How AI-enabled interactive agents will reshape our workforce of today and tomorrow.

by Archan Misra

The era of human and Artificial Intelligence (AI) co-working has arrived. While we may still be in the early days of such co-working, it is time for organisations to invest in such capabilities and plan for how they will progressively become more pervasive in our lives. The market value for such AI-driven innovation is expected to grow at a dramatic rate of 36.5 percent CAGR (Compounded Annual Growth Rate) in less than a decade to reach US$88 billion by 2032.[1] The recent breakthroughs in AI, such as generative AI (also known as GenAI), are creating dramatic new capabilities for machine-based reasoning and interaction that go well beyond the reading and production of text which has captured the popular imagination.

In this article, I discuss how the ongoing advances in technology will provide what I describe as 'cognitive augmentation' of human activities. Embedding advances—in capabilities such as physical environment sense-making and natural two-way dialogue—into software agents deployed on a variety of personal and Internet-of-Things (IoT) devices will help us avail of such cognitive augmentation via more natural conversational interactions. In time, we may all have such personal JARVIS-es[2], agents which have been designed for specific tasks and environments, thus transforming us into Iron Men and Women!

This JARVIS-like capability enables AI agents to behave more like a partner (such as a helper or co-worker) who can ably support routine activities in various contexts. They are no longer just chatbots, but are able to do a lot more. I will illustrate such possibilities with a couple of current use cases and some that may be realised in the not-too-distant-future. In the last section of my article, I will discuss some key considerations that technology leaders need to keep in mind over the next few years to harness this promise of embedding interactive agents as a natural partner to humans.

THE CURRENT STATE OF TECHNOLOGY

Imagine that you want to plan a trip... An agent will know what time of year you'll be travelling and, based on its knowledge about whether you always try a new destination or like to return to the same place repeatedly, it will be able to suggest locations. When asked, it will recommend things to do based on your interests and propensity for adventure, and it will book reservations at the types of restaurants you would enjoy.[3]

The excerpt above comes from Bill Gates. He is among many prominent industry luminaries who have argued that agents embedded with such AI-driven reasoning and proactive interventional capabilities will drive dramatic, disruptive changes in how we use and interact with technology in our daily life. In parallel, mobile, wearable, and IoT devices are increasingly being equipped with more sophisticated (and cheaper) sensors that can make sense of the device's physical environment (such as the distance, dimensions, and colour of surrounding objects). For example, the Apple iPhone from the 12 Pro models onwards and Microsoft's HoloLens augmented reality (AR) holographic device already come with sensors that can map out their physical surroundings in real time and 3D (three-dimensional [form]).

With these twin advances in sensors and AI, it is understandable that we anticipate the emergence of newer forms of 'situated agents'–that is, agents that run continuously in the background, and keep themselves updated about the state of the surrounding spatial environment without being instructed to do so. Such agents can effortlessly comprehend and respond to, and eventually even *anticipate*, instructions or queries that require the combination of both the sensing capabilities of smart devices and the reasoning capabilities of AI models. In other words, these agents are able to respond to queries that require a deep and sophisticated spatial and semantic understanding of both the physical environment (perceived through multiple sensing modes) and the digital knowledge embedded in cyberspace.

Knowing how to relate what we see in our physical spaces (such as checking the price of a pair of jogging shoes displayed in a sportswear store) to what we know from the Internet (comparing prices of the same item on several online shopping platforms) may seem to be easy tasks for humans, but it is non-trivial from a computing point of view. As an illustration, current AI virtual assistants such as Amazon's Alexa and Apple's Siri are purely voice-based, *virtual* agents, i.e., they read and respond only to voice commands and process only online information. As a result, using speech to ask Siri to play the most popular song today is child's play because it can easily retrieve such information from the Internet. However, when you ask Alexa to play the most popular song by the artist featured on the cover of the *Rolling Stone* magazine that sits on your coffee table, the virtual assistant is stymied as it does not know *where*, and to *what*, you are pointing at.

Such virtual assistants are currently neither equipped with visual perception to scan our physical environment nor can they capture our pointing gestures. These agents do not have what the computing fraternity calls 'situated awareness'–the ability to make sense of our physical surroundings. But with the right sensory inputs and software (including AI models, such as Vision-Language Models or VLMs that interpret text and image data), we can now introduce situated awareness capabilities to the likes of Alexa and Siri. In fact, several technological advances in AI research today are about improving the ability of devices to make sense of these situated cues, especially visual and gestural ones, such as that of pointing to an object. Computer scientists refer to this ability of AI models to align visual and language cues as 'visual grounding'.[4]

However, there is one catch. Many such AI models, including OpenAI's now-famous GPT-4, require network or Internet access as a lot of the actual heavyweight AI computation is done remotely, on a GPU (Graphic Processing Unit)-rich server farm. For natural interaction with situated agents, we however desire to have most, if not all, of this intensive work done on what we call 'pervasive devices', whether they are mobile phones, smartglasses, or even robots. In other words, the processing and response generation should be done 'locally'. This is because the moment you need to run anything on the cloud, you invariably lose interactivity because it will take time to complete. This 'latency' or the time lag between a request and its response[5] is highly perceptible to us, even in hundreds of milliseconds, as it is just how we humans rapidly perceive and respond to our environment. Such response speed may also be critical if we need to count on such processing to be the basis for time-sensitive actions, like when we instruct the agent controlling our electric bicycle to "swerve around that slick spill on the sidewalk".

Building optimised systems that enable situated awareness locally and swiftly is extremely demanding and requires significant research breakthroughs. There are two main challenges to overcome. First, some of the sensors that are responsible for reading the visual (e.g., differentiating colours) and spatial cues (e.g., depth and distance of objects from the sensors) are extremely power-hungry. For example, the power that a LIDAR (light detection and ranging) sensor consumes is nearly 800 times that of a microphone sensor, thus making it less suitable for continuous background sensing that may be necessary under some operational contexts. Second, newer AI models (including GPT-4 or another OpenAI tool, DALL.E 3) are computationally too complex[6] and large for devices such as smartphones or smartglasses to handle.

RESEARCH ADVANCES AND CHALLENGES

The research that my collaborators and I have been working on helps to address some of these challenges,[7] such as developing 'lightweight' AI models that perform the bulk of the more complex processing locally on the pervasive device. For example, we have developed newer and optimised AI models, amenable to local execution, that are not only able to interpret verbal commands and visual cues, but also factor in human gestures such as pointing. Such an advance allows the AI to become more efficient at following visual, voice, and gestural cues from the sensors with more precise prompts (or what we call 'resolving questions') to zoom in on the object of interest, say to focus on *Rolling Stone* when there are perhaps multiple magazines on the coffee table from the previous example.[8] Notably, these models are able to deal with the fact that the act of pointing itself is imprecise, with our pointing error higher when referring to more distant objects.

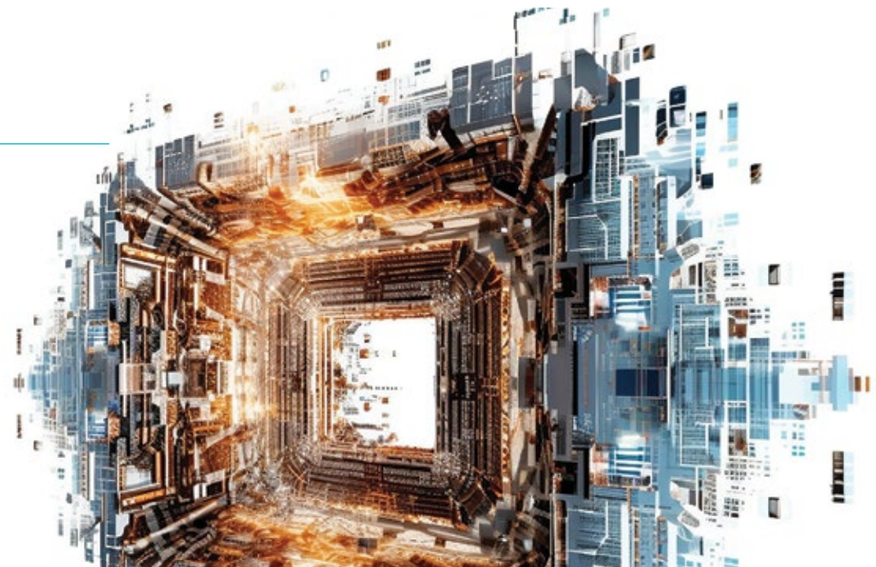
Second, we have also developed techniques that automatically skip over or approximate some of the complex stages the AI model has to complete,[9] especially for visual reasoning, when the verbal instructions suggest that it may be prudent to do so. As an intuitive example, a query like "what is the price of the object next to the laptop?" implicitly provides the hint that the visual processing can be restricted principally to finding 'medium-scale' objects such as the laptop instead of 'smaller-scale' objects (e.g., pens) or 'larger-scale' objects (e.g., cabinets).

There are, of course, still many additional challenges to overcome. Among them, agents need to perform more efficient 'video grounding'–i.e., the ability to 'interpret' a video segment rather than a single image. This will help an agent deconstruct the motion semantics needed to answer questions such as "where can I find the bag that is being carried by the lady who just walked into aisle 4?". That said, rapid advances in such combined vision-language AI reasoning capabilities suggest that interactive situated agents may begin to find their groove as quickly as in the next three years.

HUMAN-AI AGENT COLLABORATION SCENARIOS

As we work at resolving these challenges, the ways in which humans and AI agents are able to work together can only be limited by our imagination. I describe an assortment of scenarios below to demonstrate their potential. These different forms of collaborative task executions can be structured around two dimensions: the expert-novice relationship, and the physical form in which such interactive agents will be embodied.

Building optimised systems that enable situated awareness locally and swiftly is extremely demanding and requires significant research breakthroughs.



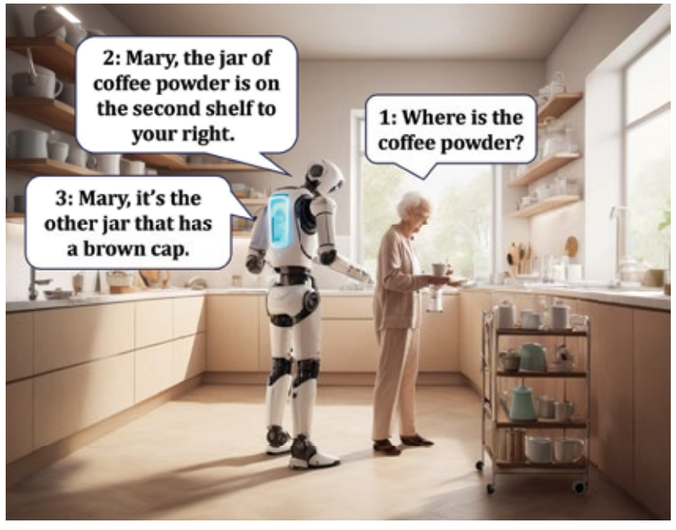


FIGURE 1
Note: Image was generated using Fooocus from the main prompt "elderly in a kitchen assisted by a robot". It is based on the Stable Diffusion model, Juggernaut XL Version 6.

Agents as experts versus novices

In many scenarios, the interactive agent serves as the assistant and is tasked with comprehending instructions provided by an 'expert' human. Conversely, in the AI agent as personal helper scenario below, the agent is the 'expert' interpreting the combination of visual, verbal, and gestural inputs to make sense of an individual's query and provide cognitive assistance.

AI agent as personal helper

Imagine there is an elderly woman, Mary, who has dementia and is experiencing cognitive impairment (refer to Figure 1). She would like to make a cup of coffee in the kitchen. Unsure of her surroundings, she asks, "Where is the coffee powder?" Her robotic assistant, which comes equipped with microphones and cameras, as well as suitable AI-based perception models, would be able to make sense of her request. "Mary, the jar of coffee powder is on the second shelf to your right," it says. When Mary picks the wrong jar, it would be able to proactively correct her with helpful attribute clues such as "Mary, it's the other jar that has a brown cap". Later, the robot may even advise her on how much water to pour into the kettle, among other tasks. The robot, with the right sensors and AI model, is able to provide cognitive augmentation to support Mary with her daily routines. Central to this scenario is the service robot's ability to support such cognitive augmentation via natural interactive, bi-directional conversations, both comprehending an individual's queries and instructions, and generating situated responses.

Embodiment of an agent

The agent itself can be embedded in various devices such as smartglasses and AR headsets worn by a worker. Alternatively, the agent can be embedded in an external embodied form, such as the robot in the personal helper scenario, or the tyre change scenario described below. This capability is especially timely as we are just beginning to transition from expensive industrial robots being used for repetitive mechanical tasks to ones that perform more common 'household' tasks, such as Abluo[10], the new restroom cleaning robot in Singapore, and Optimus, Tesla's newly announced laundry-folding robot[11]. However, such robots still operate under tightly-defined constraints and lack the integration with the interactive agents mentioned here that will allow them to exhibit greater flexibility by incorporating explicit human instructions.

AI agent-enhanced robotic tyre change

One potential use case, which integrates such agent-based interactive support with robotic manipulation, and where the human serves as the expert, may involve maintenance and aircraft repairs at an airport (refer to Figure 2). Aircraft tyres are subject to significant stress and may need to be changed after about every 120 landings, often while the aircraft is being prepared for the next flight at an airport gate. The current process of a three-person engineering crew changing and replacing one aircraft tyre (out of 14 for a typical Boeing 777), weighing as much as 250kg, may now need only one individual and one or more robots. Eventually, a robot could nimbly dismount and re-install a tyre once it is equipped with the right array of sensors, software (including the AI models), and mechanical actuators.

The engineer could then serve as an overall supervisor of the tyre change operation, and an agent embedded in the robotic platform could comprehend the instructions from the engineer and adjust its position to carry out specific operations, such as rotating the wheel or tightening certain bolts. It is important to note that this scenario serves as an example of labour *augmentation* rather than replacement–the use of robots capable of understanding human instructions serves to ramp up the operational tempo and scale of airport operations, and relieves humans from physically arduous tasks, rather than eliminating them.

WHAT ALL THIS MEANS FOR SENIOR TECHNOLOGY EXECUTIVES

While technological advances often happen faster than we can predict, their integration into real-world business processes and operations is often much more complex and thus slower. To take advantage of this emerging world of situated agents, industry professionals and senior technology executives need to be cognisant of several key principles.

Data, data, data

Every work environment in which such advanced technology platforms are situated is complex and differentiated, and proper agent functioning requires appropriately tailoring and optimising the underlying AI models. As a result, collecting the right training data from the operational environment is critical for the AI models to work effectively. For example, while both ostensibly fall into the category of 'field worker assistance', offshore workers on marine platforms will likely face operating conditions and equipment that are quite distinct from those encountered by workers despatched to repair utility equipment in a city.

It is thus important for senior technology executives to start planning now and provision their workforce to proactively begin collecting such 'situated data'. Workers could be equipped with body-worn cameras or smartglasses, even if such devices are presently unable to support such bi-directional situated interaction. Having this corpus of real-world, field data will be a critical asset in the rapid deployment of such interactive agents once the technologies mature. For the case of real-time guidance to field workers performing maintenance and repairs of industrial equipment, such data from experts will provide a baseline and thus help identify anomalies or mistakes that non-expert technicians may make during task execution.

Prioritising use cases

While such interactive agents can usher in powerful new forms of human-machine collaborative working, it is important to recognise that agents will develop only incrementally. To justify initial investments on such situated agents, senior executives need to carefully analyse their

FIGURE 2
Note: Image was generated using Fooocus from the main prompt "worker doing aircraft wheel change operation". It is based on the Stable Diffusion model, Juggernaut XL Version 6.

operations, and identify the key processes and operations that are likely to demonstrate early benefits. I believe that initial gains will come from processes that exhibit the following three characteristics.

1. Controlled, uncluttered physical environments

AI models capable of fusing verbal and visual cues presently perform well under relatively 'benign' operating conditions, including relatively little visual clutter and well-lit operating conditions. Accordingly, initial uses of such agents are likely to be more effective in relatively well-organised spaces, such as a kitchen, rather than very complex environments such as an industrial construction site.

2. Infrequent and slow robotic manipulation

While seamless comprehension of human commands can enhance a robot's efficacy in many environments, physical robots currently perform manipulation tasks much more slowly than humans (partly out of safety constraints). Accordingly, successful interactive human-robot co-working is likely to be initially confined to low-volume, physically complex, and somewhat latency-tolerant tasks (such as the tyre change scenario mentioned above) as opposed to high-volume, high-frequency, and time-critical tasks (such as rapidly sorting and packing luggage onto an aircraft).

3. Sensing data is generated by infrastructural sensors

While several examples of situated agents cite the use of AR smartglasses, such devices are often still too bulky and uncomfortable for continuous use, especially in challenging field environments like the hot and humid conditions on an airport tarmac or a marine platform. Accordingly, initial deployments of situated agents may utilise third-party equipment–i.e., sensing capabilities that are embedded in non-wearable devices, such as smartphones and robots.

Digitalisation

For the successful integration of such agent technologies for human-machine co-working, it is important to view such human-machine teams not as atomic units, but as elements of a broader transformation of enterprise business processes and workflows. Many existing workflows and organisational structures (e.g., reporting hierarchies) are designed for human-only teams and may need to be reimagined for a future where teams consist of humans, AI-powered agents, and machines. A field worker on a marine platform may require the streaming of audio-visual assistance from a remote human expert, but in due course, the assistance may be provided by one of many AI assistive agents, each customised to tackle a specific scenario. Moreover, exception handling, for cases that cannot be handled by an AI agent, may also require routing to a selective group of experts, each of whom specialises in handling specific tasks and challenges, and involves additional employees in newly created roles to perform continuous outcome monitoring and auditing to proactively identify gaps in agent capabilities.

Reskilling and retraining

While putting in place the right digital infrastructure is critical, we cannot ignore the other equally, if not more, important element of the equation–humans. For industrial operations, senior technology executives need to be serious about training and reskilling managers, in addition to the rank-and-file workers. As a small but significant example, in spite of undoubted advances, the tyre-changing robot may be able to reliably comprehend only a tightly constrained set of instructional phrases and may fail to recognise abbreviations or colloquialisms. To reap the benefits of such technologies, it is hence important to ensure that humans are trained to suitably and consciously frame, as well as constrain, their queries and instructions. In addition, investing in and deploying sufficient training tools, in the form of simulators that provide immersive training, may be important to transition a workforce that may be sceptical of the efficacy of such agents.

CONCLUSION

I have shared why AI-powered situated conversational agents are attractive, and articulated how ongoing advances in AI and IoT/sensing technologies are likely to translate my vision into practice over the next few years. However, there are limits to the contexts under which AI-human co-working can be realistically implemented within the next five to seven years. For example, when a task requires precision, speed, and flexibility, such as a complex surgery, humans would still have to be the ones to execute the task. Even so, we envision a future when AI agents are going to be embedded in pervasive devices like wearable smartglasses and service robots, such that these AI models can utilise inputs from multiple embedded sensors to provide situated and immersive comprehension of human instructions or augmentation of human perception capabilities.

Several researchers, including myself, are working to translate this vision into reality. One of the identified strategic research pillars at Singapore Management University has been labelled "Human AI Synergy" (HAIS)–it encompasses not just the sort of situated agent capabilities described here, but also a broader set of initiatives that seeks to reimagine the human-machine interface beyond current practices involving pure voice and touch-based interactions via screens. We are developing mechanisms to support conversational interfaces that help provide visually-impaired people with situational awareness of their surrounding environment, or allow natural verbal instruction-based programming of objects and their behaviour in virtual immersive environments. In addition, as part of an MIT (Massachusetts Institute of Technology)-led, NRF (Singapore National Research Foundation)-funded programme titled "Mens, Manus and Machina: How AI Empowers People, Institutions and the City in Singapore" (or M3S for short), I am working to develop the foundational AI/AR capabilities, and prototype situated agents, to support AI-augmented execution of industrial repair and maintenance tasks, as well as enhance the interactivity of online learning platforms.

So, if we indulge in a bit of imagination and science fiction, we might end up with something not unlike JARVIS.

Dr Archan Misra

*is Vice Provost (Research), Lee Kong Chian Professor of Computer Science, and Co-Director of the A*STAR-SMU Joint Lab in Social and Human-Centred Computing at Singapore Management University*

Endnotes

1 Global Market Insights, "Autonomous AI and Autonomous Agents Market Size", November 2023.

2 JARVIS refers to a fictional character in the Marvel comics. Short for "Just A Rather Very Intelligent System", it is an AI personal assistant to Tony Stark who is also Iron Man in the comic series.

3 Bill Gates, "AI is About to Completely Change How You Use Computers", GatesNotes, November 9, 2023.

4 Dulanga Weerakoon, Vigneshwaran Subbaraju, Nipuni Karumpulli, et al., "Gesture Enhanced Comprehension of Ambiguous Human-to-Robot Instructions", 22nd ACM International Conference on Multi-Modal Interactions (ICMI), 2020.

5 Dulanga Weerakoon, Vigneshwaran Subbaraju, Tuan Tran, et al., "SoftSkip: Empowering Multi-Modal Dynamic Pruning for Single-Stage Referring Comprehension", ACM Multimedia, 2022.

6 Some AI models may be required to run as many as a billion variables or parameters. In a nutshell, a parameter is a factor that the model needs to compute a response. Examples of factors are distance in the horizontal plane (i.e., the x-axis), a specific primary colour (e.g., red), and temperature.

7 Dulanga Weerakoon, Vigneshwaran Subbaraju, Joo Hwee Lim et al., "CAS: Fusing DNN Optimization & Adaptive Sensing for Energy-Efficient Multi-Modal Inference", under preparation.

8 Dulanga Weerakoon, Vigneshwaran Subbaraju, Nipuni Karumpulli, et al., "Gesture Enhanced Comprehension of Ambiguous Human-to-Robot Instructions", 22nd ACM ICMI, 2020.

9 Dulanga Weerakoon, Vigneshwaran Subbaraju, Tuan Tran, et al., "SoftSkip: Empowering Multi-Modal Dynamic Pruning for Single-Stage Referring Comprehension", ACM Multimedia, 2022.

10 Osmond Chia, "Autonomous Cleaning Bot to Start Scrubbing Public Toilets in Early 2024", The Straits Times, December 3, 2023.

11 James Pero, "Tesla's Optimus Robot Can Kind of Fold Clothes Now", Inverse, January 18, 2024.