proportion of removed waveforms (especially notable in Figure
302 4b).

303 Ideally, for the sake of simplicity during online BCI experiments, we could select a single γ
304 threshold and use it for all recording sessions. We wanted to find the lowest γ threshold that improved
305 decoding accuracy for the majority of recording sessions. We focused on lower γ thresholds because it
306 was a conservative approach, and because our intuition from single channels and results in example

307 sessions (Figure 4) showed it could have the greatest benefit for decoding. In order to ensure the γ
308 threshold we selected was cross-validated, for the remaining analyses we created two test sets for
309 decoding (training set: 60% of trials, test set 1: 20% of trials, test set 2: 20% of trials). The first test set
310 was used to determine the optimal value of the γ threshold and the second to determine its impact on
311 decoding. For each session, using the first test set, we found the maximum cross-validated decoding
312 accuracy at any γ threshold, and then searched for the lowest γ threshold that resulted in a decoding
313 accuracy within 99% of that maximum (Fig. 4c, d).

314 Since the optimal γ threshold varied between sessions, we computed the median across sessions
315 of these γ thresholds chosen with our cross-validation approach and rounded it to the nearest tested value
316 (Monkey Pe: $\gamma = 0.2$, Monkey Wa: $\gamma = 0.08$). We then used that median value as the γ threshold for all
317 sessions in the second test set and analyzed the change in decoding accuracy from decoding using
318 threshold crossings in that same test set, which we termed Δ % decoding accuracy (Fig. 4e, f). Combining
319 sessions from both animals, the average improvement in decoding accuracy was 3.9% (2-tailed, Wilcoxon
320 signed rank test; across subjects: $p < 0.0001$, Monkey Pe: $p < 0.0001$, Monkey Wa: $p = 0.057$). Thus, our
321 network's classifications tuned to the previously described γ thresholds resulted in a net benefit for
322 decoding performance across sessions compared to using threshold crossings. Although we could tune the
323 γ threshold for each session to maximize the decoding, our choice of a fixed γ threshold was more
324 consistent with the use of our network in an online decoding context, where it would be desirable to set a
325 constant γ threshold at the start of each session rather than tuning it as a free parameter. However, an
326 alternative strategy would be to collect a small data set at the start of each day to find the optimal gamma
327 value, and then continue experiments for the remainder of that day using the chosen value.

328 Given the ability of our network to improve decoding accuracy beyond that observed with
329 threshold crossings, we took advantage of our longitudinal recordings in a fixed paradigm to understand
330 how our network's performance varied as a function of time. Since there is mixed evidence in the
331 literature regarding decoding stability with long-term array implants, we were specifically interested in

how the passage of time since array implant (array implant age), and concomitant degradation of recording quality, could influence our network's performance.

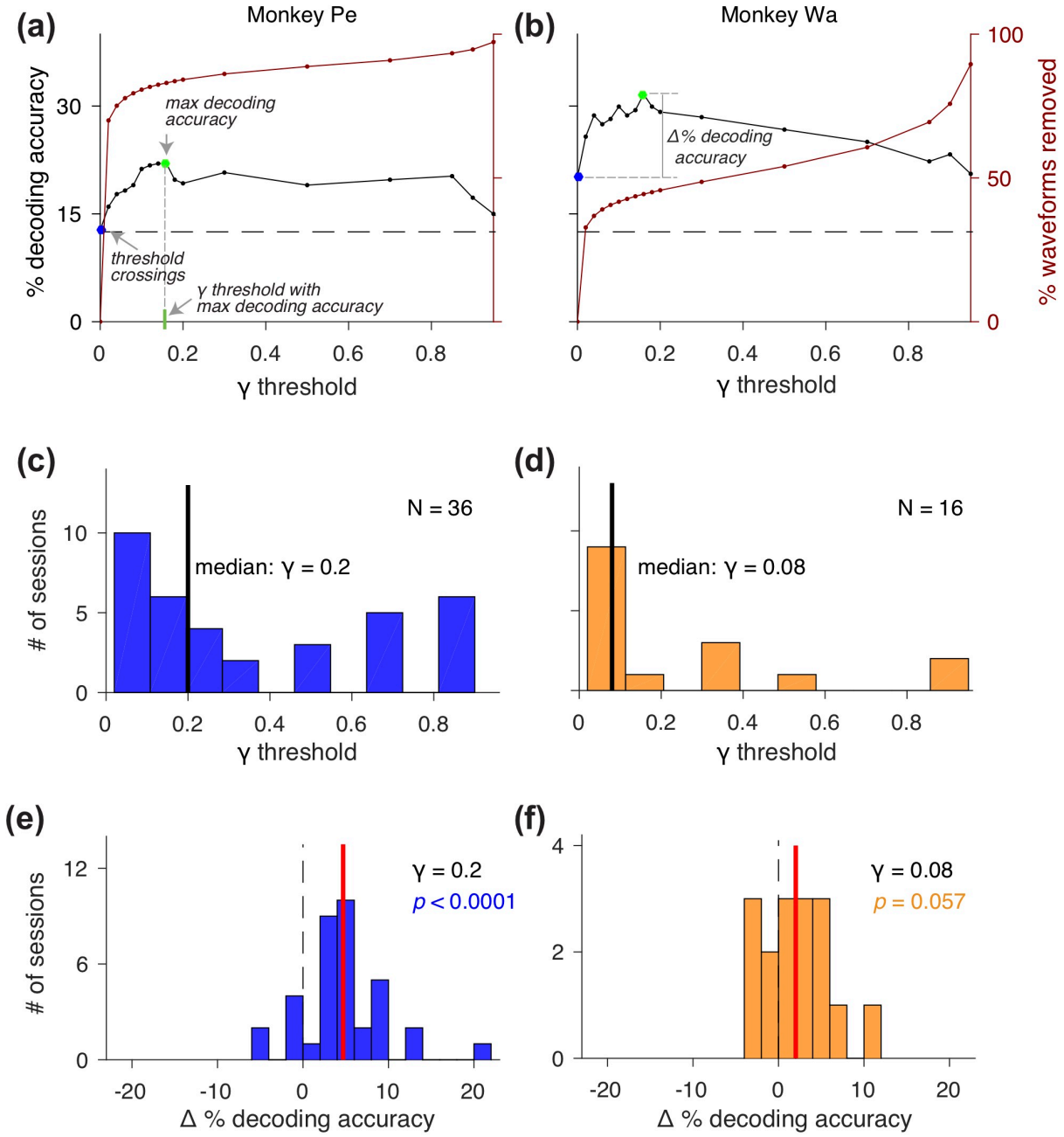


Figure 4. Using NAS spike classifications improved decoding accuracy in many sessions. (a-b) Percent decoding accuracy (black line) and percent of waveforms removed (maroon line) at different γ thresholds for an example recording session in Monkey Pe and Wa. The decoding accuracy increased for low γ thresholds and then reached an asymptote as γ increased. At the highest γ thresholds, decoding accuracy fell to chance (black, horizontal dotted line). Other labeled relevant metrics include: decoding accuracy with threshold crossings (blue dot) maximum decoding accuracy (green dot), γ threshold with the maximum decoding accuracy (grey dotted line), and Δ % decoding accuracy (the difference for a given session between the

decoding accuracy using the network classifications at a particular gamma threshold and the decoding accuracy with threshold crossings). (c-d) Distribution of γ thresholds across sessions (Monkey Pe: N =36, Monkey Wa: N = 16) that resulted in the maximum decoding accuracy for each session. (e-f) Distribution of Δ % decoding accuracy across all sessions where the γ threshold was the median from the distributions in (c-d). Mean Δ % decoding accuracy shown in red. For both monkeys, using spikes classified by the network improved decoding accuracy on average across sessions; however, this improvement was only statistically significant for Monkey Pe (2-tailed, Wilcoxon signed rank test, Monkey Pe: p < 0.0001, Monkey Wa: p = 0.057).

Impact of array age

First, we assessed how the overall quality of our neural data changed as a function of time. As the time since the array implant increased, the percentage of waveforms that were assigned a very low probability of being a spike (P(spike) < 0.02) also increased (Fig 5a; Spearman's rank order, Monkey Pe: $\rho = 0.87$, $p<0.0001$; Monkey Wa: $\rho = 0.90$, $p<0.0001$). Conversely, the percentage of waveforms assigned higher probabilities of being a spike (P(spike) > 0.70) decreased over time (Fig 5b; Monkey Pe: $\rho = -0.87$, $p<0.0001$; Monkey Wa: $\rho = -0.94$, $p<0.0001$). These classification changes were consistent with our qualitative observations that the arrays showed an increasing proportion of multi-unit activity (relative to single-unit activity) and apparent noise on the array over time. We further found that the ratio of the percentage of waveforms assigned a high P(spike) to the percentage of waveforms assigned a low P(spike) served as a proxy for a signal to noise ratio metric, and it was highly correlated with the median signal to noise ratio (across channels) of the array in each session (Supplementary Figure 2). Both signal to noise metrics decreased over time, and the decoding accuracy with threshold crossings also decreased the longer the array had been implanted (Fig 5c; Monkey Pe: $\rho = -0.86$, $p<0.0001$; Monkey Wa: $\rho = -0.69$, $p=0.004$).

Given the change in distribution of the types of waveforms present during the session over time, we tried to optimize the γ threshold based on time since array implant. We used the same method as in Fig. 4c-d to compute the γ thresholds with approximately maximum decoding (within 1%), except we calculated the median of the early sessions (recorded 0-50 days post implant) and the median of the late sessions (recorded >50 days post implant) separately. This did not prove to be a useful optimization as the median values were similar between early and late sessions in both Monkey Pe ($\gamma_{early} = 0.25$, $\gamma_{late} = 0.18$) and Monkey Wa ($\gamma_{early} = 0.06$, $\gamma_{late} = 0.08$). Decoding accuracy compared to threshold crossings (Δ %

decoding accuracy) was significantly increased in Monkey Pe (Figure 5d, 2-tailed, Wilcoxon signed rank test, $p < 0.0001$) and was not significantly helped or hurt in Monkey Wa ($p = 0.09$). In line with the trends in signal and noise over time, the decoding accuracies from later sessions (> 50 days post implant) were helped more by the network classifications in both subjects than those from earlier sessions (Figure 5d, Wilcoxon rank sum, Monkey Pe: $p = 0.01$; Monkey Wa: $p = 0.01$). When we normalized the decoding accuracy at each γ threshold by the decoding accuracy using threshold crossings for each session and separately averaged across early and late sessions, the increased benefit of using the network in the later sessions compared to the earlier sessions was clear (Figure 5e).

Altogether, these results provide additional evidence that factors such as time since the array was implanted and signal to noise ratio influence decoding accuracy and affect how noise removal impacts decoding performance. Despite decreasing signal quality and increasing noise waveforms, the effect of using our network classifications for noise removal prior to decoding was consistent in that it was most often beneficial for decoding and at worst minimally detrimental. We also confirmed that these results were not affected by poor γ threshold selection by using the maximum decoding accuracy regardless of γ threshold to calculate Δ % decoding accuracy (Supplementary Figure 3a), nor were the results substantially influenced by variability in the number of trials across sessions (Supplementary Figure 3b). Lastly, we assessed whether the benefits of our network also held in another brain region (V4) recorded from different Utah arrays but in the same sessions from the same animals in the test data set, and found similar results (Supplementary Figure 4).

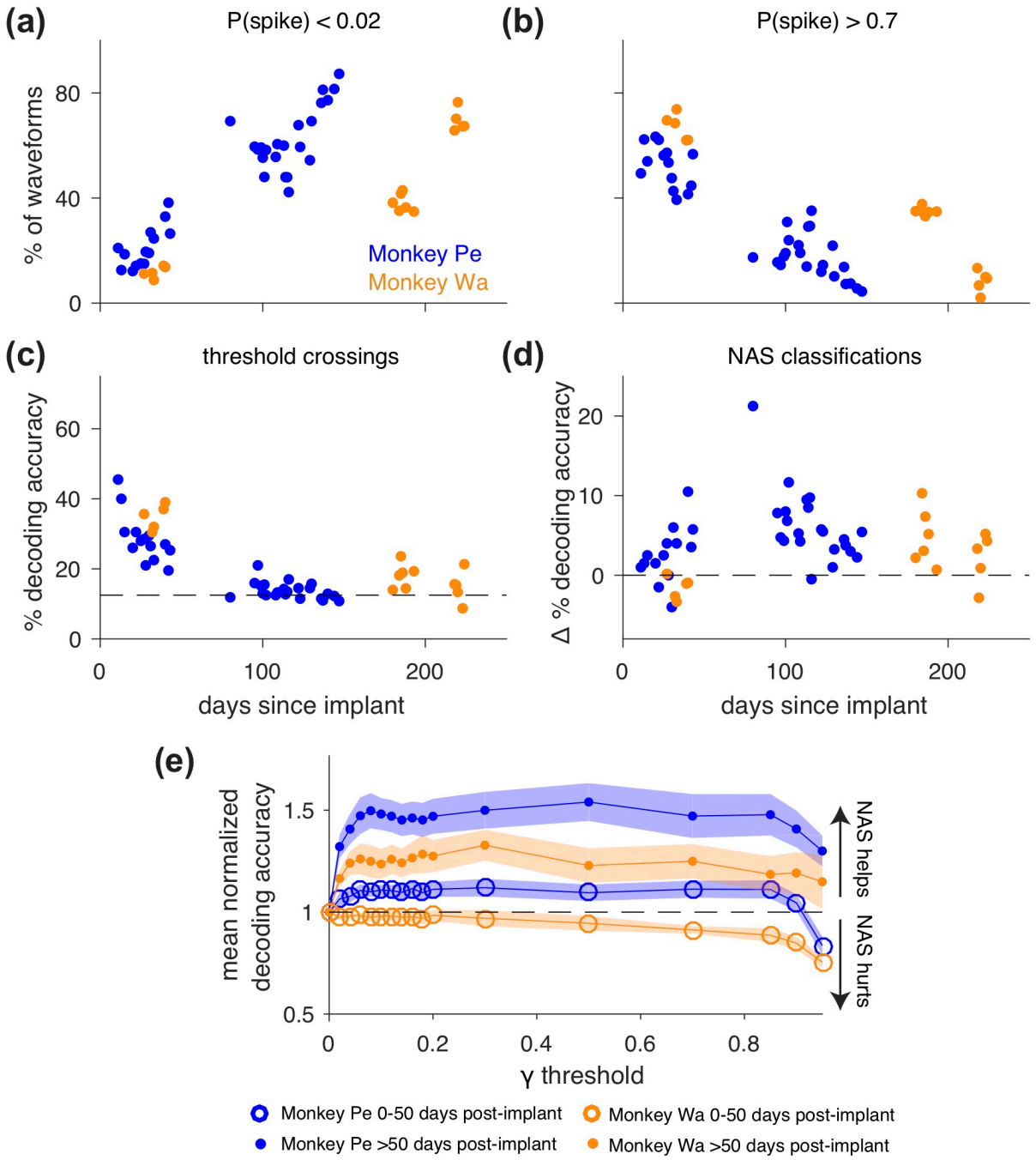


Figure 5. Noise on the array increased with the age of the array while decoding accuracy decreased in Monkey Pe (blue) and Monkey Wa (orange). (a) Percentage of waveforms with a P(spike) < 0.02 over time. A waveform with a P(spike) less than 0.02 was one that the network found very unlikely to be a spike. The percentage of these unlikely spike waveforms increased as the array became older (Spearman's rank-order, Monkey Pe: $\varrho = 0.87$, $p < 0.0001$; Monkey Wa: $\varrho = 0.90$, $p < 0.0001$). (b) Percentage of waveforms with a P(spike) > 0.70 over time. The percentage of waveforms that the network found to be strongly spike-like decreased as the array became older (Spearman's rank-order, Monkey Pe: $\varrho = -0.87$, $p < 0.0001$; Monkey Wa: $\varrho = -0.94$, $p < 0.0001$). (c) Decoding accuracy with threshold crossings (i.e. $\gamma = 0$) decreased as the array aged (Spearman's rank-order, Monkey Pe: $\varrho = -0.86$, $p < 0.0001$; Monkey Wa: $\varrho = -0.69$, $p = 0.004$). (d) Change in percent decoding accuracy using network-classified spikes relative to decoding accuracy with

threshold crossings (Δ % decoding accuracy). We computed a distribution of the maximum γ threshold (similar to Fig. 4c-d) for sessions that were 0-50 days post array implant and used the median to set the γ threshold before computing decoding accuracy for those sessions. We repeated this for sessions more than 50 days post array implant. In Monkey Pe, using NAS classifications improved decoding accuracy (2-tailed, Wilcoxon signed rank test, p < 0.0001) and in Monkey Wa, using the classifications neither hurt nor helped decoding significantly (p = 0.09). In both subjects, the network helped decoding more in the late array sessions (>50 days post implant) compared to the early sessions (2-tailed, Wilcoxon rank sum, Monkey Pe: p = 0.01; Monkey Wa: p = 0.01). (e) Mean normalized decoding accuracy across early sessions (open circles) and late sessions (filled circles) as a function of γ threshold. Using the network classifications at any γ threshold greater than zero was more helpful for late sessions than it was for early sessions.

Comparison to spike-sorting

In light of the mixed benefits of spike-sorting on decoding in the literature (10, 18, 22, 25), we sought to evaluate the relative merits of manual spike-sorting and our network classifier on decoding accuracy in our data. We took advantage of the offline spike-sorting that had already been applied to these data and calculated the decoding accuracy of spike-sorted data (Fig. 6, data aggregated across subjects in the marginal histograms). Decoding using spike-sorted waveforms was better than using threshold crossings (Fig. 6 right marginal; 2-tailed, Wilcoxon signed rank test, $p < 0.0001$) similar to decoding with our network classifications (Fig. 6 top marginal, data from Fig. 5d aggregated across subjects; $p < 0.0001$). Our network's classifications were at least as helpful as manual spike-sorting for decoding relative to using threshold crossings (Fig. 6 center; paired, 2-tailed, Wilcoxon signed rank test, Monkey Pe: $p = 0.75$; Monkey Wa: $p = 0.09$). Furthermore, the automated real-time operation of our network confers a distinct advantage over manual spike-sorting.

Together, our analyses demonstrate that our network was able to produce classifications of spikes and noise that matched our qualitative expectations and could be tuned quantitatively to adjust its permissiveness. Compared to using threshold crossings, our network's classifications most often led to an improvement or minimally harmful change in decoding accuracy, indicating that there was usually little downside to the application of our network. These outcomes are valuable particularly in the context of a BCI where real time operation and maximal decoding accuracy are prized.

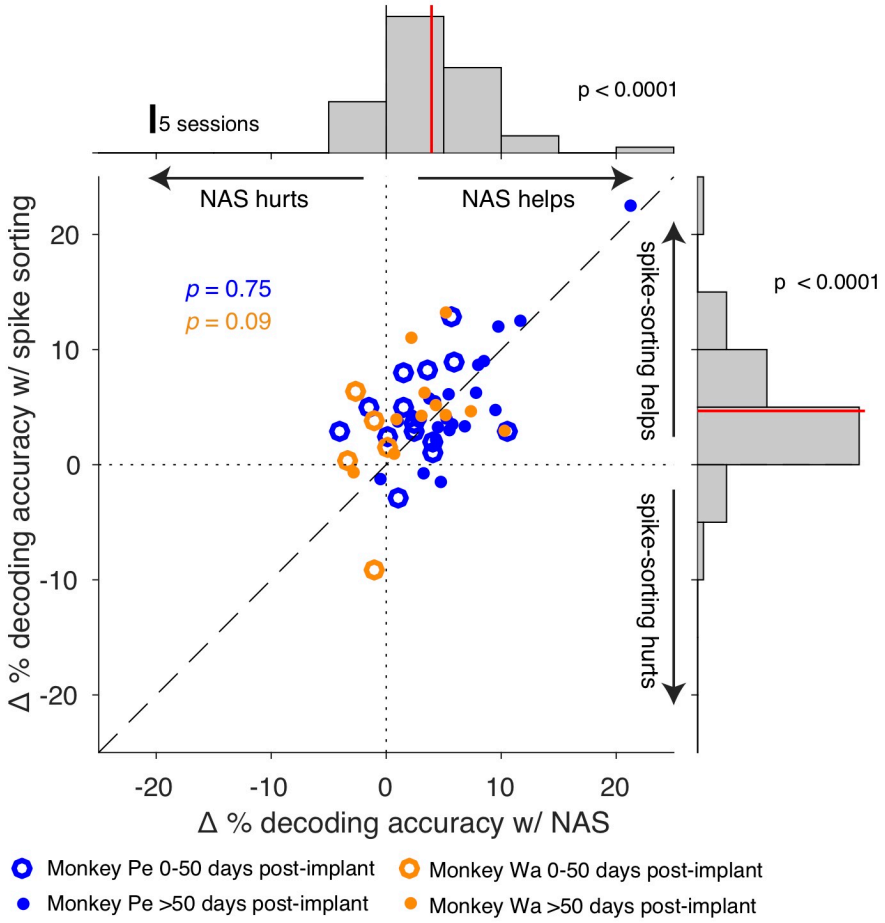


Figure 6. Using NAS classifications was comparable to using spike-sorted data for decoding. Δ % decoding accuracy was calculated for both NAS classifications and spike-sorted waveforms as the change from decoding accuracy with threshold crossings. Top: Distribution of Δ % decoding accuracy using NAS classifications aggregated across subjects (γ threshold selected as in 5d). Using NAS classifications improved decoding accuracy from that with threshold crossings (2-tailed, Wilcoxon signed rank test, p < 0.0001). Right: Distribution of Δ % decoding accuracy using manual spike-sorting aggregated across subjects. Using spike-sorted classifications also improved decoding accuracy from that with threshold crossings (2-tailed, Wilcoxon signed rank test, p < 0.0001). Center: The joint distribution of Δ % decoding accuracy with spike-sorting and NAS classified spikes. Using our network's classifications was at least as helpful as spike-sorting for decoding (paired, 2-tailed, Wilcoxon signed rank test, Monkey Pe: p = 0.75; Monkey Wa: p = 0.09).

Discussion

By leveraging our pool of previously spike-sorted data, we trained a neural network classifier to separate spikes from noise in extracellular electrophysiological recordings in real time. To assess the value of this method, we used an objective criterion – the ability to decode remembered location from prefrontal cortex in a memory-guided saccade task. We found that we could set a fixed γ threshold that resulted in an improvement or at worst a small change in decoding accuracy for the majority of sessions,

447 and this effect persisted even as recording quality decreased over time. Compared to spike-sorting
448 (performed offline with substantial manual labor involved), classifying spikes with our network had
449 similar effects on decoding accuracy relative to threshold crossings and was less tedious.

450 *Training the network*

451 Developing a neural network trained on spike-sorted data allowed us to capture the essence of the
452 rules by which experienced researchers distinguish spikes from noise without explicitly defining all the
453 variables and exceptions, a standard challenge in spike classification (15). We found that a simple neural
454 network was quite good at capturing the nuances of the spike-sorted training set. In fact, the network was
455 sensitive enough to capture a researcher's specific spike-sorting tendencies, such that networks trained on
456 a particular individual's spike-sorted data performed better on classifying test data from that individual
457 than from other individual spike-sorters. Lee *et al*. (13) similarly demonstrated the promise of using
458 spike-sorted data as a training set for retinal recordings with dense arrays. Our findings extend this work
459 in a context where the training data were less extensively curated (spike-sorting performed by multiple
460 researchers), the data were noisier (cortical recordings in an alert macaque monkey) and the recording
461 electrodes were spatially sparse (at least 150 μm apart for U-probes and 400 μm for Utah arrays), a
462 situation that is common in numerous electrophysiology laboratories. A recent related effort (37) used a
463 neural network trained on hand-sorted spiking data to determine whether channels contained any spikes
464 for the purpose of rejecting channels with only artifacts. Our work is synergistic with their findings, and
465 extends them to demonstrate that removal of noise from single channels can lead to a positive impact on
466 decoding accuracy.

467 There were potentially misclassifications in our training set, as human spike-sorting is both
468 subjective and prone to error (16, 17). However, the network was trained on many waveforms to
469 minimize the impact of any confounders, including waveforms from multiple subjects, brain regions, and
470 human spike-sorters. An ideal and even more generalizable training set might include waveforms from
471 different research groups, artificially simulated neural waveforms and noise in which ground truth could
472 be established, and/or pre-defined quantities of certain types of spikes (e.g. shifted spikes, slow and fast

spike waveforms) and noise (e.g. electrical artifacts) to target what the network learns. However, no matter how much training data is used, an ideal training set could only be defined in the context of objective metrics to assess the trained network's performance.

Evaluating the network with decoding

Since there is no obvious ground truth for classifying spikes from extracellular array recordings, one objective measure with practical benefit for BCI performance is how well we could decode information from neural activity. We used independent test data to assess the effect of classifications from our trained network on decoding accuracy. Although the expectation might be that removing noise from neural data would improve decoding, historically doing so with spike-sorted classifications has had mixed effects, in some cases hurting performance (18). Nevertheless, many BCIs currently operate with some amount of online noise removal or visual inspection (such as disabling visibly noisy channels) to pre-process neural data before decoding (1, 2, 38, 39). The question of whether to spike-sort BCI data prior to decoding or simply use all threshold crossings has been a source of debate with no clear resolution.

Our findings provide additional support, across multiple sessions over time, that removing certain noise waveforms prior to decoding provides an advantage over using all threshold crossings, consistent with findings from Fraser *et al.* (22) and Christie *et al.* (10). Contrary to our results, another study by Todorova *et al.* (18) found that using spike-sorted waveforms hurt decoding when noise waveforms were discarded. It is difficult to speculate why removing noise waveforms hurt their result but helped ours given that our methods of noise removal were different. Todorova *et al*. also found assigning noise waveforms from spike-sorting to a new unit, rather than discarding them, improved decoding relative to threshold crossings (18). Yet, when we assigned our noise waveforms to separate units, decoding accuracy was worse or unchanged compared to discarding noise waveforms both with our spike-sorted classifications and our network classifications (Supplementary Figure 5). Importantly, we defined 'noise' operationally as any waveform that did not exceed the γ threshold of our network setting. This definition of noise surely included both electrical artifacts as well as indistinct shapes that were actually neural in origin. Removal of waveforms from the former class could only help decoding, but removal of the latter

499 class might actually harm decoding if the multiunit activity contained in those waveforms has information
500 about the condition of interest for decoding.

501 Apparent noise in neuronal activity, in addition to changing the activity of single electrode
502 channels, can also be correlated across channels. In the activity of pairs of single neurons, such
503 fluctuations in trial-to-trial correlated variability are sometimes termed "noise correlation" (also known as
504 spike count correlation, or r_{sc}). Noise correlation can have a substantial impact on the ability of a decoder
505 to extract information from neuronal populations (Averbeck et al, 2006, Kohn et al, 2016), and the
506 activity of multi-unit groups of neurons has a higher noise correlation than that of the constituent pairs of
507 neurons (Cohen and Kohn, 2011). Since our network acted on threshold crossings, which typically
508 contain the activity of many individual neurons, it likely also impacted the magnitude of the noise
509 correlation between channels. Our decoder assumed that the noise on each unit was independent, a choice
510 common to BCI, which permitted training the decoder with a smaller quantity of trials than is necessary to
511 learn the covariance structure in the population. The effect of our network on decoding surely depends on
512 not only the structure of the network and the choice of γ threshold, but also the structure of noise present
513 in the population and the sensitivity of the decoder to that noise.

514 In addition to understanding how the type of noise present in the data impacts decoding, it may
515 also be useful to evaluate the types of neural waveforms that contribute to decoding. Different levels of
516 single and multi-unit activity are captured during a neural recording depending on the threshold set by the
517 researcher (i.e. the minimum voltage level at which the neural activity is marked as a spike). By adjusting
518 this threshold, we change the candidate waveforms available for decoding. Oby *et al.* (26) found that the
519 optimal threshold depended on the type of information being extracted from neural activity. Another
520 study found that higher thresholds, which captured less multi-unit activity, resulted in worse decoding of
521 direction from M1 (10). These studies in the context of our own findings highlight the need to identify
522 specific types of waveforms that contribute to the decodable information in different brain regions and
523 task contexts.

524 *Other variables that impact decoding accuracy*

525 A key concern for practical implementation of BCIs is decoding stability. Array recordings often
526 get visibly noisier over time, potentially as a result of an immune response to the implant (40) or a
527 physical shift in position (41). In our study, as the time since array implant increased there were more
528 "low probability" waveforms and fewer "high probability" waveforms, indicating the array had fewer
529 well-defined single-unit waveforms. In line with this, several other studies with chronic array implants
530 have found that the number of single-units decreased over time (42-45). Nevertheless, there are mixed
531 findings on the stability of decoding performance in chronic array implants. Some studies have found
532 decoder performance with threshold crossings remained stable over time (23, 24, 46, 47); however, other
533 studies, including our own, observed a decrease in decoding accuracy over time (48, 49). Unlike the
534 previous literature investigating the effects of threshold crossings and spike-sorting on decoding
535 performance, we used neural activity from PFC, where decoding accuracy is typically lower (50-56) than
536 in motor cortex (7, 57-59). Stability of decoding accuracy with array age may also depend on the brain
537 region of the implant, the subject's task, initial decoding performance, and the choice to train the decoder
538 daily (23, 24) or hold it constant (46). The beneficial effects of our network on decoding were particularly
539 salient in the case of older array implants, where decoding accuracy had decreased over time.

540 *Extensions for NAS*

541 Our neural network-based spike sorter is a promising tool for both offline spike-sorting and
542 improving online decoding performance. However, in designing our network we only scratched the
543 surface of many potential avenues to address these challenges. We found that using networks with
544 different numbers of hidden units and layers did not substantially alter decoding accuracy even though
545 there were some differences in how these different sized networks classified waveforms. Given that our
546 network was relatively simple, it would be possible to implement similar operations with alternative
547 algorithms such as logistic regression. We chose a neural network because it was easily trainable from
548 existing data and there are many ways to modularly build upon its complexity. Although we opted to use
549 50 hidden units and one hidden layer, it is possible that a more complex network with additional filtering
550 operations and more categories of waveform classification may result in improved decoding performance

and could help to create a more robust spike-sorter or a pre-processing step for offline analyses. However, our work had the value of exploring how a smaller network trained on spike-sorted data could identify features that were most valuable for decoding. While training on spike-sorted data was an efficient choice because of its easy availability, it may not be the best choice in regimes where canonical spike-shapes, which spike-sorted classifications prioritize, do not necessarily carry the most decodable information compared to less well-defined multi-unit activity (11, 23). If the goal is to improve decoding performance, it might be valuable to train a network to select waveforms or units that positively contribute to decoding. Ventura and Todorova (19) developed a method for identifying the information contribution of units for decoding. Such a method could be used to develop a training set of neural waveforms that are assigned a label based on how much they positively or negatively contributed to decoding.

Overall, we developed a new tool for pre-processing BCI data that classified threshold crossings in a tunable manner that was beneficial for decoding. Furthermore, we showed that spike-sorted waveform data are an underused training set with potential for building spike classification networks. A neural network-based spike classifier has the potential to reduce the need for human intervention in spike-sorting while maintaining sensitivity to nuances that many automated sorting algorithms miss. Our tunable classifier is a step toward pre-processing methods that both optimize and stabilize online decoding performance.

Acknowledgements

D.I. was supported by National Science Foundation (NSF) DBI 1659611, National Institutes of Health (NIH) T32 GM008208, and the ARCS foundation Thomas-Pittsburgh Chapter Award. M.A.S. was supported by NIH R01EY022928, R01MH118929, R01EB026953, and P30EY008098; NSF NCS 1734901; a career development grant and an unrestricted award from Research to Prevent Blindness; the Eye and Ear Foundation of Pittsburgh. R.C.W. was supported by a Richard King Mellon Foundation Presidential Fellowship in the Life Sciences (RCW). S.B.K. was supported by T32

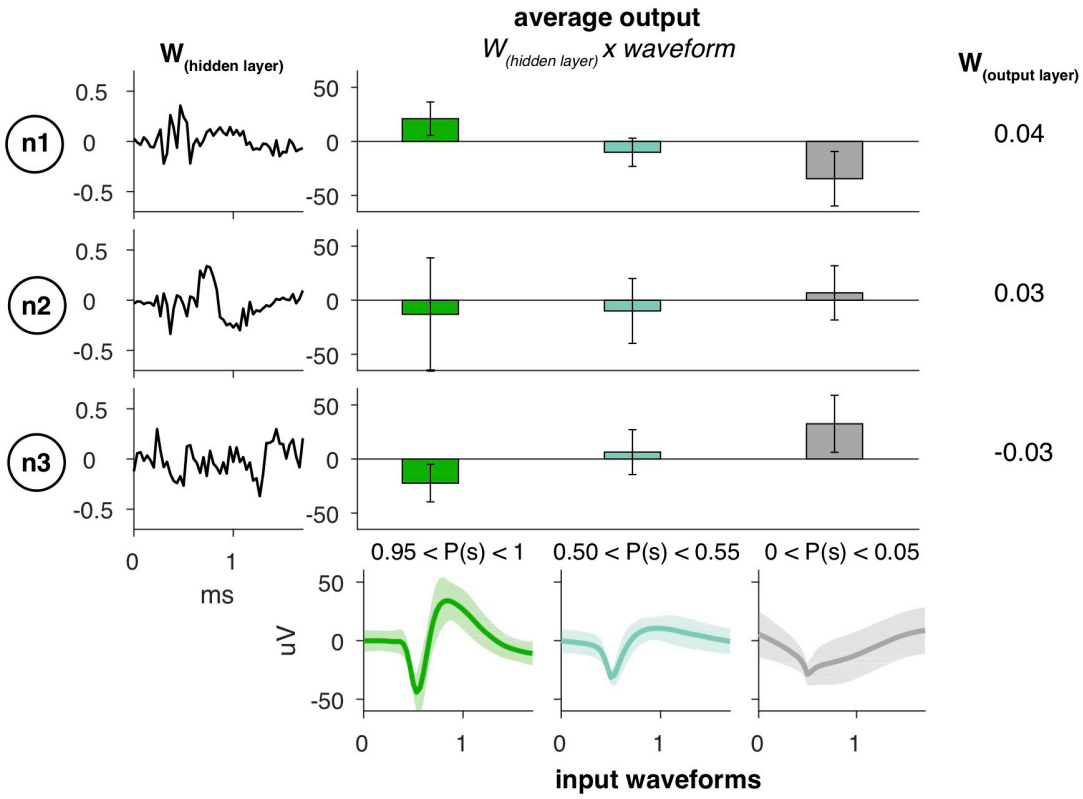
EY017271-08. We would like to thank Samantha Schmitt for assistance with data collection and Ben Cowley, Byron Yu, and Steve Chase for their valuable insight on this project.

References

1. Sadtler PT, Quick KM, Golub MD, Chase SM, Ryu SI, Tyler-Kabara EC, et al. Neural constraints on learning. Nature. 2014;512(7515):423-6.
2. Chase SM, Kass RE, Schwartz AB. Behavioral and neural correlates of visuomotor adaptation observed through a brain-computer interface in primary motor cortex. J Neurophysiol. 2012;108(2):624-44.
3. Golub MD, Yu BM, Chase SM. Internal models for interpreting neural population activity during sensorimotor control. Elife. 2015;4.
4. Schafer RJ, Moore T. Selective attention from voluntary control of neurons in prefrontal cortex. Science. 2011;332(6037):1568-71.
5. Velliste M, Perel S, Spalding MC, Whitford AS, Schwartz AB. Cortical control of a prosthetic arm for self-feeding. Nature. 2008;453(7198):1098-101.
6. Hochberg LR, Bacher D, Jarosiewicz B, Masse NY, Simeral JD, Vogel J, et al. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. Nature. 2012;485(7398):372-5.
7. Collinger JL, Wodlinger B, Downey JE, Wang W, Tyler-Kabara EC, Weber DJ, et al. High-performance neuroprosthetic control by an individual with tetraplegia. Lancet. 2013;381(9866):557-64.
8. Averbeck BB, Latham PE, Pouget A. Neural correlations, population coding and computation. Nat Rev Neurosci. 2006;7(5):358-66.
9. Lewicki MS. A review of methods for spike sorting: the detection and classification of neural action potentials. Network. 1998;9(4):R53-78.
10. Christie BP, Tat DM, Irwin ZT, Gilja V, Nuyujukian P, Foster JD, et al. Comparison of spike sorting and thresholding of voltage waveforms for intracortical brain-machine interface performance. J Neural Eng. 2015;12(1):016009.
11. Stark E, Abeles M. Predicting movement from multiunit activity. J Neurosci. 2007;27(31):8387-94.
12. Rey HG, Pedreira C, Quian Quiroga R. Past, present and future of spike sorting techniques. Brain Res Bull. 2015;119(Pt B):106-17.
13. Chung JE, Magland JF, Barnett AH, Tolosa VM, Tooker AC, Lee KY, et al. A Fully Automated Approach to Spike Sorting. Neuron. 2017;95(6):1381-94 e6.
14. Chaure FJ, Rey HG, Quian Quiroga R. A novel and fully automatic spike-sorting implementation with variable number of features. J Neurophysiol. 2018;120(4):1859-71.
15. Wood F, Black MJ, Vargas-Irwin C, Fellows M, Donoghue JP. On the variability of manual spike sorting. IEEE Trans Biomed Eng. 2004;51(6):912-8.
16. Pedreira C, Martinez J, Ison MJ, Quian Quiroga R. How many neurons can we see with current spike sorting algorithms? Journal of Neuroscience Methods. 2012;211(1):58-65.
17. Rossant C, Kadir SN, Goodman DFM, Schulman J, Hunter MLD, Saleem AB, et al. Spike sorting for large, dense electrode arrays. Nat Neurosci. 2016;19(4):634-41.
18. Todorova S, Sadtler P, Batista A, Chase S, Ventura V. To sort or not to sort: the impact of spike-sorting on neural decoding performance. J Neural Eng. 2014;11(5):056005.
19. Ventura V, Todorova S. A computationally efficient method for incorporating spike waveform information into decoding algorithms. Neural Comput. 2015;27(5):1033-50.
20. Kloosterman F, Layton SP, Chen Z, Wilson MA. Bayesian decoding using unsorted spikes in the rat hippocampus. J Neurophysiol. 2014;111(1):217-27.
21. Santhanam G, Sahani M, Ryu S, Shenoy K. An extensible infrastructure for fully automated spike sorting during online experiments. Conf Proc IEEE Eng Med Biol Soc. 2004;6:4380-4.

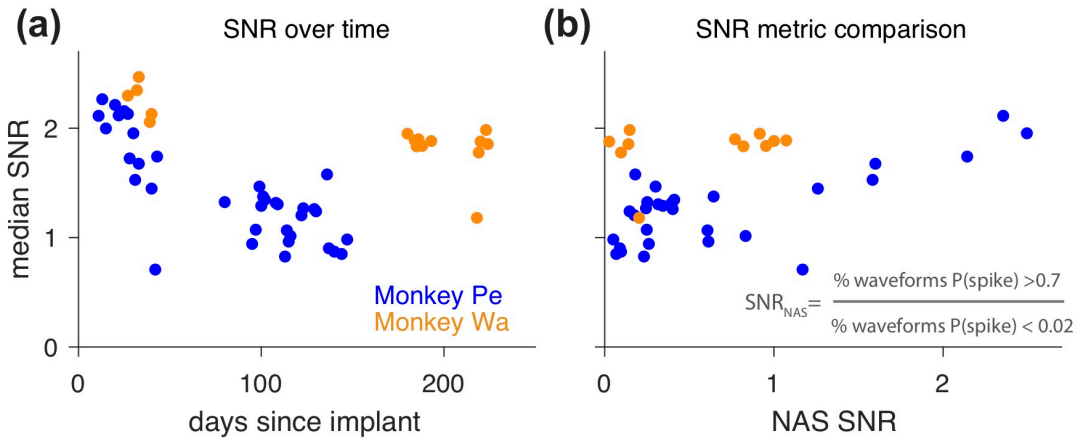
22. Fraser GW, Chase SM, Whitford A, Schwartz AB. Control of a brain-computer interface without spike sorting. J Neural Eng. 2009;6(5):055004.
23. Chestek CA, Gilja V, Nuyujukian P, Foster JD, Fan JM, Kaufman MT, et al. Long-term stability of neural prosthetic control signals from silicon cortical arrays in rhesus macaque motor cortex. J Neural Eng. 2011;8(4):045005.
24. Gilja V, Nuyujukian P, Chestek CA, Cunningham JP, Yu BM, Fan JM, et al. A high-performance neural prosthesis enabled by control algorithm design. Nat Neurosci. 2012;15(12):1752-7.
25. Dai J, Zhang P, Sun H, Qiao X, Zhao Y, Ma J, et al. Reliability of motor and sensory neural decoding by threshold crossings for intracortical brain-machine interface. J Neural Eng. 2019;16(3):036011.
26. Oby ER, Perel S, Sadtler PT, Ruff DA, Mischel JL, Montez DF, et al. Extracellular voltage threshold settings can be tuned for optimal encoding of movement and stimulus parameters. J Neural Eng. 2016;13(3):036009.
27. Bishop W, Chestek CC, Gilja V, Nuyujukian P, Foster JD, Ryu SI, et al. Self-recalibrating classifiers for intracortical brain-computer interfaces. J Neural Eng. 2014;11(2):026001.
28. Shoham S, Fellows MR, Normann RA. Robust, automatic spike sorting using mixtures of multivariate t-distributions. J Neurosci Methods. 2003;127(2):111-22.
29. Khanna SB, Snyder AC, Smith MA. Distinct Sources of Variability Affect Eye Movement Preparation. J Neurosci. 2019;39(23):4511-26.
30. Snyder AC, Morais MJ, Kohn A, Smith MA. Correlations in V1 are reduced by stimulation outside the receptive field. J Neurosci. 2014;34(34):11222-7.
31. Snyder AC, Morais MJ, Willis CM, Smith MA. Global network influences on local functional connectivity. Nat Neurosci. 2015;18(5):736-43.
32. Snyder AC, Smith MA. Stimulus-dependent spiking relationships with the EEG. J Neurophysiol. 2015;114(3):1468-82.
33. Kelly RC, Smith MA, Samonds JM, Kohn A, Bonds AB, Movshon JA, et al. Comparison of recordings from microelectrode arrays and single electrodes in the visual cortex. J Neurosci. 2007;27(2):261-4.
34. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. 2016.
35. Chollet F. Keras. 2015.
36. Diederik P. Kingma JB. Adam: a Method for Stochastic Optimization. International Conference on Learning Representations2015.
37. Saif-Ur-Rehman M, Lienkamper R, Parpaley Y, Wellmer J, Liu C, Lee B, et al. SpikeDeeptector: A deep-learning based method for detection of neural spiking activity. J Neural Eng. 2019.
38. Homer ML, Nurmikko AV, Donoghue JP, Hochberg LR. Sensors and decoding for intracortical brain computer interfaces. Annu Rev Biomed Eng. 2013;15:383-405.
39. Hochberg LR, Serruya MD, Friehs GM, Mukand JA, Saleh M, Caplan AH, et al. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. Nature. 2006;442(7099):164-71.
40. Ward MP, Rajdev P, Ellison C, Irazoqui PP. Toward a comparison of microelectrodes for acute and chronic recordings. Brain Res. 2009;1282:183-200.
41. Perge JA, Homer ML, Malik WQ, Cash S, Eskandar E, Friehs G, et al. Intra-day signal instabilities affect decoding performance in an intracortical neural interface system. J Neural Eng. 2013;10(3):036004.
42. Dickey AS, Suminski A, Amit Y, Hatsopoulos NG. Single-unit stability using chronically implanted multielectrode arrays. J Neurophysiol. 2009;102(2):1331-9.
43. Downey JE, Schwed N, Chase SM, Schwartz AB, Collinger JL. Intracortical recording stability in human brain-computer interface users. J Neural Eng. 2018;15(4):046016.
44. Fraser GW, Schwartz AB. Recording from the same neurons chronically in motor cortex. J Neurophysiol. 2012;107(7):1970-8.

674 45. Tolias AS, Ecker AS, Siapas AG, Hoenselaar A, Keliris GA, Logothetis NK. Recording
675 chronically from the same neurons in awake, behaving primates. J Neurophysiol. 2007;98(6):3780-90.
676 46. Nuyujukian P, Kao JC, Fan JM, Stavisky SD, Ryu SI, Shenoy KV. Performance sustaining
677 intracortical neural prostheses. J Neural Eng. 2014;11(6):066003.
678 47. Flint RD, Scheid MR, Wright ZA, Solla SA, Slutzky MW. Long-Term Stability of Motor Cortical
679 Activity: Implications for Brain Machine Interfaces and Optimal Feedback Control. J Neurosci.
680 2016;36(12):3623-32.
681 48. Perge JA, Zhang S, Malik WQ, Homer ML, Cash S, Friehs G, et al. Reliability of directional
682 information in unsorted spikes and local field potentials recorded in human motor cortex. J Neural Eng.
683 2014;11(4):046007.
684 49. Wang D, Zhang Q, Li Y, Wang Y, Zhu J, Zhang S, et al. Long-term decoding stability of local
685 field potentials from silicon arrays in primate motor cortex during a 2D center out task. J Neural Eng.
686 2014;11(3):036009.
687 50. Meyers EM, Freedman DJ, Kreiman G, Miller EK, Poggio T. Dynamic population coding of
688 category information in inferior temporal and prefrontal cortex. J Neurophysiol. 2008;100(3):1407-19.
689 51. Rizzuto DS, Mamelak AN, Sutherling WW, Fineman I, Andersen RA. Spatial selectivity in
690 human ventrolateral prefrontal cortex. Nat Neurosci. 2005;8(4):415-7.
691 52. Tremblay S, Doucet G, Pieper F, Sachs A, Martinez-Trujillo J. Single-Trial Decoding of Visual
692 Attention from Local Field Potentials in the Primate Lateral Prefrontal Cortex Is Frequency-Dependent. J
693 Neurosci. 2015;35(24):9038-49.
694 53. Jia N, Brincat SL, Salazar-Gomez AF, Panko M, Guenther FH, Miller EK. Decoding of intended
695 saccade direction in an oculomotor brain-computer interface. J Neural Eng. 2017;14(4):046007.
696 54. Boulay CB, Pieper F, Leavitt M, Martinez-Trujillo J, Sachs AJ. Single-trial decoding of intended
697 eye movement goals from lateral prefrontal cortex neural ensembles. J Neurophysiol. 2016;115(1):486-
698 99.
699 55. Parthasarathy A, Herikstad R, Bong JH, Medina FS, Libedinsky C, Yen SC. Mixed selectivity
700 morphs population codes in prefrontal cortex. Nat Neurosci. 2017;20(12):1770-9.
701 56. Spaak E, Watanabe K, Funahashi S, Stokes MG. Stable and Dynamic Coding for Working
702 Memory in Primate Prefrontal Cortex. J Neurosci. 2017;37(27):6503-16.
703 57. Sadtler PT, Ryu SI, Tyler-Kabara EC, Yu BM, Batista AP. Brain-computer interface control
704 along instructed paths. J Neural Eng. 2015;12(1):016015.
705 58. Masse NY, Jarosiewicz B, Simeral JD, Bacher D, Stavisky SD, Cash SS, et al. Non-causal spike
706 filtering improves decoding of movement intention for intracortical BCIs. J Neurosci Methods.
707 2014;236:58-67.
708 59. Koralek AC, Jin X, Long JD, 2nd, Costa RM, Carmena JM. Corticostriatal plasticity is necessary
709 for learning intentional neuroprosthetic skills. Nature. 2012;483(7389):331-5.
710

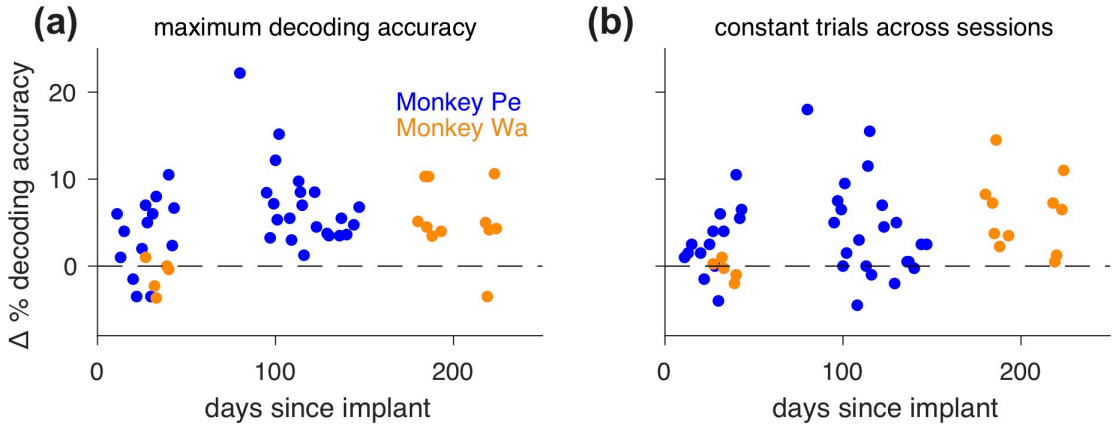


Supplementary Figure 1. Weights for three sample hidden units (n1, n2, n3) of the neural network hidden layer and their outputs. To demonstrate how the network classified waveforms, we used 1000 example input waveforms from each of three different categories (bottom three panels: mean and standard deviation of the waveforms in each category). The first category included very spike-like waveforms (i.e. those that were assigned a P(spike) value between 0.95 and 1 by the network, green), the second included moderately spike-like waveforms (0.50 < P(spike) < 0.55, blue), and the third included noise (0<P(spike)<0.05, grey). To visualize the hidden layer, we selected three out of the 50 hidden units and plotted their weights (left panels, $W_{\text{hidden layer}}$). Each waveform was linearly scaled by the weight vector of units n1, n2, and n3 and summed (middle panels: mean and standard deviation of the scalar outputs of $W_{\text{hidden layer}}$ x waveform; waveform category indicated by bar color and column, and unit # indicated by row). The output of the hidden layer was passed through a ReLU non-linearity before being multiplied by an output weight (right, $W_{\text{output layer}}$). The resultant values were summed and passed through a sigmoid non-linearity to produce a P(spike) value (not shown).

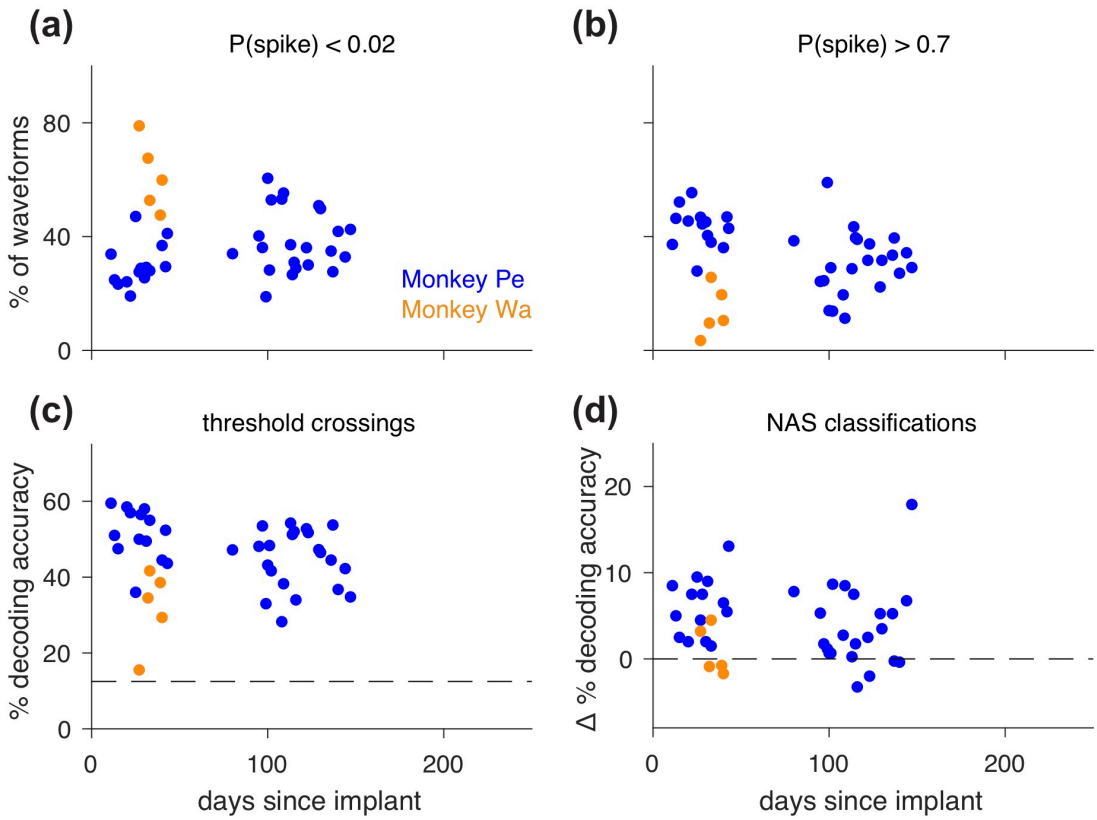
Overall, the network's hidden units were difficult to characterize; however, we observed that certain hidden units tended to have a larger, positive response to very spike-like waveforms (e.g. n1 green bar) while others had a larger, positive response to noise waveforms (e.g. n3 grey bar). The positive $W_{\text{output layer}}$ for n1 would then weigh the final summed output towards a P(spike) of 1 and the negative $W_{\text{output layer}}$ for n3 would weigh the output towards a P(spike) of 0. Unlike n1 and n3, n2 had variable responses to different types of waveforms, indicating that not every unit had a well-defined spike or noise preference. Overall, we found that different units in the hidden layer focused on features that were associated with different types of spikes in the training data. Note that in the network, the contribution of bias values was not shown in this visualization.



Supplementary Figure 2. SNR metrics decreased as the array aged for Monkey Pe (blue) and Monkey Wa (orange). **(a)** SNR was computed for each channel on the array using the method described in Kelly et al. (2007). The median SNR of the array decreased over time in both subjects (Spearman's rank-order, Monkey Pe: $\varrho = -0.77$, $p < 0.0001$; Monkey Wa: $\varrho = -0.72$, $p = 0.002$). **(b)** SNR_{NAS} (i.e. the ratio of the percentage of waveforms with a P(spike) > 0.70 to the percentage of waveforms with a P(spike) < 0.02) served as an alternate measure of SNR as it was correlated with median SNR over time (Spearman's rank-order, Monkey Pe: $\varrho = 0.78$, $p < 0.0001$; Monkey Wa: $\varrho = 0.75$, $p = 0.01$).



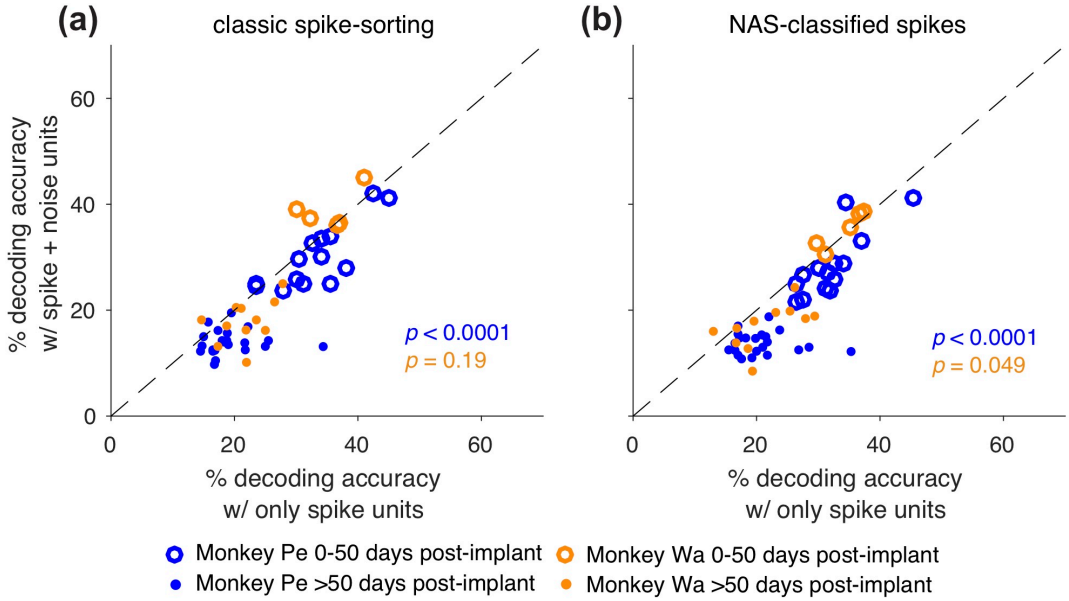
Supplementary Figure 3. Change in percent decoding accuracy from that with threshold crossings (Δ % decoding accuracy) over time, as in Figure 5 except **(a)** using the maximum decoding accuracy for each session and **(b)** keeping the number of trials used to decode constant across all sessions. In (a) the maximum decoding accuracy across all γ thresholds greater than 0 on each session was used to show the best possible network performance if γ was chosen separately in each session (and tested in a cross-validated fashion). Δ % decoding accuracy was significantly greater than zero for both subjects across all sessions (2-tailed, Wilcoxon signed rank test, Monkey Pe: $p < 0.0001$; Monkey Wa: $p = 0.01$). There was a significant difference between the early array (0-50 days post implant) and the late array (> 50 days post implant) sessions in Monkey Wa, but not in Monkey Pe (2-tailed, Wilcoxon rank sum, Monkey Pe: $p = 0.09$; Monkey Wa: $p = 0.01$). In (b), decoding accuracy was computed for each session with 25 trials for Monkey Pe and 50 trials for Monkey Wa. The γ threshold was selected using the same method described for Fig. 4d. Δ % decoding accuracy again was greater than zero for both subjects across all sessions (2-tailed, Wilcoxon signed rank test, Monkey Pe: $p < 0.0001$; Monkey Wa: $p = 0.002$). Again, there was a significant difference between the early array (0-50 days post implant) and the late array (> 50 days post implant) sessions in Monkey Wa, but not in Monkey Pe (2-tailed, Wilcoxon rank sum, Monkey Pe: $p = 0.68$; Monkey Wa $p = 0.001$).



Supplementary Figure 4. The network had a similar effect on decoding from V4 96-electrode Utah array implants in each animal as it did on decoding from the PFC array in the same sessions from the same animals (Monkey Pe- blue, Monkey Wa- orange). This analysis was done using the same steps as in Figure 5 except using waveforms from the V4 array instead of the PFC array. Monkey Wa did not have any usable late session (> 50 days post implant) data from the V4 array. As a result, only Monkey Pe was tested for trends over time.

(a) Percentage of waveforms with a P(spike) < 0.02 over time. A waveform with a P(spike) less than 0.02 was one that the network found very unlikely to be a spike. The percentage of these unlikely spike waveforms only slightly but significantly increased as Monkey Pe's array became older (Spearman's rank-order, $\varrho = 0.43$, $p = 0.01$). **(b)** Percentage of waveforms with a P(spike) > 0.70 over time. The percentage of waveforms that the network found to be strongly spike-like decreased as Monkey Pe's array became older ($\varrho = -0.48$, $p = 0.003$). The magnitude of these changes over time was much smaller in V4 compared to PFC. **(c)** Decoding accuracy with threshold crossings (i.e. $\gamma = 0$) decreased by a small amount as the array aged in Monkey Pe ($\varrho = -0.41$, $p = 0.01$). **(d)** Change in percent decoding accuracy using network classified spikes relative to decoding accuracy with threshold crossings (Δ % decoding accuracy). We computed a distribution of the maximum γ threshold (using the same method as in Figure 4c-d) for sessions that were 0-50 days post array implant and used the median to set the γ threshold before computing decoding accuracy for those sessions. We repeated this for sessions more than 50 days post array implant. In Monkey Pe, using NAS classifications improved decoding accuracy (2-tailed, Wilcoxon signed rank test, $p < 0.0001$) and in Monkey Wa, using the classifications neither hurt nor helped decoding significantly ($p = 0.81$). There was a significant difference between the early array (0-50 days post implant) and the late array (> 50 days post implant) sessions in Monkey Pe (2-tailed, Wilcoxon rank sum, $p = 0.0496$).

Thus, despite differences in the array quality between PFC and V4 (as demonstrated by the higher threshold crossing decoding accuracies in V4 compared to PFC in Fig. 5), using our network classifications still proved beneficial in one subject and had no significant impact on the other subject.



Supplementary Figure 5. Assigning waveforms classified as noise on each electrode to a new unit rather than discarding them did not improve decoding performance in most sessions for Monkey Pe (blue) or Monkey Wa (orange). In the main text analyses, any waveform classified as noise via our network or via spike-sorting was discarded. However, Todorova et al. (2014) found that creating a separate noise unit for the noise waveforms from each unit actually led to better decoding performance than discarding the noise. We sought to assess whether this result held for our data. **(a)** Using the manually spike-sorted classifications, we added a new unit to accompany each spike unit that contained all waveforms classified as noise (plotted on the y-axis). Percent decoding accuracy was worse with noise units for Monkey Pe (paired, 2-tailed, Wilcoxon signed rank test- $p < 0.0001$) and not significantly different for Monkey Wa ($p = 0.19$). **(b)** Similar to (a), except using NAS classifications. For each session, we discarded the noise units and found the γ threshold that resulted in the maximum decoding accuracy. We classified waveforms as spikes or noise using that threshold and then computed decoding accuracy with and without the noise units for each session (for the latter, the test set was independent from the set used to select the γ threshold). Discarding noise resulted in a higher decoding accuracy for most sessions in both subjects (Monkey Pe: $p < 0.0001$, Monkey W: $p = 0.049$). Thus, unlike Todorova et al. (2014), we did not find preserving the noise as separate units to be beneficial for decoding. In fact, the noise removed from most of the later recording sessions (>50 days post implant) was typically harmful for decoding.