JNphi

Examining Some Serious Challenges and Possibility of AI Emulating Human Emotions, Consciousness, Understanding and 'Self'

Sadia Tariq[*], Asif Iftikhar[τ], Puruesh Chaudhary[ϕ], Khurram Khurshid[δ]

Abstract

This research falls in the ambit of 'AI and Philosophy'. It explores why emulating the complex processes of subjective experience, emotions, consciousness, self-awareness, and the human personality, will be a huge challenge for AI research. It touches upon some finer aspects, like the huge variety of human emotions and feelings, processes of future and fringe consciousness, and the evolution of self-awareness and complex human self/personality, whose practical realization in an AI system would be very difficult, if not impossible. In the backdrop of these serious challenges it also discusses an interesting possibility of emulation in the field of Hermeneutics, focusing on hermeneutics of *ṣan'at-e ihām* (construction of ambiguity) in Urdu, Persian, Arabic, Hindi, and Punjabi poetry. The basic purpose of this work is to underscore the critical importance of this area for AI research, and to shed light on those aspects of consciousness, emotions, awareness, understanding and self whose comprehension and inclusion is necessary in designing and constructing AI systems that can parallel human mental functioning.

Key Words: mind, brain, artificial intelligence, emotions, hermeneutics, consciousness

DOI: 10.5281/zenodo.6637757

Introduction

The rudimentary origins of AI lie in the human capability of mathematical abstraction. It is the growing complexity of this mental capability which created the need, motivation, imagination and the idea within human beings to artificially construct some external

[*]**Corresponding author:** Sadia Tariq
Address: Philosophy Department, Sanjan Nagar Institute of Philosophy and Arts, House # 13, Gulberg 5, Off Zafar Ali Road, Lahore, Pakistan, Phone: +92 322 4493079.
e-mail: sadia.tariq71@gmail.com
[τ]Asif Iftikhar, Department of Humanities and Social Sciences, Lahore University of Management Sciences, DHA, Lahore Cantt, 54792, Pakistan.
[ϕ]Puruesh Chaudhary, Founder and President, Agahi, 138-B, Street 38, D-12/2, Islamabad, Pakistan,
[δ]Khurram Khurshid, Electrical Engineering Department & Artificial Intelligence Cell, Institute of Space Technology, Islamabad Highway, P.O. Box 2750, Islamabad 44000

devices or tools to support and extend the reach of this capability. The ancient Abacus made by Mesopotamian Sumerians was one of the first instruments created for assisting human calculation (computing) of increasing quantities (Garfinkel and Grunspan, 2018). The Romans made wax tablets for recording and storing symbolic information and Cryptographic Scytale for long distance communication. Metal based mechanisms were created by Greeks for calculating and predicting cosmological events like eclipses, motion of heavenly bodies, and seasons. These physical extensions of the mathematical human capability eventually gave rise to conceptions and imagination of human-like artificial assistants. Homer's Iliad mentions self-propelling tripods (chairs) and 'Golden Attendants' to help a physically disabled blacksmith (Nilsson, 2010). Similarly, Aristotle conceives of self-moving and self-motivated tools which could eliminate the need for human slaves (Nilsson, 2010). Then Leonardo comes up with the design of a humanoid robot embodied as a Knight which can move around on its own. Stories of Rabbis creating artificial servants called 'golems' to serve them can be found in the Talmud (Nilsson, 2010).

In addition to the usefulness or utility aspect there is also another reason why the human mind conceived and created Artificial Intelligence. The human spirit has this ancient and pervasive urge to reproduce itself in some extraordinary way (McCorduck, 2004). So apart from performing a very practical utilitarian function the creation of AI also served an abstract subjective desire of our minds.

Since the model and benchmark for intelligence in artificial systems has been the human mental (emotional, intelligence, and intellectual) processes hence a constant comparison between the two and emulation of the latter were logical concomitants of our obsession with creating artificial life.

Just as understanding the human mind has seen its share of hard and easy problems and the manifold stages of progress in both areas, similarly AI has its own set of hard and easy problems with varying levels of progress in each area. Conceptual and practical progress in easy problems (pattern recognition, error detection, etc.) has been faster and more obstacle free while work in hard problems like AGI, Superintelligence, sentience AI or Emotion AI, has been relatively and understandably slow, with no major breakthroughs, as of now. In this paper, we share some comparisons for the purpose of identifying some important challenges and an interesting possibility of AI agents and systems getting close to human mental functioning.

Experiential, Emotional and Conscious Processes in AI and Humans

It would not be an exaggeration if we say that the area of emotions, experience and consciousness (all interconnected) is the holy grail of AI and AGI research. It is the main obstacle in the realization of true

artificial intelligence and artificial general intelligence. Without this any claims and achievements of AI and AGI research will remain hollow and impalpable. Since the accent of mainstream AI research has been on cognition, perception, reasoning and logic hence this area was largely sidelined and got a kind of step motherly treatment for a long time. More so because the dominant myth of human beings being rational agents and decision makers has been molding human thinking and doing since the capitalist system and its cultural and economic paradigms took root. It is very recent that the critical role of emotions, feelings and experience in human thinking and intellectual functioning has been recognized and seriously studied. According to Lombardo (2011), "as a general rule upbeat emotions such as love, hope, enthusiasm, and courage positively impact human thinking–including creativity–whereas negative emotions such as fear, anxiety, sadness, and depression damp out effective and creative thinking." (Lombardo, 2011, p.28). He also connects emotions and creativity with future consciousness and wisdom.

The inception of 'Emotion AI', as a proper research field of AI, was in 1995. It was conceived and created as a tool for measuring, understanding, simulating and reacting to human emotions so as to allow a more natural interaction between humans and machines (Somers, 2019). For this purpose, many algorithms and tools have been developed for automated emotion recognition through facial expressions, posture movements, physiology, and even dialogue (Picard, 2002). Deep learning methods are being employed to "develop emotion classifiers…and dialogue models of dialogue services." (Huang *et al.,* 2020, p.1). These are trained to track human emotion and intention and respond accordingly during interaction with humans. Today, Emotion AI companies and apps are functioning in many areas like Advertising, Call centers, mental health, automotive and many others (Somers, 2019). Initiatives like 'Jibo,' the household robot companion (Conroy *et al.,* 2020) which is being used to give emotional support to people in order to address issues of moods, mental health and wellbeing, have also come up. So it is a burgeoning field with mushrooming of applications. But the basic nature and purpose of these efforts is to enable AI systems to overtly display emotional intelligence while interacting with humans and to assist humans in boosting their own capabilities to better manage their emotions (Picard, 2002).

These efforts are not yet delving into the deeper distinctions and aspects, both structural and functional, of the elaborate human emotional process of contemporary humans. For instance, we do not find any systematic work on how AI systems will identify the fine but simultaneously fuzzy demarcations that exist between emotions and feelings, and how will the perception and cognition of this distinction be installed as a symbolic or sub-symbolic program in an AI system. In our view, the existing so called 'emotionally sensitive' AI systems (including future AGIs) cannot and might never be able to distinguish,

measure or simulate Damasio's (2010) three broad categories of emotions; universal, background and social, and the huge variation that exists within each type. Interestingly, in this huge range some varieties are accompanied by corresponding brain and body responses while others are not. According to Aaron Sloman (2004), the emotion of "admiring someone's courage while being jealous of his wealth" would not be "expressible within somatic marker" and he goes on to another example of "emotions that endure over a long period of time while bodily states come and go (such as obsessive ambition, infatuation, or long term grief at the death of a loved one)" (Sloman, 2004, p.3). We would like to add to this list the deep-rooted and continuously internally operative emotions of long-term enmity and mistrust which are behaviorally and physically neither manifest nor detectable. Similarly, tracking, measuring, and simulating the nonverbal library of emotive likes/dislikes, motivations, desires, etc., and their interconnections, relationships and organization that exist as a hybrid brain-mind (neuronal synaptic connections and mental conceptual images or representations existing at the quantum level maybe) memory store, will also be a huge challenge. The reason being that it is not yet observable and amenable to our laboratory and even mathematical tools, techniques and methodologies. We are aware of many more such aspects and factors which pose a serious challenge to imitation or installation of emotional processes in AI but those will be the subject matter of another paper.

'Artificial Sentience' (AS) is a more recent frontier and a huge challenge. It is about exploring and transferring "of the functions and abilities of human experience and senses to a machine" (Lavelle, 2020, p.64). This field covers a number of distinct aspects of human experience like perception, sensation, emotion, sentiment and even consciousness (broadly) (Lavelle, 2020). The main purpose being again to endow machines with the capability to 'express' sensations and emotions in their interaction with humans. Just as the 'Principle of Artificial Intelligence' is about Strong and Weak AI similarly 'Principle of Artificial Sentience' (PAS) has been formulated in terms of 'Principle of Strong AS' and 'Principle of Weak AS'. The former states that "A machine can feel *exactly* like a human, from the point of view of sensation and emotion and of consciousness" while according to the latter "A machine can feel *approximately* like a human, from the point of view of sensation and emotion and of consciousness" (Lavelle, 2020, p.67). Since the term 'sentience' is about the capacity or capability to feel from a subjective first person standpoint i.e., 'I' and also have a sense of being aware of one's surroundings (Lavelle, 2020), hence one can see the discrepancy and the huge gulf between what 'sentience' actually denotes and the aspiration projected in the two principles of Artificial Sentience.

The present theoretical and practical work in this area is very far from the realization of the above mentioned principles of AS. There are many ideas about how they should and can be attained but as of now

there is no AI system or robot which can be called 'sentient' in the sense in which the term is used as a representation of an actual process and mechanics that evolved and exists in living things, especially its developed form in humans. In human beings, 'sentience' with all its layers and complexity is inbuilt or a default installation integrally connected to and working in tandem with other mental processes in the human body-brain-mind complex. Given this situation we find the following suggestion more realistic and achievable; "engineering artificial phenomenology (i.e., a functional equivalent of phenomenal experiences) rather than human-like phenomenal experiences." So the challenge then "becomes one of engineering a capacity fulfilling the same functions as phenomenal experiences do within cognitive processes, but remains agnostic regarding the actual qualitative dimension." (Zaadnoordijk and Besold, 2019, p.3).

Recent progress regarding installation of experiential and emotional processes in AI consists largely of thought experiments conceiving 'abstract animat' devices, which are hybrids of biological and artificial constituents (Schweizer, 2018). Paul Schweizer (2018) anticipates one such scenario wherein a robot has an artificial physical formation but a human cognitive architecture based on human Language of Thought. That is, we have managed to replicate the "*abstract* computational structure of human cognition" (Schweizer, 2018, p.84) in the robot's synthetic brain. Of course he does not mention the complex level and layers of human cognition and the even more complex intellectual functioning accompanied by developed sensitivity and emotive functioning that we have been explicating in our argument. However, he does inform us that conscious states like perceptions, cognitions and emotions are not really "tethered to *abstract* processing structure" but connect directly to "biochemical influences" (Schweizer, 2018, p.87). This means "… The hypothetical robot brain sustains a form of artificial consciousness that is qualitatively distinct from ours, and potentially very alien." (Schweizer, 2018, p.88). So it seems that even if we are able to replicate the human cognitive process and Language of Thought in the robot its conscious experience and emotional states will never be purely human but remain a "bio-machine *hybrid*". Hence the hard challenge of installing the developed human emotional processes and 'sentience' in AI.

The capturing and expressing of emotions and sentience at the behavioral and verbal levels can be imitated but this must not be confused with human-level and human-like emotional and experiential functioning. As Wang (2013) says:

> "Systems with emotion" is different from "systems with *human* emotion." Since computers are electronic, not biological, they are not going to have the biochemical and physiological aspects of human emotion. Nevertheless, there will be analogous processes and phenomena that deserve the name of "emotion",

since it plays the same functions for the AI systems as emotions for human beings." (p.193)

The phenomenon of 'Consciousness' and its study, in general, is the umbrella under which numerous mental processes, functions and capabilities are separately researched. There is no agreed upon exhaustive list of which mental processes it includes but some main ones are perception, cognition, attention, awareness, sentience, emotion, feelings, imagination, creativity, reasoning, and intelligence. The field of 'Artificial or Machine Consciousness' which is integrally intertwined with the above two subsets of AI, has grown a lot in the past few years; numerous thought experiments, simulations, computational models, and robotic experiments have been carried out. However, "… even with recent advances, no current state-of-the art AI approaches can compete with simple animals when it comes to adapting to unexpected changes to their environment or any of a wide range of behaviours commonly associated with consciousness" (Crosby, 2019, p.1). Most of the proposed testing of Artificial Consciousness focuses on the measurable connected aspects instead of the phenomenal component (Crosby, 2019). But as an initial step the possibility of creating 'minimally intelligent conscious systems' is being suggested and explored through thought experiments. The basic concept is that "if any AI systems are ever conscious, the first set of conscious entities will be the least capable that we can create." (Crosby, 2019, p.3). They will be like "human infants with cognitive and emotional deficits." (Crosby, 2019, p.3).

Another possibility which leans more on the practical side is about embedding an artificial pain nervous system and a mirror neuron system into robots for them to feel pain within themselves and in others. Practically, a 'soft tactile sensor' has been developed for this purpose. It is believed that "… Based on the capability of discrimination of the tactile sensor, artificial nervous system for pain sensation can be embedded into the robot body and brain in parallel with normal mechanoreceptor pathway with a mechanism of pain regulation" (Asada, 2019, pp.4-5). Of course it is nobody's case that the sensation of pain generated by this artificial nervous system will be anywhere near the human subjective sensation and experience of pain. The important point to be clear about is that externally embedding an artificial nervous system made up of artificial sensors and actuators and the products it will produce is a completely different process from the evolutionary process of the human nervous system and its complex pain (and also pleasure) producing mechanism. Where the types of pain range from elementary reflexive to highly abstract pain existing at the purely mental level and everything in between.

There is another distinctly human form of consciousness which will also pose a huge challenge for AI research; the process of 'future consciousness' which is an increased awareness of not only what could but also what should occur in the future (Ahvenharju *et al.,*

2018). This form employs 'memory of the future' or what is called 'episodic memory', an increasingly explored and important area in brain sciences. This form of memory is also associated with autonoetic or self-knowing which enables a person to become "aware of her own past as well as her own future; she is capable of mental time travel, roaming at will over what has happened as readily as over what might happen, independently of physical laws that govern the universe." (Conway, in press, pp.6-7). The contemporary individual equipped with this form of consciousness is able to imagine not just one 'future' but 'multiple possible futures' (Conway, in press). And since this type of consciousness is necessary for survival and keeping up with an uncertain and rapidly changing environment hence a similar process will need to be installed in AI, if it is to parallel this human capability.

A recent (somewhat ambitious) endeavor, which is a mix of theoretical, philosophical and practical work, aims to develop an autonomous, multi-task capable, powerful, highly (or super-) intelligent, and adaptive system (AMPHIA) with phenomenal consciousness (Sahner, 2019). Its substrate will involve neuromorphic architecture and like human consciousness it will emerge in a social milieu and will have its own set of ethics and values, which might be quite different from those of human consciousness. Again the emphasis is on 'cognitively' and 'functionally' conscious robot which can sense "in a sterile experience-free way" and "behaviorally mimic aspects of consciousness" (Sahner, 2019, p.5). And it is believed that such a functionally conscious robot would be very close to "what we consider conscious" which is a linear projection and an assumption given the reality of the layered complexity of human consciousness. This work also suggests some potential tests and metrics for measuring phenomenal consciousness in this AI system, one of which is worth watching out for. Let us see if and when an AI system can pass this metric wherein it becomes "apparently "interested in" or "wishes to" explore altered states (by altering its own hyper parameters or injecting noise) for no functional reason apart from what seems like "curiosity" and "desire" (Sahner, 2019, p.5). We submit that the proposed actions in this metric will require prerequisites of a sufficiently developed human-like phenomenal consciousness, the autonomy of wishing, desiring and questioning arising out of an integrated and proactive individual subjective identity or personality, complex emotional, motivational and feeling processes and a sufficiently mature intellectual process.

According to a noteworthy conundrum mentioned in this work, and also raised in other such works, "Identifying valid machine consciousness may, however, be challenging, particularly if that conscious self is endowed with high-level mental constructs foreign to our own, informed by raw phenomenal experience tethered to sense modalities we can only imagine…" (Sahner, 2019, p.7). Connected to this is the point that "Natural language descriptions of consciousness are also likely to be highly misleading because artificial systems might

have different representations of the temporal and spatial properties of objects" (Gamez, 2019, p.4). If this is so then it will be like 'alien' consciousness for us and until it establishes some kind of a two-way communication with us in some mutually understandable language, we will never be able to grasp and gauge its essential nature, and the quality and character of its mental constructs.

While discussing his ideas on mathematical consciousness, Sloman (2020) mentions an interesting claim by Turing that "computers could replicate human mathematical *ingenuity* but not mathematical *intuition*…" (p.15). Turing did not explain why he thought this but Sloman suspects that the mechanisms required for replicating mathematical intuition "cannot be provided either by current, digital, logic-based forms of computation, or by neural statistics-based learning mechanisms" (Sloman, 2020, p.15). Because, unlike digital systems, the brain could be making use of "sub-neural molecular computational mechanisms with their combinations of continuous and discrete processes…" (Sloman, 2020, p.15). Elsewhere Sloman (2018) states this same issue in the following manner:

> "If the environment with which a digital computer interacts is not a discrete-state machine, the coupled system, including any virtual machinery used, cannot be modelled with perfect precision on a Turing machine, since no discrete machine can model perfectly a process that runs through all the real numbers between 1 and 2, in order, whereas a continuously changing chemical structure might be able to…" (pp.93-94).

In our view, the above is an important limitation of what Sloman (2020) calls "currently fashionable AI theories and mechanisms" and must be seriously acknowledged and considered by mainstream AI research programs. In addition, some other varieties of consciousness apart from mathematical consciousness that he mentions are also worth noting. He talks of things like 'wondering whether', 'regretting', 'having doubts', 'wanting' (Sloman, 2020). And also of 'mutual metacognition' between humans and animals and "sophisticated mechanisms of self-monitoring" (Sloman, 2020, p.12) which involve consciousness. We agree with him when he says "Human (or more generally animal) consciousness involves far more than typical computational models of consciousness that take in a sensory array of measurements…and produces a new data-structure with labels attached, including possibly a summary label…" (Sloman, 2020, p.2). He demonstrates this through the examples of 'fearful consciousness' which involve a complex collection of reflexive changes at the physiological level and the consciousness of missing an aspect of a mathematical problem which can produce "a different kind of intense, purely intellectual, exertion!" (Sloman, 2020, p.2). So he talks of this range of conscious states involving different substrates, mechanisms and processes. The question is whether both the physiological mechanisms of 'fearful consciousness,' (also accompanied by strong

feelings) and intellectual exertion of 'mathematical consciousness,' (which might not have any physiological or somatic markers that our present empirical tools can detect,) can be measured and simulated by our present digital systems and neural networks or not. It remains to be seen whether it will come into the realm of reality or remain a possibility.

This brings us to another very important form of consciousness; the non-sensory fringe or experience, whose earliest reference goes back to Plato who "insisted that non-sensory experiences underlie our capacity to unify a multiplicity of changing sensations into a single concept" (Mangan, 2001, p.5). Some other thinkers and philosophers who also explored this form of consciousness include Kant, Leibniz, James, Dreyfus and more recently Mangan. According to Dreyfus (1972), "Fringe consciousness takes account of cues in the context, and probably some possible parsings and meanings, all of which would have to be made explicit in the output of a machine" (p.21). Mangan (2001) thinks there is actually an infinity of such non-sensory experiences and feelings. Some examples of this non-sensory fringe that he discusses include 'feeling of familiarity', feeling of 'knowing' or 'rightness'/'wrongness', or meaningfulness, 'feeling of immanence', cognitive integration, free floating anxiety, the feeling of causal connection, sense of 'mineness'… "All expressive feelings such as the sorrow of the willow or the joy of sunshine are non-sensory experiences" (Mangan, 2001, p.8). He finds (and we agree) these non-sensory fringe experiences to be quite elusive and difficult to grasp through direct attention. The phenomenology of such an experience is low-resolution and low intensity. The feelings are "amorphous, fuzzy, diffuse" (Mangan, 2001, p.10).

The problem is that our existing sensory experience and strong, clear, and conscious feelings can still not be measured, modelled, and simulated by the most sophisticated of our technological and laboratory tools and methods and consequently their replication in the most cutting edge of our AI systems and Robots is nowhere in sight, at least in the near future. So in all fairness, it seems thinking about measuring and emulating this non-sensory fringe experience and feelings would be asking too much of AI research at this stage. So our purpose here is to just underscore the importance of this integral variety of human consciousness which plays a crucial role in our thinking and decision making and to submit that if any future AI or AGI is to undertake a dynamic, more adaptive, and meaningful interaction with the real world of humans and the rest of Nature then it will need a similar process alongside the core sensory consciousness.

Self-awareness, integrated functioning, and personality in humans and AI systems

The contemporary human mind has these three in-built qualities of self-awareness, integrated functioning and personality, which are an evolved product of the many-layered, unceasing, and imperative interaction of the human form with its fluid, uncertain and complex external environment during the course of human history. All of them are interrelated and define human mental and social existence. Any serious understanding or description of the human mind for the purposes of emulation would be incomplete without a comprehensive and deeper grasp of these three indispensable qualities and their role in the construction and development of the sophisticated human "… worlds of mental contents" (DeLancey, 2002, p.223), including the 'social' mental content world and its interactions and applications.

The capability and quality of self-awareness (and then integrated mental functioning which led to the formation of human personality) in present humans is a composite of two types of awareness; a subjective awareness and experience of one's body and an awareness of one's mental self as a unified personality. The former is a developed form of the elementary two-fold generic capability that emerged in living things; a proto-sensing and experiencing of one's form as distinct from the environment around it and also experiencing an internal insufficiency and imbalance of one's form. Without this composite internal sensitivity in relation to one's form the core 'needs' to preserve the form or survive as a form and to interact with the environment would not have arisen. The living thing now had to continuously and proactively interact with and respond to its environment in order to redress its internal imbalance and preserve its form. This is how gradually the basic response system of living things evolved based on amorphous categories of preferences (later likes for what was essential for the form) and aversions (dislikes for what was harmful) as opposed to clear mathematical digits like 0s and 1s in AI.

As more complex life-forms with developed bodies and brains emerged in the evolutionary ladder, numerous tiers of a mental response system made up of sub-systems (of specific mental programmers, functions and processes of perception, cognition, problem solving, Will, etc.) were built upon the basic emotional response process and system. In human beings, especially after language, we find these emotional and mental response systems becoming very elaborate, and complex; the mental operating system of contemporary man has become a highly advanced and integrated complex of both specialized and general-purpose mental capabilities, functions and then their programmers and systems (both emotional and mental).

It is when we become aware of our experience and other mental products arising out of the above complex integrated dynamic and functioning of our response systems that there arises a layer of a unified experience, which is in fact a meta-layer of response—our

integrated self or 'I'. Our 'personality' or sense of our own selves are nothing more than an accumulated outcome or aggregate product of the on-going functioning of all our response systems. In terms of micro mental dynamics our sense of personality arises when we start associating our Will response system (which is a simple execution process in less complex and earlier living forms) with ourselves. In the human response systems, it is the Will where it all comes together. The Will represents the whole of one's response capability for all its layers. It is in the Will as an elaborate response system that the final weighing and crystallization of conclusions from all the other response systems takes place. These crystallized conclusions are then sent to the doing part of the Will for actual implementation. So it occupies a critical and leading role in the gamut of all our mental processes and response systems. And our personality becomes what our Will response system is. This idea is not new. The connection between our acts and core character has been proposed by many thinkers and philosophers

The point we want to stress about human personality is that it is not a deliberately architected product of intelligence. It is a response layer that arises during the course of the evolution of mental processes, as specialized response systems. Our experience of this layer and our verbalization of it is what gives it a separate and unique existence. Our individuality in terms of our specific mental operating system is a logical reality in Nature and our awareness of it is also valid and legitimate. But when we have a unified experience of the superstructures built on the basic fact of our individuality and verbalize it then we create this superficial and superfluous layer of 'I' and make it the cornerstone of our existence. We are aware of it because of the specific state and character of human consciousness but we have not consciously or intelligently created it. Consequently, our awareness of our personality is actually an awareness of its superstructural part, and not its fundamental part.[1] The superstructure of our personality has acquired so many layers and dimensions and has become so complex and elaborate that we have lost touch with its fundamental part and its design criteria coming from our mental genes.[2]

The above explained chain from self-awareness, integrated mental functioning to personality in humans is still a partially understood (despite decades of piecemeal but in-depth work on them in philosophy, psychology, cognitive science, consciousness studies, mind sciences, Neuroscience, brain sciences, etc.) and highly complex

[1] The fundamental part of human personality consists of motivational elements, basic positions and paradigms. This part is the design criteria coming from the dynamic between our mental genes, brain and external environment and has been made at an un-inquiring and reflexive level. Whatever reasoning and inquiry that we have done at a deliberate and conscious level has gone into the superstructure. The superstructure consists of devising systems for the pursuit of our fundamental motivations, paradigms and positions.

[2] See (Tariq et al, 2010).

process whose emulation and installation in AI is and will continue to be a huge challenge for AI researchers and engineers. But due to the observed integral connection between these processes and human intelligence and intellect there is a consensus on the necessity of understanding and installing them in AI systems. A lot of detailed but fragmented micro work is happening on understanding and designing these processes for AI but it is still in its early stages, with many gaps of knowledge and other practical obstacles in the way. According to one group of researchers the problems of implementing functions like 'self-awareness', 'self-reference', and 'self-consciousness' in a machine are "both technical and theoretical, as there is no widely accepted theory about them, and even their definitions are highly controversial." (Wang *et al.,* 2018, p.1). Since we cannot possibly be aware of all the important works happening in this area, and there are many, so we will just be touching upon a few that we came across and found of interest for our work in this paper.

Since the human 'Self' in all its complexity is too humongous a challenge to understand and emulate hence a new research trend has emerged of focusing on 'human minimal self' and 'artificial minimal self'. Where the 'minimal self' is defined as "…the pre-reflective experience of being a self, or the awareness of oneself as a subject of experience…" and "characterized by two important aspects: a sense of body ownership…and a sense of agency…" (Wang *et al.,* 2018, p.1). This work involves coming up with behavioral and computational components representing some capabilities of the human minimal self like self-exploration, curiosity, body representations, and sensorimotor simulations and predictive processes in the human brain. And it also highlights the need for defining and designing metrics for an artificial self. The rationale and anticipation is that "Although we are far from establishing whether artificial agents can ever undergo subjective experiences, these metrics may provide support and insights in the investigation of the self, in both robots and humans." (Hafner *et al.,* 2020, p.7). A noteworthy part of this work is the following question at the end which echoes an issue that we have also raised in our work: "In robotics, we can access internal states and inspect sensorimotor and prediction information. However, to what extent can this privileged point of view allow us to state—if ever possible—that a robot is undergoing subjective experience?" (Hafner *et al.*, 2020, p.7).

'Inner Speech' used by the human mind is an integral component of self-awareness and is also used as a tool for self-correction, self-focus, self-discipline, not only in relation to external issues, problems, interactions but also for internal communication between different parts and processes of the mind like emotional/sensitivity (nonverbal) and intellectual (verbal) processes. And also for modification of some mental processes. At the same time, inner speech can also be used by our egoistical or adversarial part of our personalities to reinforce and validate their negative, irrational, and insensitive thinking, ideas and

actions. So inner speech can have myriad roles and functions within the human mind. Installing such a process in a robot will again be a superficial and dumbed down replication. But recently an artificial cognitive architecture and model for inner speech has been proposed by a group of Italian researchers (Pipitone *et al.,* 2019). It is based on perception and action modules, a layered memory system and a cognitive cycle. We have not come across any reporting of its successful implementation in robots.

Regarding emulation of integrated mental functioning we came across an earlier work called 'Novamente AI Engine', an integrative AGI design and architecture which like the human mind is to be a "system that can achieve complex goals in complex environments" (Looks *et al.,* 2004, p.1). The underlying tenets of its design we found quite interesting. For instance, the mind is understood as an "interpenetration of a physical system with an abstract set of patterns [abstract programs]…" (Looks *et al.,* 2004, p.1), something like the interconnection between the 'brain' and 'mind'. Thus intelligence is seen as "a problem of finding compact programs that encapsulate patterns in the environment, in the system itself…" (Looks *et al.,* 2004, p.1). Seeing intelligence as emerging through "situated and social experience" which can lead to "autonomy, experiential interactive learning, and goal-oriented self-modification…" (Looks *et al.,* 2004, p.1) is another important tenet. The last one views minds as "self-organizing systems of agents which interact with some degree of individual freedom, but are also constrained by an overall architecture involving a degree of inbuilt executive control…" (Looks *et al.,* 2004, pp.1-2). All these tenets bring it closer to the integrated mental functioning that we explained earlier of general purpose and special purpose functioning and specialized modular response systems unified by our personality or 'I'. It was an ambitious project which also incorporated aspects from pre-existing AI works and paradigms. The aim was to create "a holistic digital mind in a direct way…" (Goertzel, & Pennachin, 2007, p.63). We could not find any latest update on the present status of this design, and its implementation stage.

A more recent work in this area is on a general AI system called "open-ended, modular, self-improving Omega AI unification architecture" (Ozkural, 2020, p.267). It is an improvement upon Solomonoff's general Alpha architecture in which Godel machine architecture and deep learning methods are merged. The Godel architecture is about providing it with a certain level of 'self-reflective' thinking. It has many important components but 'AI Kernel' is "the basic problem solver that is smart enough to bootstrap the rest of the system" and is designed to deal with "all types of data, and tasks…" (Ozkural, 2020, p.270). It can be compared to the general purpose mental processes in the human mind and the unified 'I' which integrates the various response systems, functions, programs and their languages, and enables their coordinated and efficient functioning. Of course, the human mind's process of integration of

modular and general purpose functioning and its products are of another kind and level. This AI Kernel supports real-time operation which brings it closer to human mental functioning. It uses a combination of different reference machines/languages and different types of neural networks. Then there are higher order cognitive modules like self-improvement, analysis and synthesis, etc. However, proper and mature implementation and a working system operating this architecture are still far off. So far only two of its eight reference machines have been implemented. And for it to acquire the minimum intelligence for tackling human kevel problems "a basic set of primitives" (Ozkural, 2020) are still to be decided.

Last but not least is a very important eye catching work in this area; the concept of 'self' in the general intelligence system called 'Non-Axiomatic Reasoning System' (NARS). NARS architecture is designed specifically to emulate the 'general' character of human intelligence and mental functioning. The definition of 'intelligence' here being "the ability for a system to adapt to its environment and to work with insufficient knowledge and resources…" (Wang *et al.,* 2018, p.2). This makes NARS an adaptive system like the human mind which is open to unanticipated sudden events or inputs and learns from its experience. Although this adaptability has an underlying "inviolate level" or meta level made up of inference rules and the control strategies of the system (Wang, 2007).

All the components of this architecture and their micro details, which reflect deep observation and understanding of what the human mind entails and how it works, have been carefully designed to reflect their proximity to human intelligence and mental functions. There are understandable glaring differences also and a clear discrepancy between the elaborate and complex theoretical ideas and articulations of its components' characteristics and the way it will operate, and the actual implementation stage of this system. For instance, its generalization of procedural knowledge has so far been tested on things like activating of switches and autonomous labelling and identification (Hammer & Lofthouse, 2020). Some improvements have been made in its design and the new architecture OpenNARS for applications (ONA) aims to address OpenNARS's limitations. But the 'general purpose reasoner' based on it is nowhere near the qualitatively advanced and very complex general purpose reasoning and functioning of the human mind. And it is yet to be implemented in a robot.

NARS uses 'Experience Grounded Semantics' as opposed to 'Model-Theoretic Semantics' normally used in other systems. In case of the former "truth and meaning have to be grounded on the system's experience" while in the latter it is "the model [a constant reference] according to which truth and meaning within the system is determined" (Wang, 2007, p.44), which means the 'big picture' remains unchanged.

The terms used in NARS and the concepts they represent and connote are very different from how they are used and experienced by the human mind. To take an example, within the human mind the meaning and essence of concepts like 'truth value', 'desire value', 'system's experience', 'stream of consciousness', 'satisfaction-evaluation mechanism', 'event', 'knowledge' and finally the concept of 'self' cannot be reduced to how they are conceived and used in NARS. We can see a few examples of this. In NARS, "the truth value of a statement measures its extent of evidential support, rather than that of agreement with a corresponding fact…Hence…the truth of each statement and the meaning of each term are grounded on nothing but the system's experience." (Wang *et al.,* 2018, p.3). Similarly, "… The actual experience of NARS is a stream of statements, with their truth values represented by the <f, c> pairs. Within the system, new statements are derived by the inference rules, with truth-value functions calculating the truth values of the conclusions from those of the premises…" (Wang, 2007, p.46).

If we compare the above meanings and definitions with what truth and experience means to us and how we understand them, the stark difference will be visibly clear. A stream of statements in words, symbols, and digits can never do justice to our rich and many-sided experience of a feeling, emotion, or an idea. Wang (2013) also mentions this in his detailed work on NARS. We can compare this to how our own articulation or verbalization of our experience or even pinning it down to some written form is almost always an inadequate, insufficient and unsatisfactory representation of that experience. It lacks the organic richness and feel of the actual process. Thus, the concept of truth and determining the truth value of something involve a paraphernalia of ideas, thinking, sensitivity, motivation, questioning, rejection, imagination and many other processes which are not necessarily grounded in or channelized by our experience. The human mind, more precisely its intellectual process, integrates a range of functions and methods to determine the truth or falsity of a phenomenon or any representation of it. So, in a manner of speaking, we can say that it makes use of both experience-grounded semantics and model-grounded semantics. Because we create 'big pictures', 'world views', 'macro frameworks', 'integrated reality' and 'objective reality' and these influence our micro subjective experience and it is their non-linear composite which then determines the truth value that we attach to something.

For many humans, correspondence to objective facts or objective reality is a measure of truth. So the subjective experience of our form is very important but that is not always a measure for deciding the truth of some process or phenomenon. In fact, many times our subjective experience takes us away from the truth and actual reality into the realm of a parallel reality concocted by our minds. Thus, the criteria and actual process of determining true or false, correct or incorrect, right or wrong, etc., (alongside the greys in between) within

the human mind is a manifold process and cannot be reduced and represented merely by statements and equations. As an aside, we propose that NARS based artificial general intelligence can be viewed as corresponding, in some measure, to human intelligence but not the complex and advanced intellectual functioning mentioned above. This means installing NARS in robots will not produce a successful emulation of our intellectual process. Because it can only capture and model the surface layer or superficial aspects of our intelligence process but not those advanced intellectual functions which represent our essential 'human' spirit and 'self'.

To further illustrate our main point, we look at the concept of 'self' as proposed in NARS. 'SELF' in NARS is defined as "…a concept, with built-in operations that can be directly executed from the very beginning of the system's life..." (Wang *et al.,* 2018, p.6). These internally implemented operations are referred to as "mental operations". These are based on 'mental' sensors placed "on the reasoning/learning process, which express information about the state of the system in a format (Narsese sentences) that can be directly processed by the system." and 'mental' actuators which "… though…carried out by physical processes, they are known to the system only at an abstract level, without their physical details." (Wang, 2013, pp.186-187). The role of these 'mental operations' is that they "…contribute to the system's self-concept by telling the system what is going on in its mind and allow the system to control its own thinking process to a certain extent." (Wang *et al.,* 2018, p.7). Due to the earlier mentioned experience-based foundation and functioning of NARS:

> "The notion "self" does not have a constant meaning determined by a denotation or definition. Instead, the system gradually learns who it is, and its self-image does not necessarily converge to a "true self." Since the change of meaning of a concept is done via the additions, deletions, and revisions of its relations with other concepts, the system's identity (determined by all the relations) is relatively stable in a short period, although in its whole life the system may change greatly, even to the extent of unrecognizable when compared to a previous image of itself." (Wang *et al.,* 2018, p.6).

NARS also has a 'satisfaction-evaluation' mechanism for assessing events. So events are given a 'truth value' (current status) and a 'desire value' (what the systems wants it to be). The "closeness between them is called "satisfaction"" (Wang *et al.,* 2018, p.7). The value of satisfaction can be between [0, 1]. There is a 'system-level' satisfaction also which is an aggregation of event-level satisfactions and "indicates the system's extent of "happiness" or "pleasure," (Wang *et al.,* 2018, p.7). This plays many roles within the system including 'resource allocation'. For the system to become aware of the satisfaction indicators' values, "feeling" operators reflecting these values in "the internal experience of the system" are implemented. This

is done through using "reserved terms and statements, which form the category of "emotional concepts" within the memory of the system" (Wang *et al.,* 2018, pp.7-8). These emotional concepts provide "perception of emotion" to the system itself. There is also provision for the interaction of these emotional concepts with other concepts through the inference process to generate compound concepts. So these so called 'emotions' and 'feelings' of the system which are a huge part of its internal experience also "contribute to its self-control" (Wang *et al.,* 2018, p.8). It is pertinent to mention here an interesting admission made by the originator and developer of NARS about emotions and feelings in the system. According to Wang (2013), "It will be more natural to say that the system has different feelings for "objects and things", though accurately speaking, the feelings are about the concepts representing the "objects and things" (p.191). Thus, whether it is the internal 'mental operations' which influence the AI system's self-concept or the actual operation of the 'self' concept in an AI architecture, this 'artificial 'self' is a completely different organism from the 'human self'.

We came across a suggested list of 'mental operations' made up of operator terms and the functions they are to perform within the AI system. One of the operator terms in it is 'observe' and its function is to "get an active task from the task buffer". Similarly, 'doubt' performs the function of decreasing "the confidence of a belief" (Wang, 2013, p.188). If we compare how the NARS based AI system will perform the function of 'observe' and how the human self carries out 'observation', the difference will be clearly self-evident. Both the actual process of observing; its layers, types and methods and what it means to the human 'self' are of another order, even when we exclude self-observation. The reach of human observation extends from one end of eternity to another and includes observation of both tangible manifest and intangible un-manifest realities. And the tools and methods it uses are also diverse and sophisticated including empirical tools. So it would be a bit unfair to compare this with the above mentioned 'observe' function performed by a NARS based system. Similarly, the mental process of 'doubt' has innumerable nuances, aspects, and characteristics within the human context. It is not a linear, reductive 'operation' as depicted in NARS.

We share an interesting example of the realistic stage at which the process of self-awareness and concept of 'self' are in today's AI systems. At present, an artificial agent's perception and understanding of its physical or mental 'self' exists in the form of a 'term' that is built onto its source code and basic language. In NARS and probably other architectures, the system's 'self' is represented as the term '{SELF}'. So if the following input is given to it: "I am a robot." then the system translates it in its own language as: <{SELF}-- > robot>. (Wang *et al.,* 2018, p.9). So the in-built equation that 'I' means {SELF} is the level at which the system recognizes itself. This is the

basic and actual level of its self-awareness. The inferences it can draw about itself or its actions can be seen from the following example:

Input: "I wonder whether cats are animals."

<(*,{SELF},<cat --> animal>) --> ^wonder>.

Input: "What am I?"

<{SELF} --> ?1>?

Answer: "I am somebody who wonders whether cats

are animals."

<{SELF} --> (/,^wonder,_,<cat --> animal>)>.

%1.00;0.90% (Wang *et al.*, 2018, p. 9)

One can see from the above example the linear inference drawing depending upon the priority values that the system attaches to the input statements. And the truth value of the inference/answer is depicted through the numerical symbols at the end representing 'frequency' and 'confidence'. We know that human intellect, on the other hand, based on its huge fund of knowledge (conceptual and non-conceptual, verbal and nonverbal) can draw very different, non-linear, and complex inferences and implications from any types of input statements in order to intelligently understand, intervene and also create new knowledge. And it can even revise and modify the input statements. And we have practical evidence of how it has created new knowledge and understanding through this process. Of course, in all fairness to NARS, it is not claiming to do anything like this. But AGI research, in general, is aiming for human level or better than human level intellectual capability. Our comparison should be seen as an affirmation of the challenge which AI research already recognizes.

The problem is that any sophistication and nuances of mental aspects, functions that we find in such architectures are, in fact, a reflection of human aspiration and creative capability than the actual 'mental' capability of the AI system. Because in actual reality the developed sensitivity and advanced capabilities of intelligence, intellect and creativity that we want to see in AI systems exist, as of now, only in our minds and not on the ground. The ground reality is more of an exaggeration of whatever progress has been achieved so far in computer science, IT and AI, which no doubt is huge compared to where we were even 25 years ago. So we are not in any way undermining the impressive hitherto progress in this area but just cautioning against insisting on seeing something that does not exist at present. If it becomes a concrete reality from the pool of potentials and possibilities, then for sure it must be hailed and celebrated.

A Possibility of Emulation in the Area of Hermeneutics

When we look at a painting, a poem, a piece of literature, or simply come across an idea that our mind analyzes and interprets, the "act" of interpretation too is a "construction." We hypothesize that "whatever is once constructed can also be 'copied'." We will try to demonstrate this aspect through discussion on hermeneutics of *ṣan'at-e ihām* (construction of ambiguity) in Urdu, Persian, Arabic, Hindi, and Punjabi poetry. It is hoped that the question whether such creative expression, its interpretation, and its understanding might be simulated or mimicked to some extent by artificial intelligence and with what scope and limitations will be addressed to a meaningful extent.

In ancient Greek texts as in Aristotle's logical treatises titled *Peri Hermeneias* ('On Interpretation'),[3] the noun *hermeneia* ('interpretation') and the verb *hermeneuein* are related to the concept of Hermes who transmitted Divine messages to mortals and also rendered these messages intelligible and meaningful. An important dimension of his task was explanation to make mortals understand the messages.[4]

Generally speaking, hermeneutics is the "theory or philosophy of the interpretation of meaning."[5] Our particular focus is on its philological aspects with literary texts as the object of interpretation particularly in relation to "word" in the language and context of each age.[6] Plato in *Cratylus* and Socrates in *Dialogue* assert how Hermes' tasks have to do with language and interpretation, and as an extension of the meaning of the very text, translation of a foreign language text too is an act of interpretation.[7] In the sense of providing the reader with pertinent substitutes, idioms, phrasal verbs, juxtapositions, and even new formations, the act of translation is an act of creativity. In Islamic interpretive traditions, these themes and theories were discussed in such disciplines as *uṣūl al-tafsīr, uṣūl al-Fiqh, 'ilm al-waḍa'* (roughly, semantics) and *'ilm al-balāghah* (the science of rhetoric) *etc.* Schleiermacher onward, there started to emerge certain philosophical themes which later on developed through Dilthey, Heidegger, Gadamer, Apel, Habermas, Hirsch, Bultman, Ricouer, et al. to build it up as a distinct philosophical tradition.

Ancient Greeks were not aware of hermeneutics as we know it today but had a hermeneutical approach to drama and poetry. Aristotle distinguished literary forms and recognized rhythm, period, metaphor, etc.[8] But technically speaking, according to Palmer, the

[3] In Aristotle's works, hermeneutics was in the domain of logic. In later Enlightenment, it applied to interpretation in a much wider sense.
[4] See (Palmer, 1969).
[5] See (Bleicher, 1980).
[6] See (Palmer, 1969).
[7] See (Palmer, 1969).
[8] See (Worthington, 2007).

oldest and "the most widespread understanding of the word hermeneutics refers to the principles of biblical interpretation" based upon the distinction of biblical exegesis as mere interpretation from hermeneutics as approaches to interpretation characterized by certain rules, methods and theories. Throughout the medieval era, two methods were commonly used in interpreting the Bible namely: grammatical-historical and allegorical.

The grammatical-historical method was used in interpreting the Old Testament in light of the New Testament and vice versa. One could interpret certain passages of the New Testament vis-à-vis passages of the Old Testament. Allegorical interpretation approached (particularly gnostic meanings of) text through allegories and metaphors. When Protestant reformers rejected Catholic authority, they placed emphasis on the sufficiency of text.[9] They also emphasized internal coherence and that text be interpreted in its own internal context.[10] Schleiermacher is considered to be the founder of the modern tradition of hermeneutics through which a text could be interpreted. The Schleiermacherian approach was influenced by Enlightenment philosophers, seeking to systematize knowledge, and the Romantic tradition.

Among Enlightenment philosophers, Chladenius (d. 1759) viewed hermeneutics as "the art of attaining the perfect or complete understanding of utterances, whether they be speeches (*Reden*) or writings (*Schriften*)." "Intentionality" of the author is to be grasped as neither an expression of the author's personality nor his psychological state of mind; rather as what emanates from the text in the specific genre of writing with the assumption that rules of reason remain unchangeable. These "rules," therefore, "guarantee the stability of meaning and the possibility of its objective transfer through verbal expressions." If a text was constructed in accordance with "appropriate rules of discourse" and the ideas were presented clearly by the author, "his words on the page would give rise to a correct and perfect understanding: author and reader alike shared in the same rational principles." Despite this, however, relativity of perspective for Chladenius could create contraries but not necessarily contradictories as the same historical event could still be interpreted differently by two different historians.

Among the Romantic thinkers, Friedrich Ast (d. 1841), a philologist whose major work *Grundlinien der Grammatik, Hermeneutik und Kritik* (Basic Elements of Grammar, Hermeneutics and Criticism) was used by Schleiermacher, discussed aspects of his philology in what he called "the hermeneutical circle, the relation of the part to the whole, the metaphysics of genius or individuality". Philology, to him, is not only a grammatical style of a work, rather its "basic aim is

[9] Flacius Illyricus is one of the most important Protestant theorists who laid the foundation of Protestant hermeneutics.

[10] This could be compared with the Farāhī school in Islamic Tafsīr tradition.

grasping the spirit (geist)' of the age: the outer and inner context of a work as a unity." The inner unity is the relation of various parts of a work while the outer unity is the spirit of the age. When a reader confronts a text, he not only understands the meaning of the words but also grasps the spirit of the author as well as the spirit of the age in which the text was written. Hermeneutics for Ast 'is the theory of extracting the geistige (spiritual) meaning of the text." According to him, one can grasp the overall spirit of a past text as a whole and can also fathom Geist of an individual author in relation to the whole. In his view, hermeneutics seeks to clarify the relationship of inner parts of a text to each other and to the larger spirit of the age. So, Ast believes that hermeneutics may be historical, philological, or spiritual (geistige) in its approach to a text. In historical hermeneutics, a text is to be understood 'in relation to the content of the work'. In grammatical hermeneutics, a text is to be understood 'in relation to the language', and in geistige hermeneutics, a text is to be understood 'in relation to the total view of the author and the total view of the age'. Two Enlightenment thinkers, Semler and Ernesti, had already developed the first two respectively, but the third one was an original contribution of Ast to the rise of general hermeneutics, and it is this type of hermeneutics, which was further developed by Schleiermacher.[11]

To Schleiermacher, hermeneutics deals with the possibilities of understanding and is also in a psychological sense part of the art of thinking. It involves attempts at understanding philologically what is said, and at understanding intentionality in the speaker's/author's thought. The interpreter, therefore, should be competent in linguistics and psychology. For this purpose, it will be important to understand the language common to the speaker/author and the original addressees, and to understand the context in which a statement occurred. In attempts to decipher an author's linguistic sphere, an interpreter may have to resolve issues and reach a level of understanding that is sometimes not even available to the author himself.

According to the second canon, a passage in which a word occurs constitutes a 'determinative linguistic sphere' as a context within which the meaning of the word is to be determined. Likewise, the whole of the text is a context in which a passage of it can be understood. For that, when a single passage is not enough to decipher the context, "one must turn to other passages where these same words occur, and under certain conditions, to other works of the author or even to works written by others in which these words appear. But one must always remain within the same linguistic sphere." We shall refer to this aspect as "intertextuality," and, to delve into this aspect, the interpreter resorts to his intuition as well as to his technical skill for the

[11] See (Rush, 2019)

comparison of text with other texts in the same linguistic sphere or in the same language tradition.

Despite departures from this approach where "text" retains its primacy even in interpretation, other hermeneuticians as Dilthey, Heidegger, Gadamer, et al. owe Schleiermacher much in developing their own views. Newer methods of interpretation, such as reader-response, feminist criticism, ideological criticism and postcolonial criticism, involved historical criticism that was predicated on, inter alia, a suspicion towards orthodox narratives and the supposition that through historical analyses, one may separate the "reliable" from the "unreliable" by deciphering historical agendas around texts and interpretations. Certitude in interpretation and deciphering of authorial intentionality (assumed as elusive or non-decipherable) became more problematic in reader-response and post-modernist theories that pointed up that every reader interprets the text in his or her own way. Similarly, feminists, post-colonialists, et al. expressed their apprehension of patriarchal biases or racial supremacists in interpretations that marginalized the oppressed.

For example, German scholar Hans-Georg Gadamer in his *Truth and Method* explained how ideas, tastes, and axioms impact our interpretation and that tradition gives us the historical "horizon" in the "range of vision", which grows in interaction with other horizons. The interaction, then, could augment historical and cultural "prejudices" we bring to our interpretation. Such relativism seemed to undermine the very concept of meaning itself. Gadamer points out that the criteria of "interpretive communities" too are shaped by developing history and culture.

In Urdu poetry, much has been written on the concept of *balāghat* (rhetoric) to help the reader decipher the intertextuality, irony, *tashbīh* (simile), and *talmīḥ* (allusion) in a certain verse or a couplet. For example, when Amrita Pritam writes in Punjabi:

Aj ākhāṇ vāriṣ shāh nūṇ, kithoṇ qabrāṇ vichoṇ bol
Tay aj kitāb-e 'ishq dā ko'ī aglā varqah phol

Ek ro'ī sī dhī Punjāb dī, tūṇ likh likh māre ben
Aj lakhāṇ dhīyāṇ rondiyāṇ, tenūṇ vāriṣ shāh nūṇ kehn

The reader who is aware of the tradition of the Punjab would immediately understand the intertextuality within the mentioned verses. The word "*dhī*" refers to *Heer*, "*ben*" refers to Waris Shāh's compiled work on one woman's story, and "*lakhāṇ dhīyāṇ*" would then refer to those hundreds of women who were raped and murdered during the partition of 1947. These nuances are decipherable when the reader is well-versed with the ethos, the language, the culture, and the history. Several books have been written on *balāghat* to understand and explain these nuances.

Ṣan'at-e ihām is another important principle which introduces ambiguity of meaning in poetry to create various translations, interpretations, and *mughāzlah* (flirtatious overtures). If a verse is constructed following the rules of rhetoric, it is possible for the reader to fully grasp and appreciate the multiple interpretations, given that they are aware of the ethos and the linguistic principles.

Take the example of the following couplet of *Ghālib*:

Ghunchah-e nā shuguftah ko dūr se mat dikhā keh yūṇ
Bosah ko pūchtā huṇ meṇ munh se mujhe batā keh yūṇ

Don't show from afar an unopened bud-- 'like this'
I ask about/for a kiss-- tell me with your mouth: 'like this[12]

The possibilities of meanings in this couplet could be: do not show the 'unopened bud' from a distance, show it to me from near with your lips (by smiling or it might even apply to kissing); Or don't show me a kiss through an unopened bud from a distance, rather come near me and show me with your mouth. The couplet shows the flirtatious overture of the poet and is enjoyed by the reader because of a certain mischievousness and pun that create the possibility of varying interpretations.

To further explain the idea, we can have a look at a few couplets of this author so as to eliminate the need for advancing an argument to settle authorial intentionality and interpretational disagreements. Take, for example, the following couplets:

Jab keh har pal sitam nahī hotā
Dard kyuṇ thorā kam nahī hotā

Dil ko phir terā khyāl āyā
Dawr aesā kam nahī hotā

When oppression does not happen all the time
Why does pain not subside a bit

Thought of you came to heart again
This cycle never does it end

The second couplet can be interpreted as either posing a question: would this cycle of your thought coming to the heart never end? Or it could be a statement that this cycle of remembering the beloved never ends.

Following is another example:

Ranj kī sun kabhī kahānī to
Multafit yuṇ sanam nahī hotā

[12] Translation taken from Rekhta.org
https://www.rekhta.org/ghazals/guncha-e-naa-shagufta-ko-duur-se-mat-dikhaa-ki-yuun-mirza-ghalib-ghazals?lang=ur

At some point hear the woeful story
Grants attention this way the idol not

To explore the possibilities of meaning, and interpret this couplet, we will first speculate about the intended addressee. There is ambiguity here because it could either be the beloved, God, a friend, or *nāṣiḥ* (adviser/counsellor), or even the *nāṣiḥ* counselling the poet. If we start making combinations, a number of interpretive possibilities can be deciphered. If it is the beloved, the verse could be interpreted to mean that the poet is saying to the beloved: hear my story of pain, O my beloved! Is there no possibility of you being attentive to me? The beloved in this case is addressed as *ṣanam*, which can be translated as stone-hearted, referring to their inability to empathize or feel anything towards the poet.

Another interpretation could be that *nāṣiḥ* is telling the poet that his stone-hearted beloved could not be sympathetic towards him. In this case, the first verse could be from the poet's perspective, sharing his grievance over or remonstrating about not being heard, and the second would be the response from the *nāṣiḥ*. The second verse could also be the conclusion drawn by the poet himself that his beloved would never be attentive to him.

If referring to God, the first verse could mean that the poet is beseeching God to listen to his prayer and bless him with what he has been asking for. The second would then mean that idols do not respond but God, being the ultimate being, does. The second verse could also be interpreted to mean that the the stone-hearted beloved is apathetic towards the poet and cannot, or does not, respond the way God can.

There is also an emotive appeal attached with the couplet. If you ask the poet what meaning he had in mind while writing the couplet, it is possible that he had one or two, while the audience derives three to four or more meanings. But there is also a possibility that the author had four different interpretations in mind and all four of them were indeed understood by the audience. Even if the couplet was created at a particular point in time, the audience at some other point could still decipher the intended meaning and the varying interpretations if there is awareness and appreciation of the language, the culture, and the over-all ethos.

To take the example of another couplet:

Karb kesā milā yahāṇ ham ko

Qaṣr Shāh kā eram nahī hotā

O what torment did I get here
King's palace is paradise not

One simple meaning could be that the poet has lived a life of such torment that even all the luxuries of life cannot bring him comfort. The

second hemistich could also be interpreted as the poet posing a question: why could the luxuries of life not assuage his misery or whether all the luxuries of life possess the possibility of bringing peace to such a life of torment?

Let us consider some more couplets from another ghazal of the author:

Ham rahe muẓtarib maṣāeb se

Ham nafas muntaṣib ʻajāeb se

Restless with difficulty I was
Related to soul mate fancy

One interpretation can be that the poet spent a life of misery and restlessness, but his soul mate was engrossed in worldly glory and material gain. Another interpretation could be that the poet himself was consumed by his soul mate's insignificant materialistic concerns.

Kis taraḥ āgeyā khayāloṇ meṇ

Wo tasavur keh tum ho tāib se

How in thoughts did come
That image that you are sort of repentant

This couplet can be taken as the poet questioning himself as to how the thought occurred to him that the beloved would seek redemption for their carelessness or that they would be ashamed of their misbehavior. Or perhaps the poet is wondering how the thought of repentance came to his beloved.

The above-mentioned examples portray that once an art form is constructed creatively within the ambit of certain principles, a commonality originates between the creator and the audience, giving birth to the possibility of complete understanding of the author's narrative. It is worth mentioning that prosody in Urdu is already well-developed and the measurement of syllables in the various metrical patterns is a very mathematical process. Sometimes, mistakes can be made, and human intelligence is required to rectify them, but serious work has been done already to calculate the metrical pattern of a verse through the use of algorithms.[13]

By way of example, for appreciating *Ṣanʻat-e ihām*, at a very fundamental level, we can formulate some principles, such as one that deals with the identification of the addresser and the addressee, and teach an intelligence, artificial or otherwise, how the meaning changes based on this identification. If the addresser and addressee are the same, the meaning would be different from what it would be if both are different. This principle becomes one way of engaging with ambiguity and could be applied to both hemistiches of a couplet.

[13] For reference, please see https://www.aruuz.com/taqti

A second principle relates to the concepts of takhṣīṣ and ta'mīm, specification and generalization respectively, whereby one could state something either very specifically or in a general sense. At times, a specific noun is used but the intention, understood through context, is to convey a general idea.

Another principle could be synecdoche (when a part of something refers to the whole of it or vice versa). For example, when you say, "I have come to ask for your daughter's hand", it should be understood that you are asking for a person not just the hand.

Through these and other principles, it becomes possible to narrow down the possibilities of interpretation, and one can perhaps argue that when these aspects of creative expression are enumerated and made decipherable, there may be a higher probability of artificial intelligence reliably narrowing down the interpretations. It is perhaps true that the pathos and the pain in poetry is something that can only be felt and experienced and cannot be conveyed through formulaic rules of representation. But by employing patterns and principles, experts in the domain of artificial intelligence may take a step closer to simulating and mimicking human understanding.

Concluding Comments

The aim of AI researchers, thinkers, and practitioners should be to keep improving and upgrading the specific AI systems and keep working towards more integrated, capable, and trustworthy systems which can get closer to human mental functioning and collaborate more effectively with humans. AI was neither conceived of nor should it be viewed as a substitute for human emotional and conscious functioning. In fact, the area of complex emotional and conscious functioning, nuanced interpretation and understanding and the sense of 'self' are domains which might remain beyond the pale of artificial emulation for quite some time to come, if not forever.

Besides the area of emotions and consciousness in humans is a two-edged sword. On the one hand, we find feelings and emotions of empathy, concern, care, happiness, love, which can produce constructive progress, and collaboration and on the other, anger/rage, resentment, and other adversarial and destructive emotions which can lead to serious harm and destruction. Hence we must be vigilant and careful with emulating and installing this capability in artificial systems. There must be a clear and rational system of channelizing the research and application works happening in this area along with a system of transparency and checks and balances. What we are saying is in line with the extensive work happening on AI governance, ethics, and safety. It is a response to what AI can do if it develops consciousness, emotions, understanding and a sense of an autonomous agency and self.

To sum up, this research work introduces some new ideas, thinking and understanding about human emotional, experiential and conscious processes, human personality, and the complex mental functions of subtle interpretation and understanding, which can be of use and interest to researchers, thinkers, AI developers and practitioners, working in this area. It hopes to add to the present knowledge and understanding of these highly important and critical areas of human mental functioning that have a crucial role to play in the making of present and future AI systems which hope to mimic human mental functioning and smoothly integrate with the human world.

Acknowledgments

This research has been supported and sponsored by Prince Mohammad Bin Fahd Center for Futuristic Studies (PMFCFS) and World Futures Studies Federation (WSFS).

Conflict of interest statement

None declared.

References

Ahvenharju S, Minkkinen M, Lalot F. The Five Dimensions of Futures Consciousness. Futures 2018; 104(12): 1-13.

Asada M. Artificial Pain: empathy, morality, and ethics as a developmental process of consciousness. http://ceur-ws.org/Vol-2287/paper19.pdf Accessed date: Mar 21, 2022

Bleicher J. Contemporary Hermeneutics, London: Routledge & Kegan Paul, 1980.

Conroy G, Jia H, Plackett B, Tay A. Six researchers who are shaping the future of artificial intelligence. https://www.nature.com/articles/d41586-020-03411-0 Accessed date: March 21, 2022

Conway M. [in press]. Exploring the Links between Neuroscience and Foresight. https://jfsdigital.org/exploring-the-links-between-neuroscience-and-foresight/ Accessed date: March 21, 2022

Crosby M. Why Artificial Consciousness Matters. http://ceur-ws.org/Vol-2287/paper13.pdf Accessed date: Mar 22, 2022

Damasio A. Self comes to mind: constructing the conscious brain. New York: Pantheon Books, 2010.

DeLancey C. Passionate Engines: What Emotions Reveal About Mind and Artificial Intelligence, New York: Oxford University Press Inc., 2002.

Dreyfus, H. L. What computers can't do: a critique of artificial reason, New York: Harper & Row Publishers, 1972.

Fortenbaugh WW. Aristotle's Art of Rhetoric, in Worthington, I (ed.) A Companion to Greek Rhetoric, Oxford: Blackwell Publishing, 2007, pp. 107-123.

Gamez D. Four Preconditions for Solving MC4 Machine Consciousness. http://ceur-ws.org/Vol-2287/paper6.pdf Accessed date: Mar 22, 2022

Garfinkel SL and Grunspan RH. The Computer Book: From the Abacus to Artificial Intelligence, 250 Milestones in the History of Computer Science. New York: Sterling, 2018.

Goertzel B and Pennachin C. The Novamente Artificial Intelligence Engine. In Goertzel B and Pennachin C. (eds.), Artificial General Intelligence. Heidelberg: Springer-Verlag, 2007, pp. 63-129.

Hafner VV, Loviken P, Villalpando AP, Schillaci G. Prerequisites for an Artificial Self. https://www.frontiersin.org/articles/10.3389/fnbot.2020.00005/full Accessed date: Apr 11, 2022

Hammer P and Lofthouse T. 'OpenNARS for Applications': Architecture and Control. In Goertzel B, Panov, A. I., Potapov A, Yampolskiy R. (eds.) Proc. Artificial General Intelligence: 13th International Conference, AGI 2020. Cham: Springer Nature, 2020, pp. 193–204.

Huang JY, Lee WP, Chen CC, Dong BW. Developing Emotion-Aware Human–Robot Dialogues for Domain-Specific and Goal-Oriented Tasks. Robotics 2020; 9(2), 31: 1-20.

Lavelle S. The Machine with a Human Face: From Artificial Intelligence to Artificial Sentience. https://link.springer.com/content/pdf/10.1007%2F978-3-030-49165-9_6.pdf Accessed date: Mar 30, 2022

Lombardo T. Creativity, Wisdom, and Our Evolutionary Future. https://jfsdigital.org/wp-content/uploads/2014/01/161-A02.pdf Accessed date: Mar 30, 2022

Looks M, Goertzel B, Pennachin C. Novamente: An Integrative Architecture for General Intelligence. https://www.aaai.org/Papers/Symposia/Fall/2004/FS-04-01/FS04-01-010.pdf
Accessed date: Mar 30, 2022

Mangan B. Sensation's Ghost: The Non-Sensory "Fringe" of Consciousness. http://journalpsyche.org/files/0xaa9b.pdf Accessed date: Mar 30, 2022

McCorduck P. Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence. Natick, MA: A K Peters Ltd., 2004.

Nilsson NJ. The Quest for Artificial Intelligence: A History of Ideas and Achievements. Cambridge University Press, 2010.

Ozkural E. Omega: An Architecture for AI Unification. In Goertzel B, Panov, A. I., Potapov A, Yampolskiy R. (eds.) Proc. Artificial General Intelligence: 13th International Conference, AGI 2020. Cham: Springer Nature, 2020, pp. 267–278.

Palmer RE. Hermeneutics: Interpretation Theory in Schleiermacher, Dilthey, Heidegger, and Gadamer. Evanston: Northwestern University Press, 1969.

Picard RW. Toward Machines with Emotional Intelligence. https://dam-prod.media.mit.edu/x/files/pdfs/07.picard-EI-chapter.pdf
Accessed date: Mar 30, 2022

Pipitone A, Lanza F, Seidita V, Chella A. Inner Speech for a Self-Conscious Robot. http://ceur-ws.org/Vol-2287/paper14.pdf Accessed date: Mar 30, 2022

Rush F. Hermeneutics and Romanticism. In Forster M. and Gjesdal K. (eds.), The Cambridge Companion to Hermeneutics. Cambridge University Press, 2019, pp. 65-86.

Sahner D. Evolution of Conscious AI in the Hive: Outline of a Rationale and Framework for Study. http://ceur-ws.org/Vol-2287/paper1.pdf
Accessed date: Mar 30, 2022

Schweizer P. Artificial Brains and Hybrid Minds. In Müller, V. C. (Ed.), Philosophy and Theory of Artificial Intelligence 2017, Cham: Springer Nature, 2018, pp. 81-91.

Somers M. Emotion AI, explained. https://mitsloan.mit.edu/ideas-made-to-matter/emotion-ai-explained
Accessed date: Mar 30, 2022

Sloman A. Damasio's error.
https://www.cs.bham.ac.uk/research/projects/cogaff/phil-mag-emotions-sloman.pdf Accessed date: Mar 30, 2022

Sloman A. Huge, but Unnoticed, Gaps Between Current AI and Natural Intelligence. In Muller, V.C. (Ed.), Philosophy and Theory of Artificial Intelligence 2017, Cham: Springer Nature, 2018, pp. 92-105.

Sloman A. Varieties of Evolved Forms Of Consciousness, Including Mathematical Consciousness. https://www.mdpi.com/1099-4300/22/6/615/htm Accessed date: Mar 30, 2022

Tariq S, Kazim R, Tauqir I. How Mental Genes Make the Human Mental Complex and Control Its Functioning: Transition From The Genetic Mind to the Intelligent Mind. https://www.neuroquantology.com/data-cms/articles/20191024022735pm290.pdf Accessed date: Mar 30, 2022

Wang P. The Logic of Intelligence. In Goertzel B, and Pennachin C. (eds.), Artificial General Intelligence. Heidelberg: Springer-Verlag, 2007, pp. 31-62.

Wang P. Non-Axiomatic Logic: A Model of Intelligent Reasoning. Hackensack NJ: World Scientific Publishing Co. Pte. Ltd., 2013.

Wang P, Li X, Hammer P. Self in NARS, an AGI System. https://www.frontiersin.org/articles/10.3389/frobt.2018.00020/full Accessed date: Mar 30, 2022

Zaadnoordijk L and Besold TR. Artificial Phenomenology for Human-Level Artificial Intelligence. http://ceur-ws.org/Vol-2287/paper24.pdf Accessed date: Mar 30, 2022