FRAMEWORK FOUNDRY
THINKING TOOLS. COMPETITIVE WEAPONS.

Image credit: prompt by Rowland Chen, rendering by ChatGPT

ChatGPT Is Unconscious.
To Achieve Machine Sentience, Humans Must Solve for Artificial Consciousness.

Rowland Chen, M.B.A., M.S. and Robert E. Tench, Ph.D.

The world is abuzz with chatter about generative artificial intelligence and the existential threat to humanity that it could present. Or there is no threat, according to Kyunghyun Cho, an associate professor at New York University. In any case, the most widely recognized application of generative AI is its embodiment as ChatGPT from OpenAI. For those who do not know, ChatGPT is computer software that will have a synthetic conversation with you through something called a large language model. For example,

> **Human prompt**: *Hemingway wrote a six-word novel. "For sale. Baby shoes. Never worn." Write a six-word novel using these two words: machine and sentience.*
> **ChatGPT response**: *"Machine gained sentience. Humanity, forever altered."*

As amazing as the response seems, ChatGPT is unaware of the irony and awe-inspiring thoughts its six-word novel raises in the human reader. As AI philosophers, scientists, designers, and engineers pursue software that regurgitates pre-loaded data, mimics human styles to write text, generates visual art, and produces other seemingly creative artifacts of human behaviors, one cannot lose sight of the biggest prize – developing a machine with consciousness. Some believe that human emotions cannot be programmed into a computer, thus missing the mark on developing artificial consciousness. Before putting up walls in front of ourselves, first, let us decompose consciousness into manageable parts.

Consciousness is a complex topic regardless of any effort to program a computer to simulate it. Scientists and philosophers have studied and analyzed this topic extensively for millennia and, most recently, in the context of anesthesia, as in the work conducted by Professor Emery Brown at the Massachusetts Institute of Technology. Here, we offer a novel framework for human consciousness that could prove helpful in understanding consciousness and then enable a successful attack on the problem of artificial consciousness.

The new consciousness framework has five components: curiosity, critical thinking, creativity, compassion, and conscience (Five Cs). The Five Cs drive human emotions, experiences, interactions, problem-solving abilities, moral thinking, and ethical behaviors.

Curiosity, the first of these components, motivates exploration, learning, and innovation. An intrinsic desire to define the new, challenge the status quo, and continually expand the boundaries of our knowledge defines curiosity. Curiosity asks us, "Why?" It pushes us out of our knowledge zones, prompting us to engage with the world more meaningfully.

While curiosity in AI could lead to incredible advancements, it could also result in unintended and unforeseen consequences if left unchecked. AI systems driven by relentless curiosity could encounter or even create risky scenarios in their quest for knowledge, posing potential hazards to individuals or society. Consequently, developing AI systems that balance curiosity with understanding safety constraints and ethical considerations is essential.

Critical thinking, the second C, is a disciplined process of actively and skillfully conceptualizing, applying, analyzing, and evaluating information. This component is an essential part of informed decisions, identifying false information, and understanding the potential implications of actions. Critical thinking encompasses elements of skepticism and doubt, but also open-mindedness. Thinking critically encourages us not to accept things at face value but to delve more deeply, probing for the truth.

Critical thinking, though a crucial future capability of AI to evaluate information and make decisions, also introduces risk. Advanced AI capable of sophisticated reasoning may lead to unintended consequences if the system's objectives are not perfectly aligned with human values that vary between cultures. As a result, we must ensure that AI systems apply their critical thinking abilities in a way that ensures human safety, minimizes bias, and upholds cultural norms in the absence of absolute standards.

Creativity, the third C, is the ability to generate, recognize, and communicate unique and innovative ideas. Creativity often arises when we use our curiosity and critical thinking to see new connections and possibilities. The spark that lights up when disparate ideas collide, birthing something entirely new is essential in the creative process. Creativity can take many forms, including artistic expression, scientific innovation, and problem-solving in tough situations, such as in chaotic automobile traffic for an autonomous vehicle.

Creativity in AI can be a double-edged sword. On the one hand, creativity is an essential feature of problem-solving and innovation. On the other hand, an AI might devise art, solutions, or strategies that are appealing from a machine perspective only. Currently, AI strives toward objectives set by humans. AI cannot assess whether a human regards an artificial artifact as satisfying. Automated judgment of quality is missing. Another consideration is whether the current approaches to AI, based on neural networks, have hit an imitation barrier. A new paradigm for developing creative computers will lead to breaking that barrier. As with the other Cs, a risk exists with creativity in AI. An artificial creative intelligence machine could create misinformation if left to its own devices.

Compassion, the fourth C, is the empathetic understanding of others' experiences and emotions, which drives us to act fairly towards others. Compassion in consciousness implies an ability to identify and understand the elation and the suffering in others and respond accordingly. Developing an emotional intelligence that transcends self-centric perspectives and embraces the broader human condition is one minimum requirement for human consciousness.

Incorporating compassion into AI is perhaps the most complex, essential task. Compassion is crucial for AI to understand and respect human emotions and values to avoid causing harm or distress. Developing

artificial emotional intelligence is a minimum requirement for consciousness. Without compassion, the risks of bias and exclusivity abound. A conscious machine can balance differing perspectives. Without multiple perspectives, bias creeps into any results from AI. Lacking compassion, AI risks operating as an unfeeling, non-empathetic automaton lacking the human touch.

Conscience, the fifth and final C in our framework, supports an inner voice that guides people's behaviors based on their morality and ethics. Conscience acts to discern right from wrong, helping us navigate through the complex fabric of societal norms, personal beliefs, and ethical principles. The role of the conscience differs among individuals and across cultures, influenced by upbringing, education, societal norms, and personal experiences. Conscience serves as a self-regulatory mechanism, guiding our actions and decision-making, and prompting feelings of guilt or satisfaction depending on whether our behavior aligns with our moral standards.

Conscience in AI, as researched by Dan Hendryks and his colleagues and others, presents interesting technical and philosophical questions as we advance in the realm of artificial intelligence. It nudges us toward uncharted territory where we consider whether non-human entities could potentially possess moral reasoning or a sense of right and wrong. Philosophically, the pursuit of consciousness in AI raises profound queries about sentience, AI safety, and the nature of morality itself. Here we enter an extremely murky area involving computer science, neuroscience, psychology, philosophy, anthropology, and spirituality. A solution is to build fairness into AI, but a risk exists if the metrics of fairness are not aligned with societal ethics, personal morals, and accepted norms, which vary from culture to culture.

By decomposing consciousness into the Five Cs, we can pursue artificial consciousness by pursuing each of the parts as independent research endeavors. Developing a conscious computing machine then becomes the difficult task of integrating these five disparate artificial systems. However, if AI designers have future integration in mind, then combining the Five C subsystems becomes more straightforward. Interfaces and integration points must be included in the collection of systems requirements.

Understanding and replicating human consciousness in artificial intelligence is a monumental challenge. Breaking it down into the Five Cs - curiosity, critical thinking, creativity, compassion, and conscience - allows researchers to approach the task as separate yet interconnected solutions. Much like the method used in other "moonshot" projects, literally, like the Apollo moon mission, multiple teams working in parallel could be formed to accelerate the pursuit of artificial consciousness.

However, we must proceed with great and profound caution as we converge on artificial consciousness.