Augmenting Cognition Through Edge Computing

Mahadev Satyanarayanan, Carnegie Mellon University
Nigel Davies, Lancaster University

Augmented cognition can transform human capabilities, but delivering its benefits in real time will require low-latency wireless access to powerful infrastructure resources by lightweight wearable devices. Edge computing is the only viable approach to meeting these stringent requirements. We explore the symbiotic relationship between augmented cognition and edge computing.

Viewed as autonomous mobile computing systems with built-in sensing, processing, and persistent storage, humans are the result of more than 1 billion years of evolution. Our chances of improving upon nature in a short time (say, 10 years) are negligible if we are bound by the same rules as biological evolution. However, we have a unique opportunity that is not available to nature, namely, to amplify human cognition in real time through low-latency, wireless access to infrastructure resources. These resources can be larger, heavier, more energy hungry, and more heat dissipative than could ever be carried or worn by a human user. Distributed sensing can also offer real-time inputs from vantage points other than the first-person viewpoint of a human. By seamlessly integrating these resources with human perception and cognition, we could achieve a whole that is much greater than the sum of parts.

This vision was first articulated in 2004,[19] but only now have the necessary building blocks reached a level of maturity that they can be viewed as off-the-shelf technologies. These include wearable computers with rich arrays of sensors (such as video cameras, microphones, accelerometers, and gyroscopes) and cognitive algorithms based on deep neural networks (DNNs) for

Digital Object Identifier 10.1109/MC.2019.2911878
Date of publication: 2 July 2019

computer vision, speech recognition, and natural language processing that have now reached near-human levels of accuracy. A further, crucial building block is the ability to wirelessly access cloud-like computing resources at such low end-to-end latency and high bandwidth that we are able to seamlessly integrate them into the "inner loop" of human cognition. This is the essence of edge computing, which is emerging as a new disruptive force.[20,21]

In this article, we share the experience and insights that we have gained, so far, from exploring two distinct ways of augmenting human cognition:

- *Providing just-in-time guidance and error detection for a user who is performing an unfamiliar task.* The prompt detection of errors can be valuable, even during familiar tasks, when the user is working under conditions of fatigue, stress, or cognitive overload. Informally, this is like having "an angel on our shoulder."[18]
- *Amplifying the bandwidth and fidelity of the long-term persistent memory of a human user.* Human memory is notoriously fallible, but contemporary psychology theories suggest that traces captured and displayed using pervasive devices can be employed to both reinforce and attenuate human memories, opening up the possibility of a very wide range of new applications for memory-augmentation devices.[7]

Using our insights from these two styles of augmentation, we seek to lay the foundations for edge-based augmented processing, storage, and retrieval in humans. Our work spans a wide swath of computer science, including operating systems, wireless networks, computer vision, human-computer interaction, augmented reality, data science, and health systems. In contrast to the idea of replacing the human, which is the goal of classic artificial intelligence, our goal is to enhance and extend the capabilities of a human.

WHY EDGE COMPUTING IS ESSENTIAL

Human performance on cognitive tasks is remarkably fast and accurate. For example, face recognition takes between 370 and 620 ms, depending on familiarity.[17] Speech recognition takes 300–450 ms for short phrases and only 4 ms to understand that a sound is a human voice.[1] Virtual-reality applications that use head-tracked systems require latencies lower than 16 ms to achieve perceptual stability.[10] Humans are acutely sensitive to delays in the critical path of interaction. This is apparent to anyone who has used a geosynchronous satellite link for a telephone call. The nearly 500-ms round-trip delay is distracting and leads to frequent conversational errors.

Cognitive augmentation requires sensing to be superhuman in speed, without loss of accuracy. Only then will there be time left within a very tight budget for additional processing to provide augmentation. An end-to-end latency target of a few tens of milliseconds is a safe and conservative goal, with 10 ms as the ideal. Longer delays may distract and annoy a mobile user who is already attention challenged. Since jitter is also annoying and distracting, it is important to avoid long-tailed distributions of end-to-end latency.

The most accurate cognitive algorithms are typically processing and memory intensive. Their execution speed on mobile devices tends to be slow relative to execution on a server. Table 1 illustrates this point with 2018 data from Wang et al.,[23] corroborating 2013 results from Ha et al.[13] and 2016 results from Hu et al.[14] During the 20-year period from 1997 to 2017, mobile devices consistently lagged far behind server hardware, as shown in Table 2. This stubborn performance gap is because mobile users value light weight, small size, long battery life, comfort and aesthetics, and tolerable heat dissipation over speed, memory size, and storage capacity. While mobile devices will improve in the future, so will server hardware, and the gap will remain. One can view this gap as a mobility penalty: It is the price one pays in performance for the benefit of portability.[22]

Wirelessly offloading computing-intensive operations to servers in the infrastructure helps to bridge the gap shown in Table 2. However, using servers in the public cloud is unsatisfactory

TABLE 1. Inference speed during an image-classification task.

	MobileNet (ms)	ResNet (ms)
Nexus 6 smartphone	353 (67)	983 (141)
NVIDIA Jetson TX2	13 (0)	92 (2)
Rack-mounted server	4 (0)	33 (0)

Times are per image, averaged across 100 random images. Numbers in parentheses are standard deviations. Full experimental details can be found in the source material.[23] MobileNet is a DNN optimized for mobile devices. It has a smaller memory footprint and processing demand than ResNet but is less accurate. (Adapted from Figure 3 in Wang et al.[23])

because the cloud is typically far away. The high level of consolidation necessary for economies of scale in cloud computing implies that there can be only a few large data centers worldwide. Li et al.[15] report that the average round-trip time from 260 global vantage points to their optimal Amazon Elastic Compute Cloud instances is nearly 74 ms. A wireless first hop would add to this amount. This makes it impossible to meet tight end-to-end latency goals of just a few tens of milliseconds.

Edge computing creates the illusion of "bringing the cloud closer."[20] Server hardware in edge computing is comparable to that in cloud computing but engineered differently. Instead of extreme consolidation into a few large data centers, servers in edge computing are organized into small, dispersed data centers that we call *cloudlets*. A cloudlet can be viewed as a data center in a box. By offloading to a cloudlet rather than to the cloud, a resource-challenged wearable device can simultaneously meet the goals of low end-to-end latency and resource-intensive processing.[12] This is a crucial capability for augmenting cognition.

The next two sections illustrate how edge computing can be used to enable two different types of augmented cognition. The "Augmenting Task Performance" section describes how the ease and accuracy of task performance can be improved, especially when a user is executing a task for the first time. The "Augmenting Memory" section describes how the notoriously fallible human memory can be made more accurate. We hope that our success in these efforts will stimulate and encourage research on many other forms of augmented cognition.

TABLE 2. The long-term impact of mobility constraints.

	Typical server		Typical mobile device	
Year	**Processor**	**Speed**	**Device**	**Speed**
1997	Pentium II	266 MHz	Palm Pilot	16 MHz
2002	Itanium	1 GHz	Blackberry 5810	133 MHz
2007	Intel Core 2	9.6 GHz (four cores)	Apple iPhone	412 MHz
2011	Intel Xeon X5	32 GHz (2 x 6 cores)	Samsung Galaxy S2	2.4 GHz (four cores)
2013	Intel Xeon E5-2697v2	64 GHz (2 x 12 cores)	Samsung Galaxy S4	6.4 GHz (four cores)
			Google Glass	2.4 GHz (four cores)
2016	Intel Xeon E5-2698v4	88 GHz (2 x 20 cores)	Samsung Galaxy S7	7.5 GHz (four cores)
			HoloLens	4.16 GHz (four cores)
2017	Intel Xeon Gold 6148	96 GHz (2 x 20 cores)	Pixel 2	9.4 GHz (four cores)

Adapted from Chen[3] and Flinn[12]. "Speed" metric = number of cores times per-core clock speed.

AUGMENTING TASK PERFORMANCE

GPS navigation systems have transformed our driving experience. They guide us step by step to our destination, offering us helpful, just-in-time voice guidance for upcoming actions that we need to take. If we make a mistake (for example, miss an exit), the GPS promptly recognizes and corrects it. The difficult task of navigating an unfamiliar city has been transformed into a trivial exercise in following directions.

Wearable cognitive assistance broadens the metaphor of GPS navigation. It can be viewed as an angel on our shoulder that silently observes what we are doing and offers helpful hints just in time. This concept lies at the convergence of wearable computers, edge computing, and cognitive algorithms (such as computer vision, speech recognition, natural language understanding, and other derivatives of machine learning). The wearable device provides a first-person viewpoint of a user's task. Sensor streams from the device (such as video, audio, accelerometer, and gyroscope) are transmitted over a wireless network to a nearby cloudlet for task-specific processing. The cognitive algorithms in this processing typically have memory, CPU, and graphics processing unit demands that cannot be sustained on the wearable device. Based on inferred task state, guidance in visual, verbal, or tactile form is generated, transmitted over the wireless network, and presented to the user through the wearable device.

Gabriel, shown in Figure 1, is an extensible platform-as-a-service layer

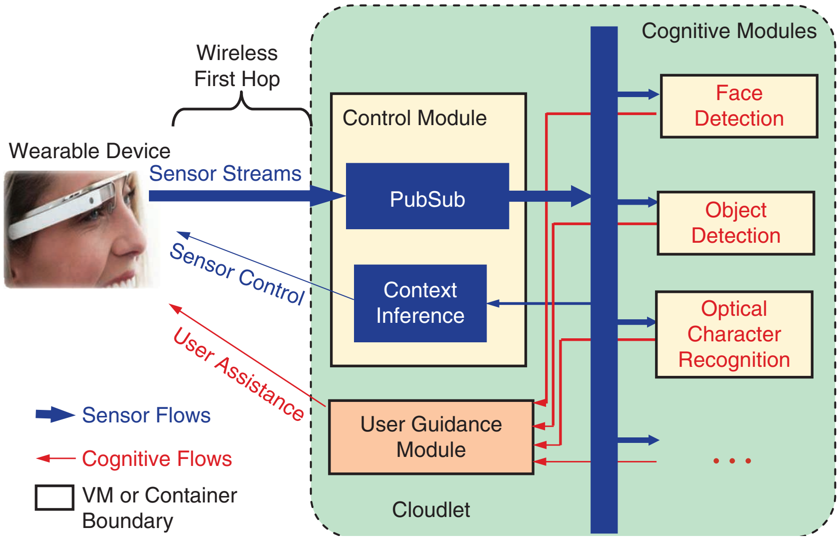


FIGURE 1. The Gabriel architecture. (Adapted from Chen et al.[4])

that we have created for wearable cognitive assistance. The front end on a wearable device performs preprocessing of sensor data (for instance, compression and encoding) and streams it over a wireless network to a cloudlet. The Gabriel back end on the cloudlet is organized as a collection of cognitive modules. The control module is the focal point for all interactions with the wearable device. A publish-subscribe (PubSub) mechanism decodes and distributes the incoming sensor streams to multiple cognitive modules (such as task-specific computer-vision algorithms) for concurrent processing. Cognitive-engine outputs are integrated by a task-specific user guidance module. This code performs higher-level cognitive processing, such as inferring task state and detecting errors. From time to time, it generates guidance.

On this platform, using a diversity of wearable devices such as Google Glass, Microsoft HoloLens, Vuzix Glass, and Osterhout Design Group's ODG R7, we have implemented more than one dozen applications.[3,4,20] Some of these are summarized in Table 3. The implementation of these applications shows considerable high-level similarity in terms of cloudlet workflow. This workflow has two major phases.

The first phase analyzes the current video frame to extract a symbolic representation of the current state of the task. This phase has to be tolerant of considerable real-world variation in the video frame due to variable lighting levels, varying light sources, varying viewer positions, task-unrelated clutter in the background, and so on. The symbolic representation is an idealized representation of the current task state that excludes all irrelevant detail. One can view this phase as a task-specific "analog-to-digital" conversion of an input video frame—the enormous state space of the input is simplified to the much smaller state space of the symbolic representation.

The second phase operates exclusively on the symbolic representation. It compares the symbolic representation to the expected task state to determine whether user guidance is needed and, if so, what that guidance should be. The guidance may have video, static images, plain text, and audio components that are streamed back to the wearable device for presentation to the user. Further details can be found in Chen et al.,[4] which also analyzes the sources of end-to-end latency in this class of applications.

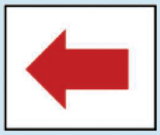
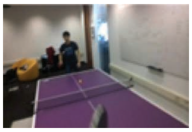
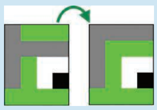
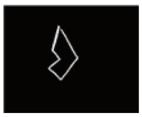
Today, Gabriel applications depend entirely on first-person sensing from a user-worn device. No use is made of additional sensing from off-body viewpoints, which has been shown to be valuable by our memory-augmentation work described in the next section Exploring how such additional sensing could help Gabriel applications is part of our future work.

AUGMENTING MEMORY

In the previous section, we focused on decision making, which is the human equivalent of processing. In this section, we focus on augmenting human memory, the equivalent to upgrading the size, speed, and storage indexing available to a computer. The importance of addressing our current inability to augment human memory cannot be overstated; for example, 47.5 million people worldwide are currently living with dementia. The loss of memory, and with it a sense of identity, is often cited as one of the most distressing aspects of the disease. Even for otherwise healthy individuals, memory augmentation offers the potential to deliver significant benefits in productivity and quality of life.

Technology has always had a direct impact on how and what humans

TABLE 3. An example of wearable cognitive-assistance applications. (Adapted from Satyanarayanan.[20])

Application name	Example input-video frame	Description	Symbolic representation	Example guidance
Pool		It helps a novice pool player aim correctly and gives continuous visual feedback (left arrow, right arrow, and thumbs up) as the user turns the cue stick. The symbolic representation describes the positions of the balls and target pocket and the top and bottom of the cue stick.	Pocket, object ball, cue ball, cue top, cue bottom	
Ping-Pong		It tells a novice to hit the ball to the left or right, depending on which is more likely to beat the opponent. Uses color-, line-, and optical flow-based motion detection to sense the ball, table, and opponent. See the video: https://youtu.be/_lp32sowyUA.	In-rally, ball position, opponent position	Whispers, "left!"
Workout		The app counts out repetitions during physical exercises. Classification is performed by using volumetric template matching on a 10–15-frame video segment. A poorly performed repetition is classified as a distinct type of exercise (for example, a good push-up versus a bad push-up).	Action, count	Says, "eight"
Face		This app jogs a user's memory of a familiar face when the user cannot recall a person's name. It detects and extracts a tightly cropped image of each face and applies a state-of-the-art face recognizer. It whispers the name of the person that is recognized.	ASCII text of a name	Whispers, "Barack Obama"
LEGO		The app guides a user through assembling 2D LEGO models. The symbolic representation is a matrix representing the color of each brick. See the video: https://youtu.be/7L9U-n29abg.	[0, 2, 1, 1], [0, 2, 1, 6], [2, 2, 2, 2]	 Says, "Put a one-by-three green piece on top."
Draw		The app helps a user to sketch better. It builds on a third-party app for desktops. Our implementation preserves the back-end logic. A new Google glass-based front end allows a user to work on any drawing surface and with any instrument. The app displays the error alignment in a sketch on Google Glass. See the video: https://youtu.be/nuQpPtVJC6o.		
Sandwich		This app helps a cooking novice prepare sandwiches according to a recipe. Since real food is perishable, we use a plastic toy in the shape of food. Object detection uses a faster regional-convolutional neural network DNN approach. See the video: https://youtu.be/USak PP45WvM.	Object: lettuce on top of ham and bread	 Says, "Put a piece of bread on the lettuce."

remember. This impact is both inevitable and fundamental—technology radically changes the nature and scale of the cues that we can preserve outside our own memory to trigger recall. We argued, in a previous article,[7] that recent developments in three separate strands of technology together enable entirely new ways of augmenting human memory.

- The near-continuous collection of memory cues has become possible through the use of lifelogging technologies, social networks, and interaction logs.
- Advances in audio and image processing now enable widespread mining of stored cues for proactive presentation.
- The pervasive nature of displays (both mobile and fixed) provides many new opportunities for displaying memory cues to trigger recall.

These building blocks provide the foundation for a new technology ecosystem that can transform the way humans remember to measurably and significantly improve functional capabilities while maintaining individual control. Example applications of memory augmentation include support for learning new skills, affecting behavior change by helping users recall previous positive (or negative) experiences, and helping to address many of the everyday cognitive failures we all experience.[5]

Memory augmentation will obviously make use of mobile devices, such as lifelogging cameras. However, they are not sufficient, and edge computing will be crucial in delivering the sensing, processing, and cuing required for affecting memory augmentation. For example, in terms of sensing, wearable devices (and their associated first-person views) have significant limitations as a platform for augmenting human cognition.[6] Lifelogging cameras are often difficult to place comfortably on the body while still maintaining clear and meaningful coverage of the environment (common problems include occlusion by hair/clothes and poor viewing angle). Moreover, these cameras are typically static and, therefore, capture a poor representation of what was actually seen at the time (Figure 2). Furthermore, psychology literature indicates that, despite seeing the first-person view, individuals may experience detachment from their current perspective, leading them to see things from the view of an onlooker. More significantly, such observer (third-person) views are a not-uncommon feature of recalled memories.[16] The ability to capitalize on sensors in the environment offers a number of advantages for memory-augmentation systems, including access to improved-quality data, professional maintenance of the sensor infrastructure, and more cost-effective solutions since the price of sensing can be shared across multiple users.

Edge resources will also be needed to support the storage and processing needs of memory augmentation systems. It is simply not possible to store and process a lifetime's memories on a mobile device, so cloud and edge support will be required. Similarly, while mobile devices, such as Google Glass, can be used to deliver memory cues, future systems are likely to make use of the full device ecosystem and present information via pervasive displays, audio devices, and environmental control (for example, stimulating recall by recreating the environmental context of a memory).

In our work, we have begun to develop an architecture for memory augmentation (see Figure 3). The architecture highlights the three distinct points of intervention for memory-augmentation systems (encoding, rehearsal, and retrieval), a range of presentation options (spanning both mobile and infrastructure), and examples of sensing systems that provide the raw data from memory cues. For example, we have built systems that capture image data using lifelogging cameras (narrative clips) and process these images to produce summaries that can be presented to users via in-home ambient displays (supporting rehearsal). The same conceptual architecture has also been

FIGURE 2. An example of the same event photographed by a lifelogging device and an infrastructure camera.[6]

used to support a diverse set of applications, including automatic summarization of meetings and delivery of lecture summaries as students walk to lectures, with those applications designed to help users restore context before their next conference or class.

Our architecture is underpinned by the notion of a memory vault in which a user's lifetime of memories resides. The architecture highlights that much of the functionality for memory augmentation is likely to reside at the edge. Extensive use is made of infrastructure sensors, and it is clearly not possible to store the entirety of a user's memories on a single mobile device.

THE ROAD AHEAD

Augmenting cognition using edge computing represents a perfect example of transformative computing, in which existing technologies (for example, lifelogging cameras, head-mounted displays, and image-processing algorithms) are leveraged to provide entirely new applications. However, it is clear that a number of challenges must be overcome before augmented cognition is available at scale.

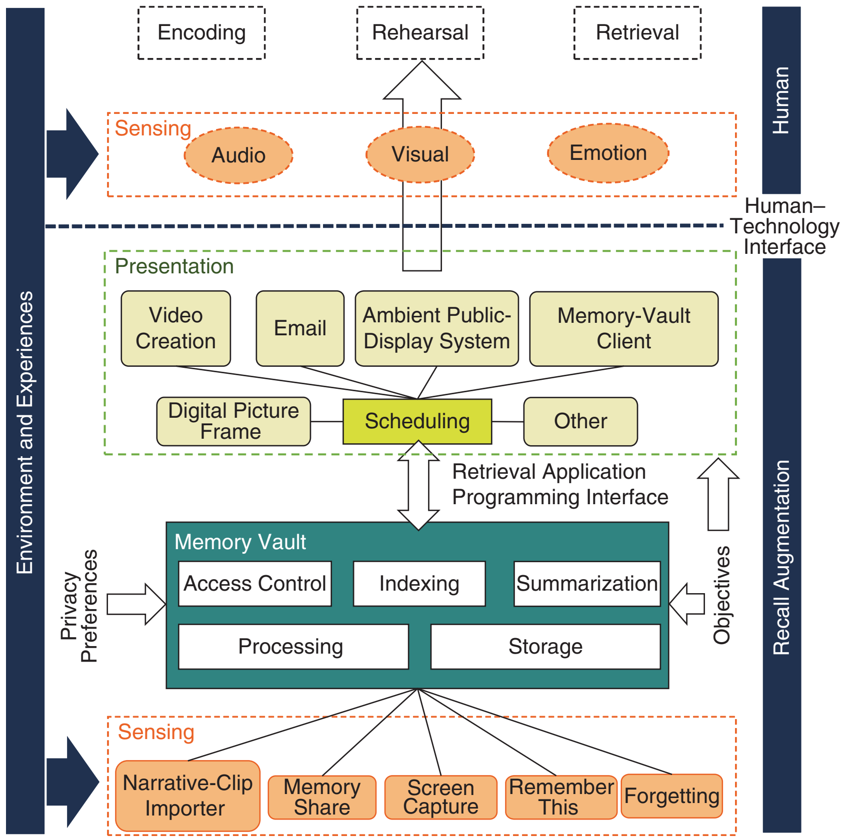


FIGURE 3. The RECALL memory-augmentation architecture.

Achieving widespread deployment of edge computing

Edge computing is key to augmenting cognition. It provides the low-latency processing, storage, and sensing infrastructure essential for this demanding class of applications. Although actual deployments of edge computing are minimal today, there is intense industry interest, and it is believed that we are on the cusp of major industry investments.[9]

Cognitive augmentation applications have the potential to play the role of killer apps for edge computing. Even imperfect implementations of these applications can provide such high value to the end user, without facing any competing alternatives, that they have the potential to create demand for edge computing. These are examples of *edge-native applications*: for example, applications that simply cannot function satisfactorily without edge computing. This is in contrast to *edge-accelerated, cloud-native applications*, where edge support is optional. The 20-year history of content-delivery networks for web access is a good example of edge acceleration. Industry today is focused on identifying new edge-accelerated use cases, rather than edge-native use cases, because the accelerated use cases involve less investment risk. They also involve much less software development, and their markets are much larger since they can function acceptably even in the absence of edge computing.

However, we believe that it is the creation of new edge-native applications that will drive edge computing. The history of technology is replete with examples of rudimentary implementations of killer apps (such as automobiles and aircraft) driving the creation of an ecosystem necessary for advancement and rapidly establishing a virtuous cycle that leads to continuous long-term improvements in both the core technology and the sustaining environment. In computing, there is strong evidence that the development of the spreadsheet, circa 1982–1983, was a major driver in the adoption of personal computing by small businesses. It was the low latency of human interaction (relative to time-sharing) that made PCs indispensable for spreadsheets. The crucial role of low latency in cognitive-augmentation applications suggests that they have the potential to play an analogous role for edge computing.

Unique security and privacy challenges

Augmented memory and decision-making applications raise a number of significant security and privacy concerns that will need to be addressed prior to widespread adoption.

Experience provenance. Traditional experience-capture systems typically use wearable devices that are assumed to be trusted, and the data produced are considered to accurately describe (within the constraints of the technology) the experience of the wearer. As edge computing is used to provide external data streams, this represents an obvious point of attack. For example, if a microphone in a meeting room is used to capture audio associated with a meeting, how do users know (without carrying out a manual review) that the audio captured is, indeed, an accurate reflection of what occurred during the meeting?[7]

Memory and decision manipulation. Both contemporary psychology theories and recent experiments suggest that cued recall can be used to both reinforce and attenuate human memories,[2] with immense security implications. A key challenge is how users can tell if their memories and decisions are being manipulated. In prior work,[7] we have suggested that this will necessitate that users monitor cues that are delivered to them so that they can look for unusual patterns of activity, which is akin to a virus checker for human memories and decision-making processes. Such systems are likely to be computationally demanding and need access to contextual data as ground truth, suggesting that edge support will be required. If a user's memory vault exists at the edge, care needs to be taken to ensure that different users' vaults are appropriately protected by using, for example, techniques borrowed from the field of application sandboxing.

Privacy mediation. The widespread use of augmented-cognition applications that collect substantial data is likely to significantly impact the privacy of bystanders. One possible approach is to attempt to denature data streams before they are processed by augmented-cognition applications. Such denaturing (for instance, face blurring or processing audio to hide speakers' identities) is also likely to be computationally demanding. Earlier work[8] has proposed that denaturing could be performed on cloudlets.

Resolving ethical dilemmas

Augmenting cognition also raises a number of nontechnical challenges. While not necessarily the focus of scientific or engineering research, these are important considerations for the community, as they will significantly impact the use of the technology.

Managing shared memories. While we often think of memories as intensely personal, much of the data that underpins these memories relates to other people. The challenges of designing appropriate security mechanisms increase significantly when sharing memories is considered. For example, in a meeting involving three people, who owns the memories of the event? Is it necessary for each person to keep a copy of the memory and manage the controls to access it? Or is it possible for a single copy to be maintained with appropriate shared ownership? As the various participants choose to delete their copies of the memory, what happens when the last interested party deletes it? Such challenges are compounded when we consider the case of managing memories after death and their use as part of the grieving process.[11]

Avoiding a new digital divide. Traditional technologies for human augmentation, such as glasses and hearing aids, generally aim to provide their users with abilities that approximate the norm. As a result, they raise few ethical challenges. However, if cognitive assistance becomes widely available, it raises important questions of fairness and equality when comparing augmented and nonaugmented humans. Crucially, care will need to be taken to ensure that a new digital divide is not created between those who can afford to augment their capabilities and those who cannot.

When our eyesight fails, we are fitted for glasses. When our hearing fails, we buy a hearing aid. What do we do when our decision making or memory is no longer sufficient for the tasks at hand? The prospect of augmented cognition is truly tantalizing, yet achieving the vision will require leveraging a wide range of technologies to support these transformative applications. We have argued that edge computing will have a key role to play, enabling us to draw on a wide range of environmental sensing and processing resources, while meeting the low-latency demands of cognitive augmentation. ■

ACKNOWLEDGMENTS

The work of Nigel Davies was, in part, funded by the U.K. Engineering and Physical Sciences Research Council under grant EP/N028228/1 (PACTMAN). Davies would like to acknowledge the contributions of members of the RECALL project in shaping many of the ideas presented in this article. The work of Mahadev Satyanarayanan was partly supported by DARPA under contract HR001117C0051 and by the National Science Foundation under grant CNS-1518865. Additional support was provided by Intel, Vodafone, Deutsche Telekom, Verizon, Crown Castle, NTT, VMware, Seagate, and the Conklin Kistler family fund. Satyanarayanan would like to thank his students, staff, and faculty colleagues in the Elijah and Gabriel projects for their many contributions to the work presented here. Any opinions, findings, conclusions, and recommendations expressed in this material are those of the authors and do not necessarily reflect the view(s) of their employers and the funding sources mentioned earlier.

ABOUT THE AUTHORS

MAHADEV SATYANARAYANAN is the Carnegie Group Professor of Computer Science at Carnegie Mellon University, Pittsburgh. Satyanarayanan received a Ph.D. in computer science from Carnegie Mellon. He is a Fellow of the IEEE and ACM. Contact him at satya@cs.cmu.edu.

NIGEL DAVIES is a distinguished professor of computer science at Lancaster University, United Kingdom, and codirector of Lancaster's Data Science Institute. His research is characterized by an experimental approach involving large-scale deployments of novel systems with end users. Contact him at n.a.davies@lancaster.ac.uk.

REFERENCES

1. T. Agus, C. Suied, S. Thorpe, and D. Pressnitzer, "Characteristics of human voice processing," in *Proc. 2010 IEEE Int. Symp. Circuits and Systems (ISCAS)*, Paris, France. doi: 10.1109/ISCAS.2010.5537589.
2. M. C. Anderson, E. L. Bjork, and R. A. Bjork, "Retrieval-induced forgetting: Evidence for a recall-specific mechanism," *Psychonomic Bulletin Rev.*, vol. 7, no. 3, pp. 522–530, 2000. doi: 10.3758/BF03214366.
3. Z. Chen, "An application platform for wearable cognitive assistance," Ph.D. dissertation, Comput. Sci. Dept., Carnegie Mellon Univ., Pittsburgh, PA, 2018.
4. Z. Chen et al., "An empirical study of latency in an emerging class of edge computing applications for wearable cognitive assistance," *in Proc. 2nd Association for Computing Machinery/IEEE Symp. Edge Computing*, Fremont, CA, Oct. 2017. doi: 10.1145/3132211.3134458.
5. S. Clinch and C. Mascolo, "Learning from our mistakes: Identifying opportunities for technology intervention against everyday cognitive failure," *IEEE Pervasive Comput.*, vol. 17, no. 2, pp. 22–33, Apr. 2018. doi: 10.1109/MPRV.2018.022511240.
6. S. Clinch, P. Metzger, and N. Davies, "Lifelogging for 'observer' view memories: An infrastructure approach," in *Proc. 2014 Association Computing Machinery Int. Joint Conf. Pervasive and Ubiquitous Computing: Adjunct publication (UbiComp'14)*, pp. 1397–1404.
7. N. Davies et al., "Security and privacy implications of pervasive memory augmentation," *IEEE Pervasive Comput.*, vol. 14, no. 1, pp. 44–53, Jan. 2015. doi: 10.1109/MPRV.2015.13.
8. N. Davies, N. Taft, M. Satyanarayanan, S. Clinch, and B. Amos, "Privacy mediators: Helping IoT cross the chasm," in *Proc. 17th Int. Workshop Mobile Computing Systems and Applications (HotMobile'16)*, pp. 39–44.
9. Nature Electronics, "Take it to the edge," *Nature Electronics*, vol. 2, no. 1, p. 1, Jan. 2019. doi:10.1038/s41928-019-0203-8.
10. S. R. Ellis, K. Mania, B. D. Adelstein, and M. I. Hill, "Generalizeability of latency detection in a variety of virtual environments," in *Proc. Human Factors and Ergonomics Society Annu. Meeting*, 2004. doi: 10.1177/154193120404802306.
11. S. Ellis Gray and P. Coulton, "Living with the dead: Emergent post-mortem digital curation and creation practices," in *Digital Legacy and*

Interaction, Post-Mortem Issues, C. Maciel and V. Carvalho Pereira, Eds. New York: Springer, 2013, pp. 31–47.
12. J. Flinn, “Cyber foraging: Bridging mobile and cloud computing via opportunistic offload,” in *Synthesis Lectures on Mobile and Pervasive Computing*, M. Satyanarayanan, Ed. San Rafael, CA: Morgan & Claypool Publishers, 2012.
13. K. Ha et al., “The impact of mobile multimedia applications on data center consolidation,” in *Proc. IEEE Int. Conf. Cloud Engineering*, 2013. doi: 10.1109/IC2E.2013.17.
14. W. Hu et al., “Quantifying the impact of edge computing on mobile applications,” in *Proc. 7th Association for Computing Machinery Special Interest Group Operating Systems Asia-Pacific Workshop on Systems (APSys 2016)*, Hong Kong, China, 2016. doi: 10.1145/2967360.2967369.
15. A. Li, X. Yang, S. Kandula, and M. Zhang, “CloudCmp: Comparing public cloud providers,” in *Proc. 10th Annu. Conf. Internet Measurement*, Melbourne, Australia, 2010. doi: 10.1145/1879141.1879143.
16. G. Nigro and U. Neisser, “Point of view in personal memories,” *Cognitive Psychol.*, vol. 15, no. 4, pp. 467–482, Oct. 1983. doi: 10.1016/0010-0285(83)90016-6.
17. M. Ramon, S. Caharel, and B. Rossion, “The speed of recognition of personally familiar faces,” *Perception*, vol. 40, no. 4, pp. 437–439, Jan. 2011.
18. T. Ray, “An angel on your shoulder: Who will build A.I.?” *Barron’s*, Feb. 27, 2018. [Online]. Available: https://www.barrons.com/articles/an-angel-on-your-shoulder-who-will-build-a-i-1519747124
19. M. Satyanarayanan, “Augmenting cognition,” *IEEE Pervasive Comput.*, vol. 3, no. 2, pp. 4–5, Apr.–June 2004. doi: 10.1109/MPRV.2004.1316809.
20. M. Satyanarayanan, “The emergence of edge computing,” *IEEE Comput.*, vol. 50, no. 1, pp. 30–39, Jan. 2017.
21. M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, “The case for VM-based cloudlets in mobile computing,” *IEEE Pervasive Comput.*, vol. 8, no. 4, pp. 14–23, Oct.–Dec. 2009.
22. M. Satyanarayanan, W. Gao, and B. Lucia, “The computing landscape of the 21st century,” in *Proc. 20th Int. Workshop on Mobile Computing Systems and Applications (HotMobile’19)*, Santa Cruz, CA, Feb. 2019. doi: 10.1145/3301293.3302357.
23. J. Wang et al., “Bandwidth-efficient live video analytics for drones via edge computing,” in *Proc. 2018 3rd IEEE/Association for Computing Machinery Symp. Edge Computing (SEC 2018)*. doi: 10.1109/SEC.2018.00019.

Digital Object Identifier 10.1109/MC.2019.2921169