Selective neuronal lapses precede human cognitive lapses following sleep deprivation

Yuval Nir1, Thomas Andrillon2–4, Amit Marmelshtein1, Nanthia Suthana5, Chiara Cirelli4, Giulio Tononi4,7 & Itzhak Fried5–7

Sleep deprivation is a major source of morbidity with widespread health effects, including increased risk of hypertension, diabetes, obesity, heart attack, and stroke1. Moreover, sleep deprivation brings about vehicle accidents and medical errors2–4 and is therefore an urgent topic of investigation. During sleep deprivation, homeostatic and circadian processes interact to build up sleep pressure5, which results in slow behavioral performance (cognitive lapses) typically attributed to attentional thalamic and frontoparietal circuits6–14, but the underlying mechanisms remain unclear3,15. Recently, through study of electroencephalograms (EEGs) in humans16,17 and local field potentials (LFPs) in nonhuman primates18 and rodents19 it was found that, during sleep deprivation, regional ‘sleep-like’ slow and theta (slow/theta) waves co-occur with impaired behavioral performance during wakefulness. Here we used intracranial electrodes to record single-neuron activities and LFPs in human neurological patients performing a face/nonface categorization psychomotor vigilance task (PVT)20–24 over multiple experimental sessions, including a session after full-night sleep deprivation. We find that, just before cognitive lapses, the selective spiking responses of individual neurons in the medial temporal lobe (MTL) are attenuated, delayed, and lengthened. These ‘neuronal lapses’ are evident on a trial-by-trial basis when comparing the slowest behavioral PVT reaction times to the fastest. Furthermore, during cognitive lapses, LFPs exhibit a relative local increase in slow/theta activity that is correlated with degraded single-neuron responses and with baseline theta activity. Our results show that cognitive lapses involve local state-dependent changes in neuronal activity already present in the MTL.

To study the neural underpinning of sleep deprivation in the human brain, we investigated the selective responses of individual neurons and how such responses change upon cognitive lapses. Twelve individuals with pharmacologically intractable epilepsy, who were undergoing depth electrode monitoring to identify seizure foci for potential neurosurgical treatment, performed a face/nonface categorization variant of the PVT in 31 experimental sessions (Fig. 1a and Supplementary Table 1). Each session included two 12-min blocks in which six images of famous people, familiar landmarks, and animals were presented (24 trials for each image) for 200 ms with long, unpredictable interstimulus intervals (2–8 s) as participants performed the face/nonface categorization task. In four individuals, pairs of PVT sessions were conducted before and after full-night sleep deprivation that was carried out for clinical purposes (time spent awake after sleep deprivation = 24.1 ± 1.6 h; mean ± s.e.m.), thereby providing a unique opportunity to examine the effects of sleep deprivation on behavior and the underlying activity of individual neurons. We were also able to acquire data for two individuals from four PVT sessions conducted before and after normal sleep to address possible circadian and learning effects. Subjects performed the task successfully and accurately (percentage correct = 94.1 ± 1.9%, no response = 2.9 ± 1.1%; mean ± s.e.m. across 31 sessions). We focused only on correct responses in subsequent analyses to maximize the chances that changes in behavior or neuronal activity in some trials were driven by the internal state of the subjects rather than by an impoverished visual stimulus.

The distribution of behavioral reaction times (RTs) during the PVT experiments was best fit by an ex-Gaussian function, representing a mixture of a normal distribution (standard RTs) and an exponential distribution for slow RTs (‘right tail’) with large variability between sessions and between participants in the predominance of the exponential component (Fig. 1b and Supplementary Fig. 1), as has been previously observed in healthy individuals21. In each experimental session, ‘fast trials’ (fastest RTs) and ‘slow trials’ (slowest RTs) were defined according to the fitted ex-Gaussian distribution (Online Methods), with slow trials comprising 16.9 ± 1.0% of correct responses (mean ± s.e.m., n = 31 sessions). We use the term ‘cognitive lapses’ throughout to refer to these slow trials characterized by delayed behavioral responses (rather than a complete absence of response), as is customary in the sleep deprivation PVT literature20–24. Other strategies for defining cognitive lapses, such as selecting the slowest
5–10% of trials, yielded similar results (data not shown). Figure 1b illustrates how sleep deprivation altered the distribution of RTs. The reciprocal of behavioral RTs (1/RT) is a sensitive marker of slower performance after sleep loss20 and was decreased after sleep deprivation (~21.3%; Fig. 1c). The parameter τ, denoting the exponential decay component of the ex-Gaussian function, increased from 122.9 ± 41.7 ms (mean ± s.d.) to 306 ± 237 ms (+128 ± 79%; Fig. 1c), whereas the parameter μ, denoting the mean of the normal distribution, was only modestly increased (+14.4 ± 8.1%, mean ± s.e.m.), indicating that sleep deprivation exerted its greatest effect on cognitive lapses, as previously reported23. We also observed a ‘time-on-task’ effect23, whereby the frequency of cognitive lapses increased with the time spent performing the task (Supplementary Fig. 2).

Given that sleep deprivation had a marked effect on cognitive lapses, we next examined whether the time spent awake (TSA) before each session could predict the measures of slow trial prevalence across the entire data set, such as elevated τ and elevated mean RT in cognitive lapses. ANOVA analysis using TSA and the number of times a participant had performed the task (training effect) as between-session factors revealed that only TSA significantly predicted slow trials (F = 4.3, P = 0.047; mean RTs for cognitive lapses across sessions, training effect and interaction, P > 0.05) and average RTs (F = 4.4, P = 0.046; other, P > 0.05). None of the variables analyzed could significantly predict mean RTs for fast trials across sessions. Time of day (circadian effect) did not correlate with RTs (fast or slow: all, P > 0.5); however, the lack of significant circadian effects here could stem from variability in the precise hours when sessions were conducted (Supplementary Table 1). In contrast, accuracy of performance in the face/nonface categorization task was not affected by TSA (F = 0.03, P = 0.86). Likewise, a relationship between TSA and slow trials was evident when conducting ANOVA on μ, σ, and τ (the parameters of the fitted ex-Gaussian distribution corresponding to the mean, left tail, and right tail, respectively): a significant effect was found for TSA on τ (F = 4.7, P = 0.04; Fig. 1d), but no effect of TSA was found on μ (F = 0.54, P = 0.47) or σ (F = 0.69, P = 0.41), verifying the specific relationship between TSA and cognitive lapses during slow trials. We did not find an effect of repeated sessions (training effect) on performance, in accordance with the literature23. In addition, subjective sleepiness (Online Methods) was significantly correlated with TSA (Spearman’s r = 0.42, P < 0.05), but ANOVA did not reveal a significant relationship between subjective sleepiness and slow trials (F = 1.5 and 2.2 for mean RTs and τ of the slow trials, respectively), replicating results from previous studies on the limitation of subjective sleepiness estimates25. Altogether, behavioral data showed that TSA was the dominant factor influencing performance on the task and primarily increased the occurrence of cognitive lapses.

Next, we examined the neuronal activity evoked by the stimuli used in the face/nonface categorization PVT task. An image of the intracranial electrodes used and a schematic of the 104 brain regions monitored in the study are provided in Figure 2a,b. The visual stimuli used in the face/nonface categorization PVT paradigm elicited robust responses in individual neurons (Fig. 2c; 1,481 units recorded in total), especially in the MTL, but occasionally also in cingulate cortex, with variability in the precise intensity, selectivity, and latency of responses across individual neurons. Whenever possible, images were selected that elicited responses in prior screening sessions in order to maximize the likelihood of effectively driving activity in the recorded neurons. Of the 611 recorded neurons in the MTL, 106 (17%) responded significantly to at least one stimulus (Online Methods). When pooling the activity of all responsive neurons (n = 162) irrespective of brain region, an average response profile emerged consisting of increased firing rates at 200–500 ms after stimulus onset (Fig. 2d), with an orderly progression of temporal latencies from the high-order visual cortex to the hippocampus and frontal lobe (Fig. 2e). Robust differences in response latencies across MTL regions were also evident when quantifying the precise timing of the responses detected in each trial separately (Online Methods and Supplementary Figs. 3 and 4), in line with previous findings26,27. Single-unit spiking responses were highly selective and could not be observed when averaging the activity of neighboring neurons not categorized as responsive (Supplementary Fig. 5). Crucially, the robust and highly selective profiles of single-neuron responses allowed for study of the effects of cognitive lapses at the single-neuron level.

We examined the relationship between cognitive lapses and underlying neuronal activity by testing how the responses of the same neurons to the same physical stimulus might change as a function of behavioral performance (comparing neuronal activity in fast trials versus slow trials in the same session). Importantly, such ‘within-session’ comparisons minimize confounding by increased epileptogenic activity after sleep deprivation. Given the relationship between cognitive lapses and TSA (Fig. 1d), we compared neuronal responses across all sessions (n = 31): 15.3% of these responses were obtained before and after sleep deprivation, and 71.4% were obtained when subjects were awake for >12 h. In individual neurons, cognitive lapses in slow trials were associated with weaker and delayed neuronal spiking discharges...
relative to responses in fast trials, with differences particularly evident around 200–300 ms following image onset (Fig. 3a). We proceeded to examine the average normalized response in fast versus slow trials across the entire data set. Slow trials were associated with attenuated, delayed, and prolonged responses to identical stimuli (Fig. 3b–d; Supplementary Fig. 6, data for individual subjects; Supplementary Fig. 7a, non-normalized peristimulus time histograms (PSTHs)). A quantitative paired comparison between the responses of each individual neuron in fast and slow trials (Fig. 3e and Online Methods) revealed that the response magnitude was attenuated by 17% in slow trials ($z(376) = -3.05, P = 0.0023$, Wilcoxon signed-rank test). Additionally, in slow trials, response latency (detected in individual trials; Online Methods) was delayed by 27 ± 6.9 ms (mean ± s.e.m.; $z(376) = 3.5, P = 4.8 \times 10^{-4}$, Wilcoxon signed-rank test). Analysis that quantified response latency as firing above baseline in PSTHs yielded similar results (data not shown). Response duration in slow trials was...
increased by 52 ± 19 ms (mean ± s.e.m.; z(376) = 3.2, P = 0.0012, Wilcoxon signed-rank test). Notably, analysis of spiking activity in neighboring nonresponsive neurons during the same trials did not reveal significantly different firing rates between fast and slow trials (P = 0.36, Wilcoxon signed-rank test). Thus, altered neuronal spiking activity during cognitive lapses was specific to responsive neurons and does not reflect a global reduction in activity at those times. Neuronal spiking responses were primarily associated with stimulus onset rather than motor responses (Supplementary Fig. 7). Correlation between the latency of MTL neuronal responses and RTs was also observed across all trials without focusing a priori on comparing fast versus slow trials (Supplementary Fig. 8).

We also examined responses to images using locally referenced LFPs recorded from the same MTL microwires with which single-unit neuronal activity was observed (Fig. 4). The robust increase in broadband LFP gamma power that occurs following sensory stimulation in multiple modalities is an extensively studied phenomenon. This LFP signal is linked to the neuronal spiking activity of local neuronal populations and typically co-occurs with a decrease in low-frequency power, also termed ‘desynchronization’ (refs. 28–31). In line with these findings, the ‘induced power’ LFP response to images (Fig. 4a) consisted of an increase in broadband gamma power (>45 Hz, 50–600 ms after stimulus) and a decrease in slow/theta power (2–10 Hz, 300–700 ms) (Supplementary Fig. 9, examples of the LFP dynamics in single trials). LFP responses were selective: some MTL microwires (n = 270 channels in 31 sessions) showed a robust response (Fig. 4a, MTL responsive channels) whereas other neighboring channels (n = 198 channels in 31 sessions) did not show significant modulations (Fig. 4c, MTL nonresponsive channels), despite the presence of high-quality signals that allowed isolation of neuronal units (Online Methods).

In responsive LFP channels, cognitive lapses during slow trials were associated with a weaker increase in gamma power in comparison to fast trials (Fig. 4b, c: −19.1%, z(270) = 2.72, P = 0.006, Wilcoxon signed-rank test) and a weaker decrease in slow/theta power in comparison to fast trials (Fig. 4b, f: −76.2%, z(270) = −5.2, P = 2 × 10−7, Wilcoxon signed-rank test). In contrast, no significant effects of cognitive lapses were observed in neighboring nonresponsive MTL channels (Fig. 4d–f: gamma: z(198) = −0.57, P = 0.57; theta: z(198) = −0.98, P = 0.33; Wilcoxon signed-rank tests). In contrast to induced power changes, the power of the evoked (average) LFP at 2–10 Hz was lower during cognitive lapses (Supplementary Fig. 10), suggesting that induced power effects reflect changes in ongoing activity rather than changes in the stimulus-evoked event-related potential.

Furthermore, during cognitive lapses, the latency of spiking responses negatively correlated with LFP gamma power (Fig. 4g; r = −0.17, P = 0.006) and positively correlated with LFP slow/theta power (Fig. 4h; r = 0.22, P = 4.5 × 10−4). The significant coupling between the degree of degradation in LFP and neuronal spiking responses suggests that these effects are tightly linked manifestations of neuronal lapses in selective circuits engaged in the task. Whether the cognitive lapses (and underlying neuronal activity) observed after sleep deprivation are qualitatively similar or different from sporadic slow responses occurring throughout wakefulness remains an open question for future studies.

Considering the growing amount of literature on increased theta power (6–10 Hz) as a correlate of sleep pressure,16,19,32 we examined theta power during baseline intervals preceding stimulus onset (Online Methods and Supplementary Fig. 11). First, we established that theta power in MTL LFPs was indeed associated with sleep pressure and cognitive lapses. We found that baseline theta power was (i) significantly correlated with TSA (Supplementary Fig. 11a; r = 0.26, P < 4.07 × 10−6), (ii) elevated after full-night sleep deprivation (Supplementary Fig. 11b; P < 2.74 × 10−5, Wilcoxon signed-rank test), and (iii) higher before cognitive lapse trials (Supplementary Fig. 11c, P < 0.0001, Wilcoxon signed-rank test). Baseline theta power also exhibited a modest albeit highly significant correlation with the level of slow/theta power (2–10 Hz) during the response interval (Supplementary Fig. 11d, e; r = 0.05, P < 4 × 10−3), suggesting that baseline theta activity might influence the degraded LFP response during cognitive lapse trials (Fig. 4). Overall, ongoing theta activity is increased with sleep pressure, and its decreased attenuation during cognitive lapses may lead to impoverished neuronal and cognitive responses.

Finally, we ruled out a potential contribution from pathological epileptiform activity. First, we confirmed that all the main findings (degraded neuronal and LFP responses) held when all data collected in regions eventually declared as being within the seizure-onset zone (SOZ; data not shown) were discarded. Second, we detected interictal spikes (ISs) across the entire LFP data set (n = 1,648 LFP channels) to test whether such events might occur more frequently around cognitive lapses (Online Methods and Supplementary Fig. 12a,b). ISs were detected in few trials (5.0 ± 0.23%) and were significantly more frequently detected within the SOZ than in other regions (Supplementary Fig. 12c; 2.7-fold increase, P < 10−48, Mann–Whitney U-test), attesting to successful IS detection. However, cognitive lapses were not associated with increased frequency of ISs when considering all data (Supplementary Fig. 12d; P = 0.46, Wilcoxon signed-rank test, n = 1,533 channels) or when considering only MTL regions in which selective neuronal effects were observed (Supplementary Fig. 12f; P = 0.48, Wilcoxon signed-rank test, n = 619 channels). In fact, when considering only sessions after complete sleep deprivation, we found a small but significant reduction in the frequency of ISs around cognitive lapses (Supplementary Fig. 12e; P = 0.025, Wilcoxon signed-rank test, n = 186 channels). Thus, we could not find a consistent or robust relationship between ISs and cognitive lapses.

Altogether, these findings show that in sleep-deprived humans engaged in a visual categorization task selective neuronal spiking responses to images are attenuated, delayed, and lengthened before cognitive lapses, and such MTL modulations of spiking activity are associated with a selectively weakened decrease in slow/theta power in responsive LFP channels. Thus, degraded neuronal activity is already evident at the perceptual stage, in which responses of individual neurons in selected trials can predict subsequent cognitive lapses. The extent to which these effects are regionally specific remains unclear, but the current results establish that, within MTL regions, cognitive lapses specifically affect responsive circuits engaged in the task. Progressive delays in neuronal activity may further accumulate in downstream decision-making and/or motor regions during cognitive lapses, ultimately leading to slower behavior. In line with the biased competition model of selective attention13, degraded sensory cortical activity during cognitive lapses may fail to elicit high-quality perceptual representations, and visual information therefore cannot be effectively fed forward to the frontal lobe regions that ultimately determine behavior. It still remains unclear whether degraded MTL activity strictly reflects impaired bottom–up signaling or whether additional top–down attention mechanisms are at play.

Brief periods of silence (OFF periods) accompanied by slow waves in field potentials are hallmarks of non-rapid eye movement (NREM) sleep in both animals34 and humans35, and they are associated with behavioral immobility and unresponsiveness. Following sleep deprivation,
awake rats exhibit local sleep-like slow/theta waves and shorter OFF periods that are associated with degraded behavioral performance\textsuperscript{19}. Given that we could only record a few neurons simultaneously in each brain region and that OFF periods in wakefulness are short (~80 ms), it was not possible to determine reliably whether such brief OFF periods occur in the human brain during sleep deprivation. However, we find that slow/theta activity, previously linked to sleepiness\textsuperscript{16,17,19,32}, is increased before and during cognitive lapses, and these changes were associated with degraded spike responses. Impaired spike responses are observed in individual neurons engaged in a cognitive task without concurrent changes in the firing of neighboring neurons, and these changes predict specific cognitive impairments in sleep-deprived humans. The tight relationship between MTL activity and perception\textsuperscript{36} suggests that visual recognition itself may slow
Cognitive lapses are associated with weaker gamma power increase and weaker slow/theta power decrease in MTL LFPs. (a) Time–frequency decomposition of induced power changes in LFPs of MTL responsive channels (n = 270 channels in 31 sessions). Columns show the average power changes for all trials (left), fast trials (lowest RTs; middle), and slow trials (highest RTs; right). Hot and cold colors mark increases and decreases in power, respectively. Black rectangles highlight stimulus-induced increased power in the gamma frequency range (>45 Hz); pink rectangles highlight stimulus-induced decreased power in the slow/theta frequency range (2–10 Hz). (b) Time course of gamma power increase (top) and slow/theta power decrease (bottom) for fast trials versus slow trials. (c) Decomposition as in a for neighboring MTL nonresponsive channels (n = 198 channels in 31 sessions). (d) Time course as in b for neighboring MTL nonresponsive channels. (e) Quantification (median) of gamma power increases (45–100 Hz, 50–600 ms) for responsive (left) and nonresponsive (right) MTL channels. Asterisks indicate significant differences (Wilcoxon signed-rank tests comparing fast trials with slow trials: ** P < 0.007). (f) Quantification (median) of slow/theta power decrease (2–10 Hz, 300–700 ms) for responsive (left) and nonresponsive (right) MTL channels. Asterisks indicate significant differences (Wilcoxon signed-rank tests comparing fast trials with slow trials: *** P < 1 × 10−7). In e and f, error bars denote s.e.m. computed across LFP channels (n = 270 and 198 for responsive and nonresponsive channels, respectively), and gray dots and lines denote 22 (responsive channels) and 17 (nonresponsive channels) individual sessions that had at least 5 LFP channels each. Green, fast trials; orange, slow trials. (g) Scatterplot of single-neuron response latency versus strength of gamma power increase showing that during slow trials increased latency in spiking responses is significantly correlated with weaker increase in LFP gamma power (Spearman coefficient r = −0.17, P = 0.007, n = 255 pictures that elicited significant responses across 87 units and 21 sessions; Online Methods). (h) Scatterplot of single-neuron response latency versus strength of slow/theta power decrease showing that during slow trials increased latency in spiking responses is significantly correlated with increased slow/theta LFP power (Spearman coefficient: r = 0.22, P = 4.5 × 10−4, n = 255 pictures that elicited significant responses across 87 units and 21 sessions).
Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS
We thank the subjects for their cooperation; M. Tran, E. Behnké, T. Fields, H. Gelbard-Sagiv, and M. Sagiv for assistance with data acquisition; and B. Salas and N. Regev for administrative help. This work was supported by a Human Frontier Science Program (HFSP) Organization long-term fellowship (Y.N.), the Planning and Budgeting Committee of the Israeli Centers of Research Excellence (I-CORE) program and the Israel Science Foundation (grant no. 51/11, Y.N.), the FP7 Marie Curie Career Integration Grant (Y.N.), the Adelis Foundation (Y.N.), Société Française de Recherche et Médecine du Sommeil (SFRMS, T.A.), the Institute of Cognitive Studies of École Normale Supérieure (IEC–ENS) (ANR-10-LABX-0087 and ANR-10-IDEX-0001-02 PSL*, T.A.), National Institute of Mental Health (NIMH) grant no. R01MH099231 (C.C. and G.T.), National Institute of Neurological Disorders and Stroke (NINDS) grant no. P01NS083514 (C.C. and G.T.), National Institute of General Medical Sciences grant no. R01GM16916 (G.T.), and NINDS grant nos. R01NS033221 and R01NS084017 (I.F.).

AUTHOR CONTRIBUTIONS
Y.N., C.C., G.T., and I.F conceived and designed the research, I.F performed surgeries, Y.N. and N.S. collected data, Y.N., T.A., and A.M. analyzed data, and Y.N., T.A., C.C., G.T., and I.F wrote the manuscript. All authors provided ongoing critical review of results and commented on the manuscript.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. in Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem (eds. Colton, H.R. & Altevogt, B.M.) Ch.3 (National Academies Press, 2006).
2. Duffy, J.F., Zitting, K.M. & Czeisler, C.A. The case for addressing operator fatigue. Rev. Hum. Factors Ergon. 10, 29–78 (2015).
3. Goel, N., Rao, H., Durmer, J.S. & Dinges, D.F. Neurocognitive consequences of sleep deprivation. Semin. Neuro. 29, 320–339 (2009).
4. Lyznicki, J.M., Doee, T.C., Davis, R.M. & Williams, M.A. Sleepiness, driving, and motor vehicle crashes. Council on Scientific Affairs, American Medical Association. J. Am. Med. Assoc. 279, 1908–1913 (1998).
5. Borbély, A.A. A two process model of sleep regulation. Hum. Neurobiol. 1, 195–204 (1982).
6. Chee, M.W. et al. Lapses during sleep deprivation is associated with distributed changes in brain activation. J. Neurosci. 28, 5519–5528 (2008).
7. Drummond, S.P. et al. The neural basis of the psychomotor vigilance task. Sleep 28, 1059–1068 (2005).
8. Drummond, S.P. et al. Sleep deprivation–induced reduction in cortical functional response to serial subtraction. Neuroreport 10, 3745–3748 (1999).
9. Padilla, M.L., Wood, R.A., Hale, J.A. & Knight, R.T. Lapses in a prefrontal–extrastriate preparatory attention network predict mistakes. J. Cogn. Neurosci. 18, 1477–1487 (2006).
10. Portas, C.M. et al. A specific role for the thalamus in mediating the interaction of attention and arousal in humans. J. Neurosci. 18, 8979–8989 (1998).
11. Thomas, M. et al. Neural basis of alertness and cognitive performance impairments during sleepiness. I. Effects of 24 h of sleep deprivation on waking human regional brain activity. J. Sleep Res. 9, 335–352 (2000).
12. Tomasi, D. et al. Impairment of attentional networks after 1 night of sleep deprivation. Cereb. Cortex 19, 233–240 (2009).
13. Weissman, D.H., Roberts, K.-C., Visscher, K.M. & Woldorff, M.G. The neural bases of momentary lapses in attention. Nat. Neurosci. 9, 971–978 (2006).
14. Wu, J.C. et al. Frontal lobe metabolic decreases with sleep deprivation not totally reversed by recovery sleep. Neuropsychopharmacology 31, 2785–2792 (2006).
15. Van Dongen, H.P. & Dinges, D.F. Investigating the interaction between the homeostatic and circadian processes of sleep-wake regulation for the prediction of waking neurobehavioural performance. J. Sleep Res. 12, 181–187 (2003).
16. Berréÿri, G. et al. Neural and behavioral correlates of extended training during sleep deprivation in humans: evidence for local, task-specific effects. J. Neurosci. 35, 4487–4500 (2015).
17. Hung, C.S. et al. Local experience-dependent changes in the wake EEG after prolonged wakefulness. Sleep 36, 59–72 (2013).
18. Pigarev, I.N., Notthdurft, H.C. & Kastner, S. Evidence for asynchronous development of sleep in cortical areas. Neuropeport 8, 2557–2560 (1997).
19. Yazovskiy, V.V. et al. Local sleep in awake rats. Nature 472, 443–447 (2011).
20. Basner, M. & Dinges, D.F. Maximizing the sensitivity of the psychomotor vigilance test (PVT) to sleep loss. Sleep 34, 581–591 (2011).
21. Basner, M., Hao, H., Goel, N. & Dinges, D.F. Sleep deprivation and neurobehavioral dynamics. Curr. Opin. Neurobiol. 23, 854–863 (2013).
22. Doran, S.M., Van Dongen, H.P. & Dinges, D.F. Sustained attention performance during sleep deprivation: evidence of state instability. Arch. Ital. Biol. 139, 253–267 (2001).
23. Lim, J. & Dinges, D.F. Sleep deprivation and vigilant attention. Ann. NY Acad. Sci. 1129, 309–322 (2008).
24. Ma, N., Dinges, D.F., Basner, M. & Rao, H. How acute total sleep loss affects the attending brain: a meta-analysis of neuroimaging studies. Sleep 38, 233–240 (2015).
25. Van Dongen, H.P., Maislin, G., Mullington, J.M. & Dinges, D.F. The cumulative cost of additional wakefulness: dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. Sleep 26, 117–126 (2003).
26. Andrillon, T., Nir, Y., Cirelli, C., Tononi, G. & Fried, I. Single-neuron activity and eye movements during human REM sleep and awake vigil. Nat. Commun. 6, 7884 (2015).
27. Mormann, F. et al. Latency and selectivity of single neurons indicate hierarchical processing in the human medial temporal lobe. J. Neurosci. 28, 8865–8872 (2008).
28. Nee, E. et al. Coupling between neuronal firing rate, gamma LFP, and BOLD fMRI is related to interneuronal correlations. Curr. Biol. 17, 1275–1285 (2007).
29. Edwards, E. et al. Comparison of time–frequency responses and the event-related potential to auditory speech stimuli in human cortex. J. Neurophysiol. 102, 377–386 (2009).
30. Fisch, L. et al. Neural “ignition”: enhanced activation linked to perceptual awareness in human ventral stream visual cortex. Neuron 64, 562–574 (2009).
31. Pfurtscheller, G. & Aranibar, A. Event-related cortical desynchronization detected by power measurements of scalp EEG. Clin. Neurophysiol. 42, 817–826 (1977).
32. Finell, I.A., Baumann, H., Borbély, A.A. & Achermann, P. Dual electroencephalogram markers of human sleep homeostasis: correlation between theta activity in waking and slow-wave activity in sleep. Sleep 12, 523–529 (2000).
33. Desimone, R. & Duncan, J. Neural mechanisms of selective visual attention. Annu. Rev. Neurosci. 18, 193–222 (1995).
34. Steriade, M., Timofeev, I. & Grenier, F. Natural waking and sleep states: a view from inside neocortical neurons. J. Neurophysiol. 85, 1969–1985 (2001).
35. Nir, Y. et al. Regional slow waves and spindles in human sleep. Neuron 70, 153–169 (2011).
36. Suthana, N. & Fried, I. Percepts to recollections: insights from single neuron processing in the human medial temporal lobe. Neuron 76, 788–801 (2012).
37. Harris, K.D. & Theile, A. Cortical state and attention. Nat. Rev. Neurosci. 12, 509–523 (2011).
38. Joshi, S., Li, Y., Kalwani, R.M. & Gold, J.I. Relationships between pupil diameter and neuronal activity in the locus coeruleus, cingulate, and cingulate cortex. Neurotherapeutics 8, 221–234 (2016).
39. Wilhelm, B. et al. Daytime variations in central nervous system activation measured by a pupillographic sleepiness test. J. Sleep Res. 10, 1–7 (2001).
40. Parikh, V., Koak, R., Martinez, V. & Sarter, M. Prefrontal acetylcholine release controls cue detection on multiple timescales. Neuron 56, 141–154 (2007).
ONLINE METHODS

Subjects. Twelve individuals (aged 19–52 years, 5 females and 7 males) with pharmacologically intractable epilepsy underwent monitoring with depth electrodes for seizure focus identification and potential surgical treatment\(^\text{35}\). Subjects provided written informed consent before participation in the research study, under the approval of the Medical Institutional Review Board at the University of California, Los Angeles, USA. Electrode location was based only on clinical criteria. For each subject, localization of the SOZ was performed on the basis of recordings during hospital monitoring in combination with prior functional and anatomical neuroimaging. Four individuals participated in a full-overnight sleep deprivation session to increase the propensity for interictal epileptiform discharges and seizures in order to aid clinical diagnosis. During these nights, medical staff verified via video that the subjects remained awake and entered the subject’s room whenever there were signs of sleepiness or eye closure.

Data acquisition. For each individual, 8 to 12 flexible polyurethane depth electrodes were placed in some of the following regions: hippocampus; amygdala; entorhinal cortex; perihippocampal gyrus; anterior fusiform gyrus; temporal gyrus; fusiform gyrus; temporal–parietal–occipital junction; anterior, middle, and posterior cingulate; supplementary motor area; inferior frontal gyrus; orbitofrontal cortex; parietal cortex; posterior temporal cortex; and temporal–parietal cortex. Electrode location varied between patients and was based on clinical profiles (details in Supplementary Table 1). Electrode positions were verified by postimplant computed tomography (CT) co-registered with preimplant magnetic resonance imaging (MRI) imaging using BrainLab stereotactic and localization software (http://www.brainlab.com/)\(^\text{42}\). Each depth electrode terminated in a set of eight insulated 40–μm platinum/iridium microwires\(^\text{41}\) (impedances 200–500 kΩ) protruding from the tip and located 4–5 mm from the most distal microelectrode contact (see Fig. 1d in ref. 35). Microwire signals were simultaneously recorded continuously (Cheetah Recording System (Neuralynx, Tucson, AZ) for 1 subject, Neuropore Recording System (Blackrock, Salt Lake City, UT) for the other 11 subjects), sampled at 28 kHz (the 1 subject) or 30 kHz (the 11 subjects), band-pass filtered in hardware between 1 Hz and 9 kHz, and referenced locally to a ninth noninsulated microwire (LFP signal).

Unit identification and spike sorting. Units were identified using the ‘wave_cluster’ software package\(^\text{37}\) as described previously\(^\text{35}\): (i) extracellular microwire recordings were high-pass filtered above 300 Hz, (ii) a lower threshold at 5 s. d. above the median noise level was computed, and (iii) detected events were clustered using superparamagnetic clustering and were categorized as noise or single- or multiunit clusters. Classification of single- and multiunit clusters was based on the consistency of action potential waveforms and the presence of a refractory period for single units, i.e., less than 1% of interspike intervals (ISIs) within 3 ms, as in ref. 35. Overall, 1,481 units were identified (561 putative single units, 920 multiunit clusters).

Face/nonface categorization PVT paradigm and subjective sleepiness ratings. Subjects participated in 31 sessions of a visual face/nonface PVT paradigm (Fig. 1a). We used pictures of familiar people, familiar landmarks, and animals (instead of detection of the ‘bull’s-eye’ target used in classical PVT) to elicit robust responses in MTLE neurons. When possible, pictures were chosen on the basis of their effectiveness in eliciting responses in the recorded neurons by means of a ‘visual screening’ experiment performed earlier that day\(^\text{44}\). Each session included either one (n = 11) or two (n = 20) 12-min blocks. During each block, four face images and two nonface images (places or animals) were presented on a laptop computer for 200 ms while subjects performed a face/nonface categorization task. Each picture was presented 24 times in a pseudorandom order (total of 144 trials), with long pseudorandomized interstimulus intervals of 2–8 s (uniform distribution), as in classical PVT design\(^\text{45}\). Subjects were instructed to press one of two buttons (for face versus nonface) as quickly as possible. Subjective sleepiness was assessed at the beginning of each experimental session using the Stanford Sleepiness Scale (SSS) and a visual analog rating of sleepiness, which were combined to a single sleepiness score (Supplementary Table 1).

Analysis of behavioral data. The distribution of RTs for trials with a correct answer to the face/nonface categorization task (95% of trials) was fit in each session separately by an ex-Gaussian function\(^\text{45,46}\), representing a mixture of a normal distribution (standard RTs) and an exponential distribution modeling slow trials representing cognitive lapses (Supplementary Fig. 1). Three parameters were used to fit this model to the behavioral data: μ, σ (both describing the Gaussian component), and τ (describing the exponential component). We then quantified the effects of sleep pressure and subjective sleepiness on these parameters in each session.

Cognitive lapses were defined as the trials contained in the exponential part of the distribution. To define the boundary between the Gaussian and exponential parts, we simulated 10,000 values using the Gaussian component (μ and σ) and computed the RT cutoff at 99.9% of the simulated values. RTs above this threshold value were therefore considered as belonging to the exponential tail. In certain cases, the tail of the Gaussian distribution contained a large proportion of all RTs. We limited cognitive lapses to a maximum of 20% of the slowest trials in these cases to avoid highly heterogeneous numbers of cognitive lapses across sessions. Fast trials were defined as the n trials with the lowest RTs, where n was the number of cognitive lapses for the same experimental session. Only trials with correct answers were considered in this analysis.

Time-on-task effects (Supplementary Fig. 2) were evaluated for statistical significance by dividing each session (n = 31) into four equal parts according to the trial order. In each part, we computed the average response time for all trials with correct answers as well as the proportions of fast and slow trials as defined above. To overcome intersession variability and to compare RTs across sessions in this context, each RT (for each quartile and session) was normalized by the value for the first quartile (100%). Supplementary Figure 2 shows the gradual increase in the mean RT and the proportion of slow trials in the course of the face/nonface categorization PVT experimental blocks.

Analysis of epileptiform interictal spikes. To rule out potential contribution of epileptiform IISs around cognitive lapses\(^\text{47}\), we performed automated detection of IISs in 10 of the 12 subjects in whom the SOZ was determined clinically. IIS detection methodology followed our previous work\(^\text{35}\) as follows: continuous LFP data for each microwire were band-pass filtered between 50 and 150 Hz (fourth-order Butterworth filter). The envelope of the band-pass-filtered data was then extracted using the Hilbert transform. We extracted the mean and s.d. of this high-frequency envelope for consecutive intervals, each lasting 10 s. For each of these intervals, epochs longer than 5 ms during which the envelope exceeded 8 s.d. above the mean were marked as an IIS. Examples of IIS detections are available in Supplementary Figure 12. We then marked those trials that were associated with IISs within a window from −2 to 4 s with respect to trial onset for each LFP channel separately.

Analysis of spiking activity. Single- and multiunit clusters were evaluated for their response to each picture separately. Neurons were declared ‘responsive’ for a given image if their mean firing rate at 200–500 ms following stimulus onset was significantly greater than the baseline firing rate (−600 to 0 ms) (P < 0.005, paired t-test). Only neurons with a baseline firing rate higher than 2 spikes per second were included in this analysis to ensure a good estimate of baseline firing. We also repeated the main analyses in the paper using different criteria for responsiveness, including P < 0.05 and P < 0.001, as well as not discarding neurons with a baseline firing rate below 2 spikes per second, and this did not affect the main results. Neurons were declared ‘nonresponsive’ for a given image when the P value was above 0.2.

Automatic response detection and quantification. Response latency and termination were determined by Poisson spike-train analysis along the lines of the procedure described in refs. 26,27. In this procedure, the ISIs of a given unit are processed continuously over a window from −600 to 1,000 ms with respect to stimulus onset and the onset of a spike-train is detected on the basis of its deviation from a Poisson process (i.e., exponential distribution of ISIs), where the unit’s baseline firing rate is used to estimate the distribution. A response was detected whenever a series of short ISIs represented a significant (P < 0.005) deviation from this exponential distribution (examples in Supplementary Fig. 3). Variations of the critical P value defining significant deviations did not significantly alter subsequent results. Only responses with at least three spikes
were further considered. We determined the presence of a response given these criteria for each trial and, whenever a response was present, we defined its onset latency (or response termination) as the time between image onset and the first (last) spike of the response (green and red dots, respectively, in Supplementary Fig. 3). Only latencies within the first 600 ms were considered, and the overall latency of each neuron was taken as the mean across all trials. Response duration was defined in each trial as the difference between the times of response termination and response onset.

We then quantified (Fig. 3e) differences between fast trials and slow trials in terms of (i) response magnitude, defined as the mean firing rate over the window from 200 to 500 ms with respect to trial onset normalized by the baseline activity from −600 to 0 ms for the same trials, (ii) response latencies, and (iii) response duration; the latter two were defined as explained in the previous paragraph. These quantifications were performed separately in responsive neurons (Fig. 3e) and nonresponsive neurons (reported in text). To further compare responsive units with nonresponsive units, we paired each of the conditions (pictures) for which there was a significant response (n = 469 conditions in 162 neurons) with an equal number of data points randomly selected from nonresponsive units recorded simultaneously in the same brain region. Of the 162 responsive units, 142 could be paired in such a manner and are included in Figure 3e, which shows the comparison between fast and slow trials across these 142 units and the stimuli to which they responded (n = 376).

To highlight differences between the responses in fast and slow trials despite neuron-to-neuron variability in response timing and amplitude, response PSTHs were aligned for visualization purposes only (Fig. 3b–d), such that time zero was defined for each stimulus condition as follows. The response PSTH across all trials was computed and smoothed with a Gaussian kernel (σ = 50 ms), and the moments during the rise and fall slopes at which the response reached 0.3 of its peak magnitude were defined as the ‘beginning’ (time zero; vertical lines, Fig. 3b–d) and ‘end’ (sloped lines, Fig. 3b,c), respectively. Furthermore, response amplitude (color scales, Fig. 3b,c; y axis, Fig. 3d) was normalized by dividing the PSTHs of both fast and slow trials by one joint scalar, which is the maximum of the peak firing rate of the two PSTHs. Finally, Figure 3e shows a joint color representation of slow and fast trial PSTHs within each response (stimulus), where hue (green to orange) represents the difference between the normalized firing rate responses in fast and slow trials. Intensity represents the magnitude of the normalized response for fast or slow trials, depending on which was the greatest.

Analysis of local field potential responses. Apart from Supplementary Figures 8 and 12, all analysis of field potentials was restricted to MTL micro-wires. Field potentials obtained from locally referenced microwire LFPs (Fig. 4) were segmented around each picture presentation, and a time–frequency decomposition was performed on the unfiltered data by applying a fast Fourier transform and Hamming window tapering and normalizing the power spectrum to a distribution (sum = 1). When comparing baseline (−2,000 to 0 ms) theta activity (6–10 Hz) to poststimulus (300–700 ms) slow/theta power (2–10 Hz), as seen in Supplementary Figure 11d, power values were calculated using the EEGLAB toolbox with parameters (for example, FFT, tapering, normalization) as described above. Similar results (for example, r = 0.061, P < 10−20 for Supplementary Fig. 11d) were obtained when quantifying the power at 6–10 Hz in the response interval (same frequencies as in baseline). The relationship between baseline theta and slow/theta power in response intervals was further evaluated in individual LFP channels by computing the correlation coefficient (R values in Supplementary Fig. 11e) between baseline and response across all trials and examining the distribution of R values across all LFP channels.

Analysis of baseline theta activity. Locally referenced LFPs in MTL responsive channels (as above) during the prestimulus baseline periods (−2,000 to 0 ms) were segmented, and trials with artifacts (as above) or with ISIs were excluded from further analysis. Baseline theta power (Supplementary Fig. 11a–c) was calculated as the percentage of total power contained within the band at 6–10 Hz by applying a fast Fourier transform and Hamming window tapering and normalizing the power spectrum to a distribution (sum = 1). When comparing baseline (−2,000 to 0 ms) theta activity (6–10 Hz) to poststimulus (300–700 ms) slow/theta power (2–10 Hz), as seen in Supplementary Figure 11e, power values were calculated using the EEGLAB toolbox with parameters (for example, FFT, tapering, normalization) as described above. Similar results (for example, r = 0.061, P < 10−20 for Supplementary Fig. 11d) were obtained when quantifying the power at 6–10 Hz in the response interval (same frequencies as in baseline). The relationship between baseline theta and slow/theta power in response intervals was further evaluated in individual LFP channels by computing the correlation coefficient (R values in Supplementary Fig. 11e) between baseline and response across all trials and examining the distribution of R values across all LFP channels.

Statistics. Error bars in all figures denote s.e.m. (s.e.m. = s.d./(n − 1); n, number of data points) unless otherwise stated. Student’s t-tests were performed after confirming normal distributions via Kolmogorov–Smirnov tests. Nonparametric Mann–Whitney U-tests and Wilcoxon signed-rank tests were used whenever normality was not confirmed. Correlations were examined using the nonparametric Spearman’s rank correlation coefficient or Pearson’s method for normally distributed pairs of variables. To examine the circadian influence (i.e., time of day) on RTs (see main text), we could not use ANOVA (circular variable). Instead, we computed the correlation coefficient between the time of day and RTs using the ‘circ_corr’ function in the Circular Statistics Toolbox for MATLAB.

Preprocessing and data analysis were performed in MATLAB using custom-developed analysis routines and using the publicly available software package EEGLAB.

Data availability. The data that support the findings of this study are available from the corresponding author upon reasonable request. A Life Sciences Reporting Summary is available.

1. Fried, I. et al. Cerebral microdialysis combined with single-neuron and electrophysiological recording in neurosurgical patients. Technical note. J. Neurosurg. 91, 697–705 (1999).
2. Suthana, N.A. et al. Specific responses of human hippocampal neurons are associated with better memory. Proc. Natl. Acad. Sci. USA 112, 10503–10508 (2015).
3. Quiroga, R.Q., Nadadzy, Z. & Ben-Shaul, Y. Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. Neural. Comput. 16, 1661–1687 (2004).
4. Quiroga, R.Q., Reddy, L., Kreiman, G., Koch, C. & Fried, I. Invariant visual representation by single neurons in the human brain. Nature 435, 1102–1107 (2005).
45. Dawson, M.W. Fitting the ex-Gaussian equation to reaction time distributions. *Behav. Res. Methods Instrum. Comput.* **20**, 54–57 (1988).
46. Lacouture, Y. & Cousineau, D. How to use MATLAB to fit the ex-Gaussian and other probability functions to a distribution of response times. *Tutor. Quant. Methods Psychol.* **4**, 35–45 (2008).
47. de Curtis, M. & Avanzini, G. Interictal spikes in focal epileptogenesis. *Prog. Neurobiol.* **63**, 541–567 (2001).
48. Delorme, A. & Makeig, S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **134**, 9–21 (2004).
Corresponding author(s): Yuval Nir

1. Sample size
Describe how sample size was determined.

Rare data from neurosurgical patients undergoing sleep deprivation were collected over 7 years. The main findings are highly significant statistically, and can be observed in data of individual participants.

2. Data exclusions
Describe any data exclusions.

N/A

3. Replication
Describe whether the experimental findings were reliably reproduced.

The main findings were reliably reproduced across individual sessions and subjects as can be seen in dots (Figure 3e, Figure 4e,f) and Supplementary Figure 6.

4. Randomization
Describe how samples/organisms/participants were allocated into experimental groups.

N/A

5. Blinding
Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

N/A

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters
For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

Confirmed

- The exact sample size \( n \) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
- A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- A statement indicating how many times each experiment was replicated
- The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
- A description of any assumptions or corrections, such as an adjustment for multiple comparisons
- The test results (e.g. \( P \) values) given as exact values whenever possible and with confidence intervals noted
- A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
- Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.
Software

7. Software

Describe the software used to analyze the data in this study. Analysis was performed in Matlab using custom-developed analysis routines and using the publicly available software package EEGLAB.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). Nature Methods guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company. N/A

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species). N/A

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used. N/A

b. Describe the method of cell line authentication used. N/A

c. Report whether the cell lines were tested for mycoplasma contamination. N/A

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use. N/A

Animals and human research participants

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study. N/A

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants. Twelve patients (age 19–52 years, 5 females) with pharmacologically intractable epilepsy underwent monitoring with depth electrodes for seizure foci identification and potential surgical treatment. Patients provided written informed consent prior to participation in the research study, under the approval of the Medical Institutional Review Board at the University of California, Los Angeles, USA. Electrode location was based only on clinical criteria, and Dr. Itzhak Fried performed all surgeries. For each subject, localization of the seizure onset zone was based on recordings during hospital monitoring, in combination with prior functional and anatomical neuroimaging.