# Conversational Bricks and The Future of Architecture: Will <Stores> Survive as the Epicenter for <Retail> Activity in Society?

Brian Subirana, Nava Haghighi, Richard Cantwell, and Sanjay Sarma
Massachusetts Institute of Technology, Auto-ID Laboratory, Cambridge, USA
Email: subirana@mit.edu, {nava, rcantwel, sesarma}@mit.edu

*Abstract*—**The most advanced IoT technologies enable giving physical objects human-like personalities and allowing humans to "converse" with them in any environment. In this paper, we explore the disruptive potential of adding conversational capabilities to any construction material such as bricks and in particular, explore retail and conversational commerce with the aim of designing physical spaces that can compete with e-commerce. We define conversational architecture as the design of buildings enabling human engagement with objects, physical environments, and virtual entities using conversational speech. In a time where digital content and virtual spaces are becoming more relevant through personalization and anticipation of needs, conversational commerce technologies can create a seamless experience between the digital and the physical, making both experiences more rich, while helping bring the relevance of digital experience to physical. In this paper, we will discuss open areas of research in conversational architecture through examining the enabling technologies and open problems that need to be addressed. We contend Conversational Architecture is a building technology that may ensure long-term sustainability of collective architecture and hope to accelerate research and policy discussions in this new and emerging field.**

*Index Terms*—**Internet of Things, Conversational Architecture, Conversational Commerce, Customer Experience, Building Technology, Conversational Bricks, Talking Bricks, Architecture, Retail**

## I. Introduction

With over 20 major retailers closing their stores in 2017, the retail industry is at a turning point and companies are trying to re-define the value of their brick and mortar stores. The shift to a virtual world and the growth in the digital landscape points at a climate change that requires architects and designers to re-evaluate traditional design concepts and methods and develop a new typology that will keep the customers engaged, providing them with an experience that is unparalleled in the digital world. We feel the issues underlying the survival of physical stores are similar to those related to other collective spaces such as universities, convention centers, concert halls, stadiums, libraries or even restaurants. Will <Stores> survive as the epicenter for <Retail> activity in the society? Will <Universities> survive as the Epicenter for <Research>? Will <Convention Centers> survive as the epicenter for <Conferences>? And can conversational bricks change the future of architecture?

Let's imagine a scenario that will soon be possible with current technologies. It is your anniversary and your virtual assistant reminds you to stop by your partner's favorite store to pick up a gift. You walk in and the store welcomes you and congratulates you on your anniversary, then using your partner's style preferences and knowledge of in-store inventory, the app begins to plan you trip. It takes into consideration the watch your partner admired last time you visited the store, and also the fact that your partner had asked to receive notification when a certain fitness tracker was back in stock. Meanwhile a barcode pops up alerting you that you can receive a 15% off your in-store purchase today and a free coffee from the café that they just opened. You start your journey by looking at the watch, and make your way to the fitness tracker section. Your assistant reminds you that your partner likes to customize their accessories based on their outfit and points that the fitness tracker has a spare band in your partner's favorite color. You decide to get the fitness tracker for both of you when your assistant pulls up your new year's resolution of running every day. At the end of your trip, you stop to get your free coffee. When you get to the register, you tap your phone at the scanner, which loads your customer information, payment details and promotion coupon. On the way home, you ask your assistant to make a reservation at the restaurant you had your first date at 4 years ago, and to reserve the same table. You make a quick stop at the florist to pick up the flowers your assistant had ordered earlier that week, and make your way home to celebrate this special day with your partner.

With the advancements in sensing technologies and artificial intelligence, this vision of the future retail store has become highly feasible in the near future. With the digital revolution, there has been a climate change in the field of retail, and conversational commerce is a

Manuscript received March 13, 2018; revised July 22, 2018.

doi: 10.18178/ijscer.7.3.238-245

disruptive storm that has opened new possibilities. We see the store of the future as the cash-less, wall-less, always open interactive hub, providing personalized experiences and anticipating the needs of those who walk into it. The critical question is how do we seamlessly integrate the new technologies into the building of the future. What will the architecture look like when the stores are open 24/7, when you know exactly who is in the store and at what time, and customers can walk into any store, get what they need and leave without the need for customer service or cashiers?  How is this going to change the definition of privacy and how will the new policies protect the customers? What are the new infrastructures that need to be considered for the large amounts of data that is processed in stores? How can we seamlessly integrate the new hardware components such as cameras and sensors into architectural components? What is the future of architecture when the bricks start talking to you?

## II. Enabling Technologies

In order to better understand conversational commerce, we need to understand the technologies that currently enable it and what needs to be done to enhance the experience. The following section analyzes the state of the art and future directions for research in the enabling technologies for conversational commerce using the “Internet of Things (IoT) & Retail Technologies” framework developed at the MIT AutoID lab (Figure 1). The framework groups technologies into three areas: Information, Life and Matter.

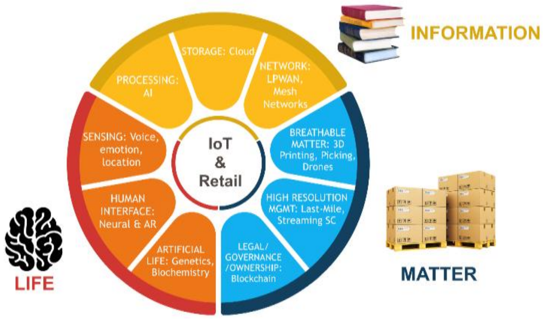
INFORMATION
STORAGE: Cloud
PROCESSING: AI
NETWORK: LPWAN, Mesh Networks
SENSING: Voice, emotion, location
IoT & Retail
BREATHABLE MATTER: 3D Printing, Picking, Drones
HUMAN INTERFACE: Neural & AR
HIGH RESOLUTION MGMT: Last-Mile, Streaming SC
ARTIFICIAL LIFE: Genetics, Biochemistry
LEGAL/ GOVERNANCE /OWNERSHIP: Blockchain
LIFE
MATTER

Figure 1 Top 9 technologies changing IoT and Retail, Preliminary framework developed by Prof. Brian Subirana and Prof. Sanjay Sarma

We envision the building as a conversational commerce ‘buddy’ that recognizes and responds to the shopper, expresses what a shopper would perceive as emotion, presents geographically relevant information, and grows and evolves with the user to make for more satisfying, enjoyable and effective interactions.

In this paper, we will focus on the two categories that will help make buildings come alive; “Sensing” and “Processing”.

### *A. Sensing: Voice*

Imagine having a natural, intuitive, contextual, conversational, voice based interaction with buildings. In order to have a natural human conversation with the building, the building needs to listen (Automatic Speech Recognition), understand and process (Natural Language Understanding), and respond (Text to Speech) in a human way (Figure 2). Deep learning has fueled dramatic improvements along all 3 steps. In this section we will review the current state of the technologies and open areas for improvement.

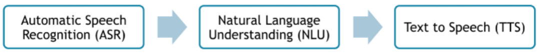
Automatic Speech Recognition (ASR)
Natural Language Understanding (NLU)
Text to Speech (TTS)

Figure 2 3 steps in Conversational Voice-based Interactions

Automatic Speech Recognition (ASR), is the task of converting speech to the words the speaker intended. It is a complex task as there can be many sources of variation in speech such as language, grammar, hesitation, accent, pitch, background noise, etc. Due to these sources of variation, the pronunciation of the same phone or word is never guaranteed to be the same each time, thus its recognition is also bound to be uncertain, consequently ASR is modeled as a probabilistic process [1]. ASR has been an active area of research for over 60 years now. Recent improvements have been driven by the introduction of deep learning. Deep learning techniques were originally applied to acoustic modeling [2] and then expanded to language modeling. In contrast to past steady incremental improvements, deep learning drove dramatic improvements, as measured by the drop in the Word Error Rate (WER) metric. On the industry benchmark Switchboard corpus for instance, WERs that 5 years ago were greater than 15%, have been driven down to 5.9% by Microsoft, which deemed this as “achieving human parity” [3], and 5.5% by IBM [4].

Although ASRs have shown considerable progress, speech recognition is still an open problem. Figure 3 shows where the focus of ASR problems lies today [5] – researchers need to solve for huge vocabularies, free-style tasks, noisy far field speech, spontaneous speech, and mixed languages.

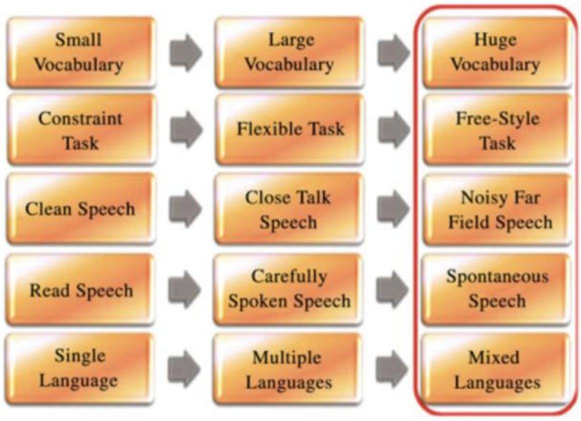

Figure 3 The ASR problems we work on today are much more difficult than what we have worked on in the past due to the demand from the real-world applications.

“Meeting speech” with multiple speakers, overlapping speakers, and far field speech (with noise and reverberation) is also a challenge.

Natural language understanding (NLU) and dialog management (DM) are two essential components of conversational systems. What are they? NLU detects a dialog through parsing speech and filling a semantic template, and DM predicts responsive system actions. “For example, given a user utterance ‘any action movies recommended this weekend?’, NLU predicts intent request_movie and slots genre and date; thereafter, DM predicts system action request_location.” [6]

Deep learning techniques have led to significant improvements in the tasks of domain/intent classification, slot filling and system action prediction. Note that adding a NLU layer requires “knowledge of syntax (the rules governing sentence structure in a given language), semantics (the study of meaning at various levels—words, phrases, sentences and so on), how human dialogues are structured, and access to online resources, including large-scale knowledge bases, from which responses can be crafted.” [7]

A good overview of future directions in this domain is given in a review article [8] that interprets the “evolution of NLP research as the intersection of three overlapping curves – namely Syntactics, Semantics and Pragmatics Curves – which will eventually lead NLP research to evolve into natural language understanding”.

Text to speech (TTS) systems generate speech audio from text received from a response generation component. The main goals for the system are intelligibility and naturalness. There are two main stages in TTS: text analysis, in which text is normalized and converted into a representation of phonemes and prosodic information, and waveform synthesis in which the internal representation is converted to a waveform. [1]

Waveform synthesis has taken two main forms: concatenative, in which parts of pre-recorded speech is diced into individual phonemes and then spliced back together to create new phrases, and parametric in which a computer generates sound – recently this has taken the form of statistical parametric speech synthesis which uses a statistical model to create the proper waveform for each sound [9]. Transitioning from hidden Markov models to deep neural networks has improved the parametric approach. Note that with the concatenative approach one would need to record a whole new database in order to modify the voice (whether to change the speaker or alter emphasis/emotion), which has led to demand for the parametric approach. But the parametric approach was traditionally less superior to the concatenative approach, until in September 2016 Google announced WaveNet, which uses statistical parametric synthesis relying on deep neural networks to produce speech in both English and Mandarin that listeners rated as superior to the best existing systems [10]. However, the WaveNet system is far from real-time and not production ready.

Baidu followed up with Deep Voice [11], a production-quality real-time text-to-speech system constructed entirely from deep neural networks, followed up by Deep Voice 2 which “has a single neural TTS system that can learn hundreds of unique voices from less than half an hour of data per speaker, while achieving high audio quality synthesis and preserving the speaker identities almost perfectly” [12].

Another form of waveform synthesis is articulatory synthesis. This “requires a detailed computer model of how acoustic waves are generated and modified in the human vocal tract. Although theoretically the most flexible type of speech synthesis, articulatory models are complex and not widely used. A notable example is the open-source gnuspeech project.” [7]

There are other areas in TTS that needs to be explored further, specifically as it relates to naturalistic speech with prosody. As an ACM review paper[9] explains:

“Researchers in speech synthesis, however, would like to move beyond merely “intelligible” to speech that sounds more natural. Their work could make synthesized speech easier to understand and more pleasant to hear. It could also allow them to synthesize better voices for people unable to speak for themselves, and create text-to-speech systems for less-common languages. ‘Practically all the systems work well at the sentence level,’ says Alex Acero, senior director of Siri at Apple. Ask a machine to read you a newspaper article or an email message from your mother, however, and the result will be flat. ‘Yes, you can understand it if you pay attention, but it’s still not the same as having someone read it to you,’ he says. Computerized speech cannot handle prosody—the rhythm and intonation of speech that conveys meaning and adds emotional context. ‘That is incredibly important for humans,’ says Acero. 'That’s why when you send text messages, you add emojis.’”

### *B. Sensing: Emotion*

In order for the Conversational Commerce “buddy” to have an accurate understanding of the human state, it needs to understand human emotions and integrate the current emotional state of the user into its decision-making. According to Rosalind Picard who pioneered the field of affective computing “The latest scientific findings indicate that emotions play an essential role in rational decision making, perception, learning, and a variety of other cognitive functions...if we want computers to be genuinely intelligent, to adapt to us, and to interact naturally with us, then they will need the ability to recognize and express emotions, to have emotions, and to have what has come to be called ‘emotional intelligence’.” [13]

“Emotion is a complex psychological state that involves three distinct components: a subjective experience, a physiological response, and a behavioral or expressive response.” [14] Emotion aware Artificial Intelligence has its roots in the field of Affective Computing, which originated in MIT’s Media Lab in the mid-90s. This interdisciplinary field draws from signal processing, machine learning, social and cognitive sciences and more.

The first step to responding to emotions is to be able to detect them. Current approaches can be distinguished by whether they measure physiological responses or behavioral/expressive responses. Physiological data tends to have better correlation with inner feelings, as audio/visual cues can detect the outward expression of emotion but not the inner feeling of it, and furthermore people vary in their expressiveness of emotion [15]. However, measuring physiological data typically requires intrusive on-body sensors, though a recent novel approach describes using only wireless signals that are reflected off people's bodies [15].

Multimodal systems outperform unimodal systems, with unimodal systems acting as the building blocks - in their article "A review of affective computing: From unimodal analysis to multimodal fusion" [16], Poria et al. describe a two-step framework: processing unimodal data separately and fusing them all together using various techniques.

Features extracted from audio/visual/physiological data can be classified using machine learning – the classifiers can take two approaches: categorical or dimensional [16]. Categorical models for emotion assign discrete labels like "happiness", "sadness", "fear" etc. These tend to be limited in terms of describing the complete complex range of emotions that can occur. Dimensional models which represent emotions in a two (e.g. arousal-valence), three (e.g. arousal-valence-dominance) or more dimensional space can capture better the continuum of emotions and are not limited by language labels. An affective representation model called the Hourglass of Emotions goes beyond the limitations of dimensional models by allowing for the modeling of compound emotions and up to four emotions experienced at the same time [17].

To sum up the trends in the field: using multiple modalities, different fusion techniques and deep learning techniques on datasets that in the past tended to be acted data but now are crawled from the internet, has driven significant advances in the field. However, a number of research challenges still remain before emotion enabled devices become pervasive [16]. The number of emotional dimensions to use is unclear. Given that emotion expression tends to be subject-dependent, the generalizability of systems can be difficult to determine. De-noising a continuous stream of data, identifying whether different modalities refer to the same content, effectively modeling temporal information, reducing multimodal data dimensionality to meet performance constraints etc. are all challenges.

### *C. Sensing: Location*

We would like the buildings to guide the user to exactly where they are looking to go, calculating an optimal route for them, providing them relevant location-based offers, or allowing them to trigger requests for assistance or seeing where they are from their exact location. These require an indoor positioning system (IPS) that is accurate, cost effective, scalable and maintainable.

In outdoor environments, Global Positioning System (GPS) technology has been tremendously successful in identifying location and has been adopted widely. Unfortunately in indoor environments "usability of the GPS or equivalent satellite-based location systems is limited, due to the lack of line of sight and attenuation of GPS signals as they cross through walls."[18]

Thus a variety of indoor technologies have come to be adopted – each has its own characteristics and benefits/tradeoffs. Thus deploying each comes to the exact nature of the use case, and depends on factors like size, hardware infrastructure, level of accuracy needed and budget. We would require an indoor positioning system (IPS) that is accurate, cost effective, scalable and maintainable.

According to Brena et al.'s review paper [18], overall, "there is not yet an overall satisfying solution for the Internal Positioning System (IPS) problem. Either very precise solutions are very expensive, or not real time, or cheap proposals are too inaccurate. If we take a standard problem for IPS like locating merchandise in shelves while walking, not a single technology or combination of technologies is both feasible and satisfying." They sum up present and future trends in the field as follows:

Current Trends:

- Reuse Existing Infrastructure in Indoor Environments (e.g., Access Points, Lamps, and Sound Systems) for Location Purposes
- Technology Fusion (Hybrid Positioning Systems): Leveraging the complementarity of several technologies, e.g. Wi-Fi with BlueTooth
- Use of Mobile Devices as an Essential Component of a Positioning System: Leveraging the large number of embedded sensors
- Crowdsourcing: Using open distributed collaboration of many users to build or refine location systems. E.g. currently Wi-Fi fingerprinting requires a thorough mapping activity which can be very expensive. So, if it is possible to automatically construct signal maps by the spontaneous activity of users, that would be a major improvement.

Future Trends and Open Problems:

- Indoor/Outdoor IPS: Outdoor positioning systems will merge with IPS in a seamless way to locate a person with a smartphone anywhere.

Consideration of Privacy and Security Issues in the Development of IPS

### *D. Processing: Artificial Intelligence*

Our vision for a conversational commerce platform includes a cognitive learning element that senses, thinks, anticipates and remembers on a shopper's behalf. In essence, a "shopper genome" that encodes information across a comprehensive set of shopper characteristics. Or a "mini-me" that is learning to be "me", taking over for me, open to me, owned by me, and accessible wherever I am. Our digital avatars will need to leverage Artificial Intelligence (AI) technologies and techniques.

The field of Artificial Intelligence research dates back to the 1950s. Over the subsequent decades interest in the subject has waxed and waned. What distinguishes today's machine learning algorithms from the expert systems of yore is that it yields systems that learn automatically from experience. Rather than relying on a rigid, hand-coded, predefined set of rules that continually runs the risk of obsolescence, today's machine learning algorithms adapt and re-learn automatically as new data points are added, in structured or unstructured format.

The key drivers of machine learning progress include the availability of large datasets with which to train the algorithms (driven by the Big Data trend – everything is increasingly networked and generating large amounts of data), the availability of low cost computational power, and the advancement in machine learning algorithms, the latter being accelerated by the open source trend.

Machine learning algorithms can broadly be divided into three paradigms – supervised learning, unsupervised learning and reinforcement learning [19]. Table I highlights the differences between them [20].

Future directions for AI include evolving it to true human level intelligence, "cognitive" AI tends to be a collection of AI subsystems put together to mimic human cognition (including elements of reasoning, control, learning, memory and adaptability). Observing learning in naturally occurring systems points to further opportunities for evolving AI, e.g. "humans clearly learn many different skills and types of knowledge, from years of diverse training experience, supervised and unsupervised, in a simple-to-more-difficult sequence (e.g., learning to crawl, then walk, then run). Learning systems typically operate in isolation to analyze the given data, people often work in teams to collect and analyze data" [19] Explainable AI will be needed to understand, trust and manage next-gen AI. Ethical and privacy issues will also need to be resolved.

TABLE I. SUPERVISED, UNSUPERVISED, AND REINFORCEMENT LEARNING

| | Description | Retail Use Case |
|---|---|---|
| Supervised Learning | One guides the system by tagging the output. | Analyze products customers buy together: Build a supervised learning model to identify frequent item sets and association rules from transactional data. |
| Unsupervised Learning | The input data is not 'tagged' requiring the system to infer the naturally occurring boundaries or classifications. | Recommend products to customers based on past purchases: Build a collaborative filtering model based on past purchases by "customers like them". |
| Reinforcement Learning | Seen in dynamic systems that can take an action in the real world and measure the outcome to correct its future behavior. | Reduce excess stock with dynamic pricing: Build a dynamic pricing model that adjusts the price based on customer response to offers. |

## III. PRODUCT ARCHITECTURE

We are in Chapter 1 of conversational commerce. Currently conversational commerce technologies are primarily used in virtual assistants. As-Is architectures vary by the virtual assistant provider with the two dominant providers being Amazon (with Alexa) and Google (with Google Home). We envision a future where there will not just be Google Home and Alexa but numerous virtual assistant technologies.

The current approach, i.e. having a plethora of varying architectures, leads to the following challenges:

- For consumers – no standard way to discover and engage with retailers through voice
- Low skill discovery & retention – only 31% of the skills on Alexa have more than 1 review, indicating low usage
- 3rd party skills need to be invoked with rigid formulas
- 3rd party entities must develop an app per virtual assistant technology, and go through a tech company's platform

Our future vision is to democratize this process. There are two possible ways to solve this problem. One is to have multiple assistants, Alexa, Google, Cortana, as well as those representing other entities such as Target, all triggered from one listening/speaker device. The other approach is to have one personal virtual assistant that contacts other assistants on your behalf. The former provides an opportunity for entities to personify in ways they could not have done otherwise. We will identify the opportunities and issues that arise with personification of commercial entities in the next section.

## IV. FUTURE OF ARCHITECTURE AND SPACE

Non-residential architecture has always been a physical manifestation of a virtual entity. However, the physicality of space has limited the interactions of those entities with their users to the constraints of the space. With the development of digital identities, these virtual entities have managed to move beyond the physical boundaries of space. Conversational commerce will bridge the gap between the physical and digital through establishing a voice personality that is independent from physical or digital spaces and can be accessed anywhere and anytime. In this section, we will discuss voice personalities, the implications of conversational commerce on physical spaces, and the boundary-less experience enabled by conversational commerce.

### A. Voice Personalities

For a long time, physical identity has been the key identifier of virtual entities such as brands. With the rise of digital and due to its added convenience, the existence of physical spaces has been threatened [21]. Additionally, much of the digital interaction is mediated through entities such as Amazon which weakens the customer relationship even further. Conversational commerce has the capacity to bridge the gap and strengthen the customer relationship through connecting the digital experience to the physical, while providing a physical experience that anticipates their need and brings a higher degree of affiliation with the brand, and increases customer loyalty (Figure 4).

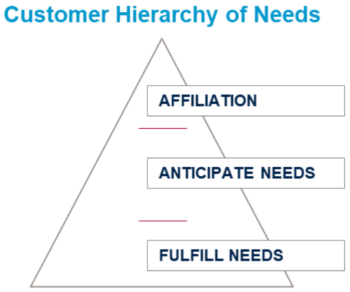

Figure 4 Developed by Capgemini in conjunction with The MIT Initiative on the Digital Economy (IDE)

Conversational Commerce will also provide the opportunity for the virtual entity to re-invent its identity through the medium of voice, and develop its voice personality. This entity will remain the same throughout the digital and physical experiences and therefore plays a big role in shaping the image of the virtual entity. But what is this personality and who has the ability to define it? What is its gender, ethnicity, and age? Does it have an accent? Is it a human, a godly presence, or a machine? Can it be taught culturally appropriate behavior and concepts such as respect, closeness, and personal space? Does it grow and change with its user or is it the same entity for everyone? Does it behave differently in private vs public spaces? Does it interact with children differently from adults? These are some of the open questions that need to be answered in order to create a successful conversational commerce 'buddy'.

### *B. Architecture that Speaks*

Conversational commerce will not only help establish a new aspect of identity for virtual entities, it will also open up new architectural possibilities. For example, the users will have the ability to communicate with building components and the building can now directly offer services to its users. Such abilities will require a larger internal support system and infrastructure to house the additional mechanical and electrical components, while eliminating the need for many back of house programs. The elimination of much of back of house spaces opens up more costumer facing spaces, allowing those spaces to transform to a multi-sensory and interactive experience. An example of this could be the store becoming a place for branded experiences rather than simply shopping in-store [22].

Alternatively, conversational commerce technologies will also change much of the human interface components of buildings such as thermostats and doorknobs, but also check-out lines and way-finding elements, and while those components become obsolete, new technological components will emerge. Lastly due to a new, more personalized approach to commerce, the spaces that traditionally would have a large inventory on hold, will now be able to deploy a more accurate estimate of need, and adjust their supply chain model. As a result, much of the old supply chain components will also become obsolete giving way to re-appropriation of those spaces for more customer facing functions.

### *C. Boundary-less Experience*

The voice personalities, and the new conversational commerce spaces will reveal possibilities for a new type of experience. One that is not limited to a particular space or time and that seamlessly transitions from virtual to physical experiences. This experience will anticipate needs and is shaped around its user. Virtual entities need to rely on a set of core enablers, processes, end points, and capabilities to achieve their desired level of smart digital maturity. Figure 5 depicts Capgemini's point of view on a Smart Digital Store where a given virtual entity connects in-store technologies at the retailer to IoT devices and smart devices (including virtual assistants).

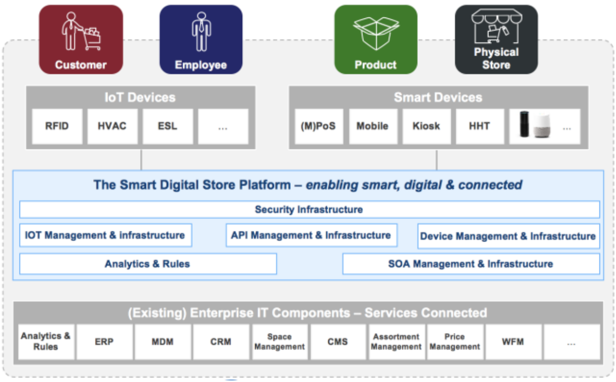
Customer
Employee
Product
Physical Store
IoT Devices
RFID
HVAC
ESL
...
Smart Devices
(M)PoS
Mobile
Kiosk
HHT
...
The Smart Digital Store Platform – *enabling smart, digital & connected*
Security Infrastructure
IOT Management & infrastructure
API Management & Infrastructure
Device Management & Infrastructure
Analytics & Rules
SOA Management & Infrastructure
(Existing) Enterprise IT Components – Services Connected
Analytics & Rules
ERP
MDM
CRM
Space Management
CMS
Assortment Management
Price Management
WFM
...

Figure 5 Source: Capgemini presentation titled "Deep Dive: Retail IoT platform behind the Smart Digital Store"

Imagine walking into the grocery store and telling the walls that you want to make crepes. The assistant recognizes that you saw a chocolate crepe recipe earlier that day on Instagram, and asks if you would like to make that recipe. Then you tell your assistant that your budget for dinner is $5 and that you also have a guest you need to cook for. The assistant lights up the path and the ingredients you need to pick up. It registers that you have picked up those ingredients and deducts it from your account. This experience can be beneficial for both the customer and the provider, however it requires mutual values. Our recent research (conducted collaboratively by Capgemini in conjunction with The MIT Initiative on the Digital Economy (IDE), indicates that value is exchanged through participation and information, where the firm provides access and transparency into the enterprise and the customer provides information. It is at the intersection of these exchanges that mutual value is created (Figure 6)

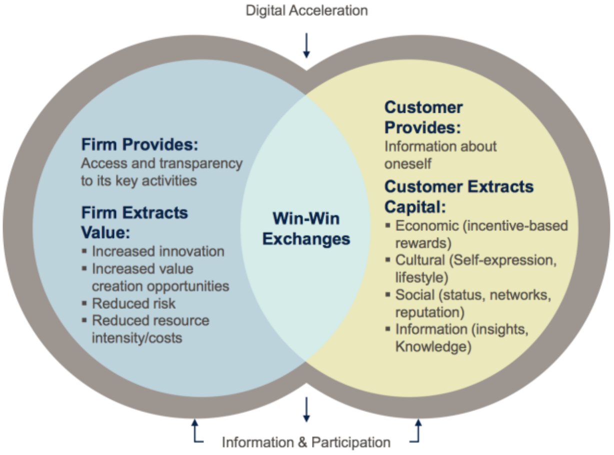
Digital Acceleration
Firm Provides:
Access and transparency to its key activities
Firm Extracts Value:
▪ Increased innovation
▪ Increased value creation opportunities
▪ Reduced risk
▪ Reduced resource intensity/costs
Win-Win Exchanges
Customer Provides:
Information about oneself
Customer Extracts Capital:
▪ Economic (incentive-based rewards)
▪ Cultural (Self-expression, lifestyle)
▪ Social (status, networks, reputation)
▪ Information (insights, Knowledge)
Information & Participation

Figure 6 Source Capgemini/ The MIT Initiative on the Digital Economy (IDE) 2017

Lastly, the virtual entity needs to create a cohesive experience that transitions seamlessly from digital to physical and in between by cohesively designing the conversational commerce, physical spaces, and digital platforms. While conversational commerce provides the opportunity for the seamless integration, it is imperative that the virtual entity invests their resources to ensure all aspects of the experience are well integrated.

## V. Canon

New technologies need new policies that protect the rights of all stakeholders. Voice-based person-to-machine interactions will provide a wealth of new data, and with that, new legal challenges will arise. Biometric identification will also raise a host of ethical questions. “Voice” Net Neutrality as a framework may provide one potential solution. We argue that Net Neutrality can be distilled to 3 core components: Explainability, Fairness, and Configurability

- Explainability entails having explicit standards
- Fairness in this context is the idea of equal access
- Configurability requires that users can adjust settings based on preferences, such as privacy

However, achieving NN in a voice-only environment necessitates decisions, both legal or otherwise, on criteria including but not limited to the following:

- Neutral Navigation
- PII Removed
- Physical Characteristics of the Voice
- Privacy
- “Incognito” Mode for the Voice
- How to Delete “Browsing” Histories
- Data Transferability
- Transparency
- Voice Synthesis with Permission
- Command Standardization
- Psychiatric Diagnosis
- Learned Models
- Awareness
- Bi-Directional Neutrality
- Cookie and Session Transfer Neutrality
- Consumer Protection

Voice Net Neutrality is an extension of the struggle for individual rights. Parallels can and should be drawn between previous civil rights efforts and the future of net neutrality. This means that any robust net neutrality should include ideas of equal access, anti-discrimination, and privacy rights.

While it is debatable which standards are necessary, and what measures are needed legally, the fact remains that net neutrality will likely cease to exist without new standards for the voice. Therefore, a set of standards needs to be developed as a framework to ensure Net Neutrality in the age of voice. Companies cannot prevent disintermediation in Conversational Commerce, but they can seek to control it through forming a coalition of retailers and technological partners to develop CC Standards.

## VI. Conclusion

Historically, paradigm shifts in architecture have been caused by invention and discovery of new materials and technologies. When steel was originally introduced to architecture, it was primarily a direct replacement for masonry facades. Overtime when the capabilities of the material was discovered it started being used as a replacement for masonry structural systems, and ultimately it was embraced to its full potential when it was used in the modern American high-rise. The ability to have a conversation with buildings and products, and the advancements in technology associated with this conversational ability, have opened new possibilities in architecture. Currently this new technology is in its infancy stages where it is used as an add-on or replacement for current technologies. It is the responsibility of the architects, designers, engineers, and the architecture industry as a whole to embrace the conversational commerce technologies to their full potential and develop new typologies that reflect the unique opportunities this medium provides. Through re-imagining the physical space and experiences, we can re-invent the future of retail.

## Acknowledgment

Valuable input has been provided by our sponsors Intel, Cap Gemini, Coles, Target, Via Technologies and OTTO Group.

## References

[1] M. McTear, Z. Callejas, & D. Griol, *The conversational interface: Talking to smart devices*, New York: Springer. 2016.

[2] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, … B. Kingsbury. “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups,” *IEEE Signal Processing Magazine.*, vol. 29(6), pp. 82–97, 2012.

[3] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, … G. Zweig, “Achieving Human Parity in Conversational Speech Recognition,” *Audio Speech and Language Processing IEEE/ACM Transactions,* vol. 25, pp. 2410-2423, 2017.

[4] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, … B. Roomi, “English conversational telephone speech recognition by humans and machines,” 2017.

[5] D. Yu, L. & Deng, *Automatic speech recognition: A deep learning approach*; Springer, 2014.

[6] X. Yang, Y.N. Chen, D. Hakkani-Tür, P. Crook, X. Li, J. Gao, & L. Deng, “End-to-end joint learning of natural language understanding and dialogue manager,” *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference*, 5690–5694, 2017.

[7] C. McLellan, (September 2016). How we learned to talk to computers, and how they learned to answer back. [Online]. Available: http://www.techrepublic.com

[8] E. Cambria, & B. White, “Jumping NLP curves: A review of natural language processing research,” *IEEE Computational Intelligence,* 2014.

[9] N. Savage, “Thinking deeply to make better speech. Communications of the ACM,” 2017.

[10] A.V.D. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, … K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” 2016.

[11] S.O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, … M. Shoeybi, “Deep Voice: Real-time Neural Text-to-Speech,” 2017.

[12] S. Arik, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, … Y. Zhou, “Deep Voice 2: Multi-Speaker Neural Text-to-Speech,” 2017.

[13] R.W. Picard. “Affective computing,” *MIT Press,* 1997.

[14] D.H. Hockenbury, & S.E. Hockenbury, “Discovering psychology,” *Macmillan,* 2010.

[15] M. Zhao, F. Adib, & D. Katabi, “Emotion Recognition using Wireless Signals,” *In Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, pp. 95–108, 2016.

[16] S. Poria, E. Cambria, R. Bajpai, & A. Hussain, “A review of affective computing : From unimodal analysis to multimodal fusion,” *Information Fusion,* 37: 98–125, 2017.

[17] E. Cambria, A. Livingstone, & A. Hussain, "The hourglass of emotions," *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* 7403 LNCS, Pp. 144-157, 2012.

[18] R. Brena, J. Garc á-V ázquez, C. Galv án-Tejada, D. Mu ñoz-Rodriguez, C. Vargas-Rosales, & J. Fangmeyer, "Evolution of Indoor Positioning Technologies: A Survey," *Journal of Sensors,* 2017.

[19] M. Jordan, & T. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science,* 349(6245), pp. 255–260, *2015*.

[20] A. Rao, (November 2015). Demystifying Machine Learning Part 2: Supervised, Unsupervised, and Reinforcement Learning. Next In Tech, Pwc. [Online]. Available: http://usblogs.pwc.com/emerging-technology/demystifying-machine-learning-part-2-supervised-unsupervised-and-reinforcement-learning/.

[21] J. Skorupa, (March 2017). Has the Apparel Retail Bubble Burst?. Retail Info Systems. [Online] Available: https://risnews.com/has-apparel-retail-bubble-burst

[22] B. Quinn, "Buying into the Future," *Frame,* vol. 115, pp. 166-171, 2017.

**Brian Subirana** is the director of the MIT Auto-ID Laboratory. He holds a PhD in Artificial Intelligence (Computer Vision) from MIT CSAIL, an MBA from MIT Sloan and has over 200 publications. Prior to becoming an academic he was with The Boston Consulting Group. He currently researches applications of disruptive IoT/AI technologies focusing in five industries: digital learning, electric vehicles, retail, supply chain, and cryptocurrencies/blockchain. He is particularly interested in inventing business strategies and technical platforms that optimize value from IoT/AI technologies while benefiting from convergence between the five industries above.

**Nava Haghighi** is a researcher and graduate fellow at the Integrated Design and Management (IDM) program at MIT where she focuses on human-centered design and innovation. Her research interests are in developing new experiences enabled by technological advancements, as well as human-computer interaction and how machines can exist as a natural extension of humans. She holds a bachelor of architecture from California Polytechnic University, San Luis Obispo, where she focused on material innovation and fabrication technologies. As part of her undergraduate thesis, she explored physical spaces as a medium for facilitating unique, interactive, ad-hoc experiences and developed a new architectural typology the occupiable product. She has worked as an experience designer and design consultant working with Tesla, T-Mobile, Lexus, and Nissan to understand the future of retail and redesign customer experiences based on new technologies.

**Richard Cantwell** is an MIT Affiliate at the School of Engineering and Auto-ID Laboratory at MIT. His focus is creating business value in the networked digital economy. As Vice President - Cisco Systems, he led strategy and innovation consulting at the intersection of business process and technology. Prior to that, Cantwell directed consumer marketing, advertising, new product development and business transformation at Procter & Gamble, Gillette and Johnson & Johnson. He led P&G/Gillette's pioneering involvement in RFID and helped establish the Auto-ID Center at MIT. He served as Chairman of its Board of Overseers and then Chairman of the Board of Governors for GS1 EPCglobal, developing many of the key global industry RFID standards for product identification, supply chain visibility and electronic data exchange. He received the New England Business and Technology Association Award for Innovation, appeared on Computerworld's list of the Most Powerful People in Networking, was selected one of the Top 25 Consumer Product Visionaries by Consumer Goods Technology, has been named one of Advertising Age's Top 100 Marketers and was a Cannes Advertising Award winner. Cantwell is a graduate of Harvard University and received his master's degree in business administration from Dartmouth College.

**Sanjay Sarma** is MIT's Vice President for Open Learning and the Fred Fort Flowers and Daniel Fort Flowers Professor of Mechanical Engineering. Dr. Sarma founded the Auto-ID Center at MIT (now MIT Auto-ID Labs) and developed many of the key technologies behind the EPC suite of RFID standards now used worldwide. He was also the the founder and CTO of OATSystems, which was acquired by Checkpoint Systems (NYSE: CKP) in 2008. He sits on the boards of GS1, EPCglobal and several startup companies. Dr. Sarma received his Bachelors from the Indian Institute of Technology, his Masters from Carnegie Mellon University and his PhD from the University of California at Berkeley. In between degrees, Sarma worked at Schlumberger Oilfield Services in Aberdeen, UK, and at the Lawrence Berkeley Laboratories in Berkeley, California. He has authored over 50 academic papers in computational geometry, sensing, RFID, automation and CAD, and is the recipient of numerous awards for teaching and research including the MacVicar Fellowship, the Business Week eBiz.