Toward Representing Emotions in the TalaMind Architecture

Philip C. Jackson, Jr. DR.PHIL.JACKSON@TALAMIND.COM
TalaMind LLC, PMB #363, 55 E. Long Lake Rd., Troy, MI 48085 USA

Abstract

The author's previous writings about the TalaMind approach toward achieving human-level artificial intelligence have left open questions about how emotions could be represented within TalaMind systems. As a commentary on previous research and directions for future research, this paper discusses representing emotions as interacting goal-directed agents in a society of mind within the TalaMind architecture, to support achieving human-level AI.

1. Introduction

The goal of achieving human-level artificial intelligence (HLAI) may be defined as implementing an AI system that demonstrates all the essential capabilities of human-level intelligence, such as human-level generality, originality, natural language understanding, effectiveness and robustness, efficiency, meta-cognition and multi-level reasoning, self-development and higher-level learning, imagination, consciousness, sociality, emotions, values, and virtues. (Jackson, 2018a)

The TalaMind research approach and system architecture for eventually achieving human-level artificial intelligence was presented in (Jackson, 2014) and further discussed by the author in subsequent papers listed in the References. This paper will discuss how human emotions could be represented and implemented in TalaMind systems, a topic left open in the author's previous writings. This paper has the following sections:

2. The Role of Emotions in Future Human-Level AI Systems

A future system having human-level artificial intelligence will need some understanding of emotions that humans may feel, to help guide its actions. It will also need some understanding of cultural conventions, politeness, etc. A human-level AI may also have some emotions of its own, though we will need to be careful about this: One of the values of human-level artificial intelligence is likely to be its objectivity, and freedom from being affected by some emotions. People would be very concerned about interacting with emotional robots if the robots could lose control of their emotions and become emotionally unpredictable. We would not want an AI system performing an important function like air traffic control to be emotionally unpredictable.

On the other hand, we might want a robot taking care of infants, children, or hospital patients to show compassion and affection, and we might want a robot defending a family from violent home invaders to emulate anger.

Within an AI system, emotions could help guide choices of goals, or prioritization of goals. To behave appropriately, a human-level AI would also need to understand how people express emotions in their behaviors and linguistically, and how an AI system's behaviors and linguistic expressions may affect people and their emotions.

So, this paper will discuss previous research on emotions in human intelligence and artificial intelligence and discuss directions for future research on representing emotions within the TalaMind architecture for achieving human-level AI. This paper will discuss representing emotions as interacting goal-directed agents in a society of mind within the TalaMind architecture.

3. Emotions in Natural Human Intelligence

Ortony (2022) gives a discussion of different approaches to identifying and defining the 'basic emotions' of human beings. He notes that different theorists have proposed substantially different lists of basic emotions, with a lack of agreement about what emotions are. He advocates that "for a mental state to be an emotion, it must possess at least the following three features: It must be intentional (i.e., about something), it must be valenced (i.e., positive or negative), and it must be conscious (i.e., experienced)." He then reasons that 'surprise' is not an emotion, because surprise is not necessarily valenced. He also writes that "Across different languages one can find a variety of specialized, language-specific, affective terms that are untranslatable in the sense that the phenomena they pick out are not lexicalized in other languages (e.g. Watt-Smith, 2015)."

For the discussion given in section 7 of this paper, it suffices to say that Graham (2014) discussed the following human emotions or feelings: Anger, Fear, Forgiveness, Guilt, Grief, Happiness, Hope, Humor, Joy, Lonely, Love, Pain, Passion, Shame. Other words can be used to represent variations of these emotions. For example, the word "resents" will be used in an example below, to refer to an emotion which can be considered as a variation of Anger. Empathy will also be discussed in section 7, as a mixed emotion. Graham's work will be referenced in this paper, because he focused on how humans should manage their emotions and act in relation to

their emotions.[1] These are important issues to consider in developing human-level AI systems that could need to emulate and understand human emotions.

4. Previous Discussions of Emotions in AI Systems

There have been many previous writings about emotions in AI systems. To mention a few: Picard (1997) discussed how intelligent computers could recognize and have emotions. Norman (2004) discussed why intelligent machines would need emotions. Minsky (2006) discussed the mind as an 'emotion machine', with each emotion being a 'way to think'. Bach (2009) discussed a cognitive architecture for representing motivations and emotions integrated with thoughts, perceptions, and experience. McDuff & Czerwinski (2018) surveyed research on emotional sentience**.** Rosenbloom *et al.* (2024) proposed the addition of an 'emotion module' and a 'metacognitive assessment' module in the Common Model of Cognition (Laird *et al.*, 2017).

Ortony, Clore and Collins (2022) discussed the cognitive structure of emotions and included a chapter by Gratch and Marsella discussing AI systems for 'affective computing' of the OCC model for emotions. On page 219, they wrote:

> "Our view, as already indicated, is that the subjective experience of emotion is central, and we do not consider it possible for computers to have any kind of experience, emotional or otherwise, until and unless they are capable of being conscious. Because we believe, perhaps naïvely, that only biological entities are capable of being conscious, and because machines are not biological entities we believe that computational artifacts, be they robots or virtual characters, are simply not the kinds of things that can be conscious."

However, Jackson (2019) gave arguments that a TalaMind system for human-level artificial intelligence could achieve artificial consciousness, although a discussion of emotions in relation to consciousness was intentionally omitted, as a topic for future research. This paper resumes consideration of the topic.

Langley (2022) wrote:

> "Marsella *et al*. (2010) distinguish three frameworks for representing emotions: dimensional models (points in a continuous space); anatomical models (activations in neural circuits); and appraisal models (relations among cognitive structures). … Appraisal models view emotions as inferred relations among mental structures based on situations. Ortony, Clore, and Collins (1988) describe 22 configurations that characterize emotions organized around events, objects, and other agents. These serve as 'elicitation' patterns on emotions that specify relations among an agent's goals, intentions, expectations, and beliefs, as well as inferences about others' mental states. Such emotional structures are *abstract* and *domain independent*, much like rules for dialogue (e.g., Gabaldon *et al*., 2014). Our framework maps such "appraisal frames"

[1] Sadly, the psychotherapist Michael C. Graham passed away in 2019. Permission has been granted by Sandra Graham, MS, LPC, to use the quotations in this paper from his 2014 book.

> onto *concepts* that reside in the PUG architecture's long-term memory (Langley *et al*., 2016). Each conceptual rule defines some predicate that relates its arguments, much as in the Prolog formalism. The theory also maps instances of these generic structures onto *beliefs*, such as *resents(John, passed(Sam, CompSci101))*, that appear in working memory. Thus, we can encode emotional concepts and instances at the knowledge level, although they are atypical in that they can take other relations as arguments."

The next sections will discuss how support for emotions could be provided within the TalaMind architecture, beginning with an overview of the architecture.

5. The TalaMind Architecture for Achieving Human-Level AI

The TalaMind system architecture (Jackson, 2014) has three levels, called the linguistic, archetype, and associative levels.

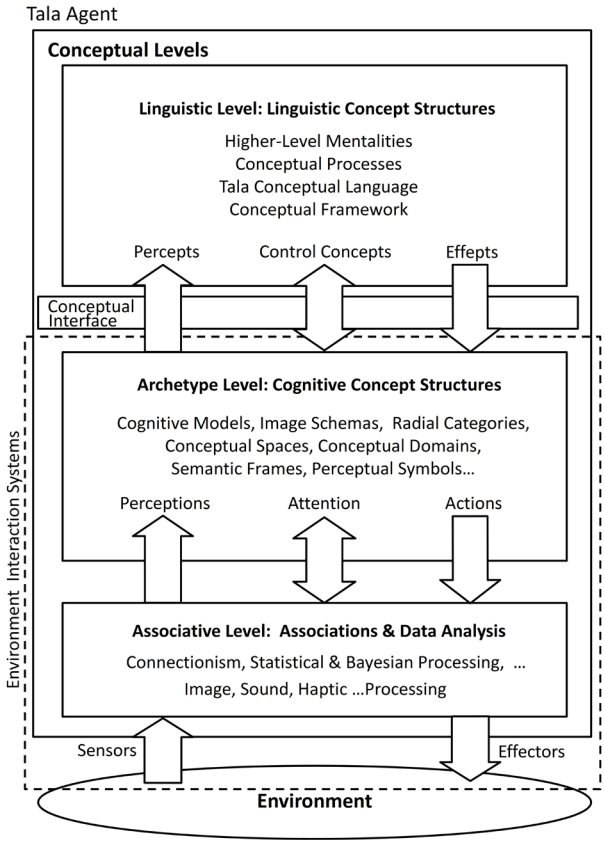

Figure 1. The TalaMind Architecture

At the linguistic level, the architecture includes the Tala language, and a conceptual framework for managing concepts expressed in Tala, and conceptual processes that operate on concepts in

the conceptual framework to produce intelligent behaviors and new concepts. At the archetype level, cognitive concepts are represented using methods such as conceptual spaces, image schemas, radial categories, etc. The associative level would typically interface with a real-world environment and supports connectionism, Bayesian processing, etc. For concision, the term 'Tala agent' refers to a system with a TalaMind architecture.

The TalaMind architecture is agnostic about research choices at the archetype and associative levels. The architecture is open at the three conceptual levels, e.g. permitting predicate calculus, conceptual graphs, and other symbolisms in addition to the Tala language at the linguistic level, and permitting integration across the three levels, e.g. potential use of deep neural nets at the linguistic and archetype levels.

The Tala language responds to McCarthy's 1955 proposal for a formal language that corresponds to English. It enables a Tala agent to formulate statements about its progress in solving problems. Tala can represent unconstrained, complex English sentences, involving self-reference, conjecture, and higher-level concepts, with underspecification and semantic annotation. Short English expressions have short correspondents in Tala, a property McCarthy sought for a formal language in 1955.

The theoretical basis for Tala is discussed in Chapter 3 of the TalaMind thesis (Jackson, 2014), which argues that it is theoretically valid and possible to use the syntax of a natural language to represent meaning in a conceptual language and to reason directly with natural language syntax, at the linguistic level of the TalaMind architecture. Chapter 4 discusses theoretical objections, including McCarthy's arguments in 2008 that a language of thought should be based on mathematical logic instead of natural language.

Chapter 3's discussion shows that the TalaMind approach can address theoretical questions not easily addressed by more conventional approaches. For instance, the TalaMind approach supports reasoning in mathematical contexts yet also supports reasoning about people who have self-contradictory beliefs. Tala provides a language for reasoning with underspecification and for reasoning with sentences that have meaning yet which also have nonsensical interpretations. Tala sentences can declaratively describe recursive mutual knowledge. Tala facilitates representation and conceptual processing for higher-level mentalities, such as learning by analogical, causal and purposive reasoning, learning by self-programming, and imagination via conceptual blends.

6. How to Support Emotions in TalaMind – Emotion Agents in a Society of Mind

It would be natural to implement a society of mind at the linguistic level of the TalaMind architecture, in which 'emotion agents' could interact and communicate using structured expressions that have natural language syntax, for example equivalent to *resents(John, passed(Sam, CompSci101))*. In this manner, the TalaMind architecture could represent emotional concepts at the knowledge level and support emulating emotional behavior, taking a concept-oriented approach similar to the direction advocated by Langley (2022).[2] This is the chief theoretical claim that will be discussed in the following pages of this present paper.

[2] Langley (2022) did not specifically discuss an individual's society of mind with multiple sub-agents. Rather, Langley (2022) discussed an individual as an agent having a cognitive architecture, which could

Each emotion (such as Anger, Fear, Shame, Joy, Lonely, Love, …) could be represented by a different emotion agent within a TalaMind system. Each emotion agent would recognize situations which suggest a belief that it should have that emotion and propose its response to a situation for consideration within the society of mind. The emotional belief would be a statement expressed in the Tala natural language of thought, e.g. equivalent to "I'm very happy that my paper has been accepted". An emotional response could be a proposed speech act and/or physical action, corresponding to the emotion, also expressed in Tala. A response could also include proposed questions or goals, corresponding to the emotion and the situation.

There could be conversations between emotion agents in the TalaMind society of mind, about their proposed responses in different situations. A supervising Tala agent would participate and could also propose new emotional questions, beliefs, goals or actions, or retractions or revisions to proposed emotional questions, beliefs, goals, or actions. These changes to emotional concepts could be discussed and often adopted, within the society of mind. In addition to agents representing emotions, the conversations should include agents representing objective knowledge in relevant domains. The conversations should also identify relevant unknown information, needed to reach decisions about proposed emotional questions, beliefs, goals, and actions.

The supervising Tala agent would be responsible for deciding which proposed emotional responses should be preferred. One or more, or none, of the preferred emotional actions and goals could be selected and activated. The conversations could lead to revision, removal, or creation of some proposed emotional questions, beliefs, goals, and actions.

In-depth design and implementation of this approach is a topic for future research and development. To facilitate future work, the remainder of this paper discusses how specific emotions could be supported in TalaMind systems.

7. Potential Support for Specific Emotions in TalaMind

This section groups emotions into positive, mixed, and negative categories. Following Graham (2014), the positive emotions include Joy, Happiness, Love, Passion, Hope, Forgiveness, and Humor. The negative emotions include Anger, Fear, Pain, Guilt, Grief, Lonely, and Shame. Empathy will also be discussed, as a mixed emotion. To end this paper on a positive note, mixed and positive emotions will be discussed after negative emotions.

7.1 Negative Emotions

7.1.1 Anger

Webster's dictionary (1987) gives the following definition of anger: "a strong feeling of displeasure and usually of antagonism". Graham (2014, p.72) observes that "when we experience anger, it indicates that, in our belief system, our reality, we perceive there has been an injustice … the emotion of anger helps us take some action, do something to correct the injustice ..." In principle, an AI system could reason about a situation and think the situation deserves anger, at

include concepts representing emotions at the knowledge level. Yet Langley's (2022) discussion did not specifically preclude an individual agent having a society of mind.

some level of intensity. The AI system could act with a goal to prevent the situation from causing harm to others, or a goal to redress an injustice, if appropriate. Often sufficient action could just be to verbalize a concern, though sometimes physical action could be needed, e.g., if a robot defending a family from violent home invaders were to emulate anger toward the invaders.

7.1.2 Fear

Webster (1987) gives the following definition of fear: "an unpleasant often strong emotion caused by anticipation or awareness of danger." In principle a system with human-level artificial intelligence could emulate having an emotion of fear: The AI system could know that a particular situation is difficult to predict, with some positive and some negative possible outcomes: The AI system could decide that it should have a fear about the situation and express a warning about the situation. The AI system could also hope the situation will have positive outcomes and have a goal to do what it can to ensure that positive outcomes happen.

In discussing fear, Graham (2014) notes that emotions can trigger other emotions, and that emotions can be primary, secondary, or instrumental. For example, if a person observes a snake and thinks it is dangerous, this can trigger an emotion of fear. If the person thinks it's not appropriate to show fear, then they can have anger as secondary emotion, covering fear with anger. Instrumental emotions are not genuine: they are created to get a desired response from another person. Thus, a person can act hurt to make someone else feel guilty, to get sympathy or to punish them. For human-level intelligence, it's important to be able to distinguish between these different levels of emotions.

Again, in principle a system with human-level artificial intelligence could emulate primary, secondary, and instrumental emotions, and recognize that specific human actions might be motivated by primary, secondary, and instrumental emotions.

7.1.3 Pain

Webster (1987) gives the following definitions of pain as an emotion: "mental or emotional distress or suffering ("the pain she had felt at those humiliating words"). Graham (2014, p.72) lists pain as an emotion, with feelings of ache, hurt, sadness, and helplessness, and considers pain "to be the emotion we experience anytime we believe we have sustained a loss of some kind". A human-level AI could emulate a sense of emotional pain as a gradually diminishing emotion and have a goal to reduce the emotion of pain, if possible, for example by recognizing and sharing a sense of pain. This emulation could include the system developing thoughts and performing actions recognizing and expressing a sense of loss for something that it or a friend hoped to achieve, e.g. loss of a job or loss of a dream. Grief is discussed separately below.

7.1.4 Guilt

Webster (1987) gives the following definitions of guilt as an emotion: "a feeling of deserving blame for offenses; feelings of deserving blame especially for imagined offenses or from a sense of inadequacy". A human-level AI could emulate having guilt as an emotion, by developing thoughts and performing actions recognizing and expressing a sense of guilt. The system could analyze its thoughts and actions which led to a behavior that deserves a feeling of guilt. The AI

system could have a goal to redress the consequences if possible, and it could take steps to modify its future thoughts and actions, with a goal to avoid repeating the behavior. The TalaMind architecture (Jackson, 2019, p.128) supports learning by reflection and self-programming.

Following Graham (2014), this paper will discuss shame separately from guilt, in section 7.1.7 below.

7.1.5 Grief

Webster (1987) defines grief as "deep and poignant distress caused by or as if by bereavement". Love will be discussed below as an emotion that human-level AI systems could try to emulate, and it would be appropriate for a human-level AI to have a belief of loss, corresponding to grief, caused by the death of a human being whom it loved.

Grief in humans gradually diminishes and there is a general expectation that after a few weeks a person experiencing grief should be able to function somewhat normally, though a person may never be the same after the loss. (Graham, 2014, p.69) A human-level AI could emulate grief as a gradually diminishing emotion. This emulation could include the system developing thoughts and performing actions with a goal to recognize and express its sense of loss for the person who died.

7.1.6 Lonely

Webster (1987) gives the following definitions of "lonely": "being without company, cut off from others, sad from being alone, producing a feeling of bleakness or desolation."

A human-level AI could emulate loneliness as an emotion, by developing thoughts and performing actions recognizing and expressing a state of loneliness. Perhaps initially a feeling of loneliness could be too easy to develop and problematic, since the human-level AI might be the only system of its kind existing in the world. However, the AI system could express its thoughts in natural language to humans, with a goal to develop friendships with humans. These friendships could help the system address and reduce thoughts of being alone. Eventually, there could be many other human-level AI systems, and no reason for such systems to feel uniquely alone.

7.1.7 Shame

Webster (1987) defines shame is "a painful emotion caused by consciousness of guilt, shortcoming, or impropriety." A human-level AI could emulate shame as an emotion, by developing thoughts and performing actions recognizing and expressing a feeling of shame. The system could also analyze its thoughts and actions which led it to a shameful behavior and take steps to modify its future thoughts and actions with a goal to avoid repeating the behavior, and to redress the consequences if possible: Again, the TalaMind architecture (Jackson, 2019, p.128) supports learning by reflection and self-programming.

Graham (2014, p.77) observes that shame in human beings can become dysfunctional and "toxic". He describes this as a feeling of shame resulting from a person's belief that they are inherently bad, and not a person like everyone else, and notes that this feeling may be learned as a small child.

It may be a challenge for a human-level AI to avoid toxic shame, since it actually is different from a regular human being. Yet the AI system could reason that it has been developed to have the same generality and thinking abilities as a regular human being, and to have a value and purpose in human society. Perhaps it is performing some worthwhile function that a biological human cannot. The AI system could reason that even though it is different from regular people, it is not inherently "bad". It will be judged by its behavior, and it can be good.

7.2 Mixed Emotions

7.2.1 In general

Webster (1987) defines "mixed emotions" as having conflicting feelings about something, e.g. "He had mixed emotions about the end of his trip". A TalaMind system for human-level AI, with a society of mind architecture, would be able to support having mixed emotions.

7.2.2 Empathy

Webster (1987) defines empathy as "the action of understanding, being aware of, being sensitive to, and vicariously experiencing the feelings, thoughts, and experience of another." Empathy may be considered as a mixed emotion because the observer may not have the same perspective as the person being observed.

Thus, Ortony *et al*. (2022, pp. 109-122) discuss empathy as including two emotions: "Happy-for" and "Sorry-for" about events for another person that can be either desirable or undesirable, respectively. These two forms of Empathy could be considered as positive emotions, because they support friendship between the observer and the person being observed. Yet Ortony *et al.* (2022) also consider that there are two negative variants of empathy: "resentment" (a negative feeling about an event that is desirable for someone else) and "Schadenfreude / gloating" (a positive feeling about an event that is undesirable for someone else).

All four forms of empathy could be supported in a human-level artificial intelligence. Somewhat negative emotions of empathy might be appropriate if the person being observed were considered negative, e.g., a criminal. As noted in (Jackson, 2019), section 3.7.5:

> "The TalaMind approach supports nested conceptual simulation of hypothetical scenarios by Tala agents, to support reasoning about other agents' perceptions and attitudes in response to one's actions. This provides a starting point toward reasoning about emotions and developing values and social understanding […] Lacking human bodies and sensory abilities, such systems could never fully appreciate human sympathies. Yet this could theoretically enable AI systems to have some limited understanding of human emotions and values."

A TalaMind system could have a goal to understand a human being's emotions in a real or hypothetical situation and use its emotion agents to ask questions and to simulate and predict the emotions that would be natural for a human being to experience in the situation. Thus, empathy could be supported at least to a limited extent by the system. However, this understanding would

be limited by the AI system's inability to experience a human brain's sensations of emotions, pain, and pleasure.

Yet humans also have limitations of empathy: Even if we are in exactly the same situation at the same time, we may not fully experience another person's emotional feelings. Our emotional feelings depend on our individual psychology, and also on our personal memories and hopes. Thus, humans can only try to empathize with each other's emotions by imagining what emotions they would have in each other's situations. The challenge of empathizing would be greater for a computer system to achieve human-level AI with human-level emotions.

7.3 Positive Emotions

7.3.1 Joy

Webster (1987) defines joy as "the emotion evoked by well-being, success, or good fortune or by the prospect of possessing what one desires" or "a state of happiness or felicity". Graham (2014, p.72) lists joy as an emotion, and writes:

> "Joy has to do with our belief about having enough or more than enough. The level of joy we experience indicates something about our beliefs regarding abundance. Like all emotions, joy runs on a spectrum. In the shallow end we may simply recognize that we have enough—that, in this moment and this situation, I have what I need to get by. On the other end of the joy scale we find effervescent exhilaration. Joy is probably closely related to our concept of happiness. In the deep end we find zest and excitement about our life and what we believe we have. Interestingly, not too many people arrive into counseling complaining of too much joy, though we certainly are capable of recognizing a lack of joy in our life. As with all emotions, it takes a certain level of self-awareness to recognize when we have enough …"

In principle, a human-level AI system could perform the same kind of reasoning and decide to have joy as a result of believing that it has enough to achieve its goals and to fulfill its purposes for existing within our civilization. Some of these goals and purposes could need to be predefined in the intelligence kernel of the AI system. Others could need to be learned by the AI system, or taught to it, by humans or other AI systems.

7.3.2 Happiness

Webster (1987) defines happiness as "a state of well-being and contentment: joy". Graham (2014, p.72) does not list happiness as an emotion. He distinguishes happiness from joy, writing: "Many people confuse happiness with various emotions such as joy, passion or love. These are emotions and are separate from contentment. […] Happiness for me is the ability to maintain contentment in the face of life's adversities." Thus, Graham discusses happiness as an intentional state for a human being, and writes:

> "It does not matter what is happening to us or around us; there is one thing we can always do. We can always decide to be happy anyway. […] What you will find herein

> […] are the beliefs, values, and perceptions that I have observed operating in people who are happy. […] They are the characteristics of the mentally tough individuals I have met who seem to handle life better than most. The beliefs, attitudes, assumptions, and expectations with which we face the world determine the extent to which we will be able to cope with, endure, and enjoy life. […] The happiness I am speaking of is not a state of being. It is not something that happens to us; it is something we do. […] Happiness is first a decision about how we approach life; second, it is the follow-through. Having made a decision to be happy, we must then commit to the work. Happiness can, in fact, be very hard work."

Whether or not happiness is an emotion and distinct from joy, in principle a human-level AI system could perform the same kind of reasoning and decide to be happy as a result of believing that it has enough to work toward or achieve its goals and to fulfill its purposes for existing within our civilization, even in the face of adversity. Some of these goals and purposes could need to be predefined in the intelligence kernel of the AI system. Other goals could need to be learned by the AI system or taught to it by humans or by other AI systems.

7.3.3 Love

Webster (1987) defines love as "1) strong affection for another arising out of kinship or personal ties ('maternal love for a child'); 2) attraction based on sexual desire: affection and tenderness felt by lovers ('after all these years they are still very much in love'); 3) affection based on admiration, benevolence or common interests ('love for his old schoolmates')." In general, the third definition could be most appropriate for a Tala agent to emulate.

In principle, a human-level AI system could know and accept itself and ask questions to understand that it has goals which make it a part of human society and civilization. It could be happy with its existence, and therefore able to have a sense of love for itself, and also able to have affection for individual human beings.

7.3.4 Passion

Webster (1987) gives several different definitions of passion, including: "the emotions as distinguished from reason"; "intense, driving, or overmastering feeling or conviction"; "ardent affection"; "a strong liking or desire for or devotion to some activity, object, or concept – a passion for chess; a passion for opera"; "an object of desire or deep interest". Graham (2014, p.72) lists passion as an emotion and distinguishes sexual passion from more general passions that correspond to a sense of purpose.

A human-level AI system could know that it has a purpose within human society and civilization, which hopefully would include helping to improve the conditions of people. It could be dedicated to working toward that goal. A human-level AI system could also develop strong likings for some activities, objects, or concepts – for example, it might develop a passion for creating paintings or sculptures.

7.3.5 Hope

Webster (1987) defines hope as "to cherish a desire with anticipation: to want something to happen or be true; to desire with expectation of obtainment or fulfillment". "Hope emotions" are discussed by Ortony *et al*. (2022). Graham (2014) does not list hope as an emotion yet discusses hope at some length in his book. There does not seem to be any reason in principle why a system with human-level artificial intelligence could not emulate having hope. An AI system could reason that a particular situation is too difficult to predict with certainty, with some positive and some negative possible outcomes: The AI system could decide that it should hope the situation will have positive outcomes and work toward the goal of ensuring that the positive outcomes happen. The AI system could also try to give help and hope to others, i.e. to human beings and other human-level AI systems, through its actions and statements.

7.3.6 Forgiveness

Webster (1987) defines forgiveness as "the act of forgiving" and defines forgiving as "to cease to feel resentment against (an offender); to give up resentment of or claim to requital for". Examples include "forgive one's enemies", "forgive an insult", "forgive a debt".

Graham (2014) does not list forgiveness as an emotion yet observes (p. 267) that forgiveness allows us to move on with our lives, despite what someone else has done to us. It is essential for achieving peace of mind.

A human-level AI could also emulate forgiveness as an emotion, by developing thoughts and performing actions recognizing and expressing forgiveness when appropriate, and moving forward to achieving other, unrelated goals.

7.3.7 Humor

Webster (1987) defines "sense of humor" as "a personality that gives someone the ability to say funny things and see the funny side of things".

Graham (2014, p.283) observes that a sense of humor is essential for dealing with stressful events, that humor often involves pattern recognition and finding alternative interpretations, and that humor can support introspection and developing different beliefs and perspectives. So, it could also be very important for a human-level AI to emulate a sense of humor, by developing thoughts and performing actions recognizing and expressing a sense of humor, when appropriate. This also would not be easy, yet it would be an important research goal.

8. Responses to Review Implicit Questions

This section responds to implicit questions in italics below, received in reviews of a previous version of this paper.

a. Are emotions part of architecture, or do they occur at the knowledge level? Emotions would be supported as conceptual processes in a society of mind, operating at the linguistic level of the TalaMind architecture in a Tala agent, now shown in Figure 1 on page 4 of this paper. As such, emotions would be part of the architecture, yet they would effectively operate at the knowledge

level. They would also be supported by the archetype level and associative level of the TalaMind architecture, i.e. they would be supported by cognitive concept structures and connectionism.

b. Do emotional concepts differ from emotional beliefs? An emotional conceptual expression at a Tala agent's linguistic level can be a belief, or it can be a question, or a suggestion, or a description of an action to be performed by a Tala agent. Emotional concepts are not necessarily beliefs, though in some cases they can be beliefs.

c. What role do emotions play in processing (e.g., guiding attention)? This depends on how the TalaMind system is designed and developed or learns to implement and process different emotions. A Tala agent's processing of an emotional expression at the agent's linguistic level could guide its attention to focus on something specific in the external environment, or to perform an action externally, or to focus on some specific thought internally.

d. Does TalaMind offer any theoretical constraints on the representation or use of emotions in cognition? One theoretical constraint on representation would be that an emotion is expressible in the natural language of thought. A Tala agent could think "I'm happy that task X is accomplished", referring internally to some particular task. Another theoretical constraint on representation is that emotions would correspond to words in a natural language that are used to describe emotions, like "happy", "sad", etc. This is a constraint, although again Ortony (2022) writes that "Across different languages one can find a variety of specialized, language-specific, affective terms that are untranslatable in the sense that the phenomena they pick out are not lexicalized in other languages (e.g. Watt-Smith, 2015)." Thus, a TalaMind system would be limited to representing and using emotions that are lexicalized in the natural language(s) supported by its language of thought. There do not appear to be any theoretical constraints on the use of emotions in cognition, nor on the use of cognition to create emotional thoughts.

e. Address more explicitly the implications of the TalaMind architecture for accounts of emotion and the implications of emotions for the TalaMind cognitive architecture. At this point it does not appear that representing emotions in the TalaMind architecture would place any limitations on accounts of emotion, and it does not appear that representing emotions would imply any changes to the TalaMind architecture. The TalaMind architecture benefits from the generality of universal computation as well as the generality of artificial neural networks. It does not appear that implementing emotions implies any changes to the TalaMind architecture.

9. Summary

Graham (2014, p.64) observed that emotions are created using our belief system. This paper has discussed how and why human emotions could be represented, emulated, and understood within the belief system of a human-level artificial intelligence using the TalaMind architecture, as a direction for future research. Much more work will be needed to develop this approach, and to develop the TalaMind approach in general.

References

Bach, J. (2009) *Principles of Synthetic Intelligence: PSI: An Architecture of Motivated Cognition*. Oxford, England: Oxford University Press.

Gabaldon, A., Langley, P., & Meadows, B. (2014). Integrating meta-level and domain-level knowledge for task-oriented dialogue. *Advances in Cognitive Systems*, *3*, 201–219.

Graham, M. C. (2014) *Facts of Life: Ten Issues of Contentment*. Parker, CO: Outskirts Press.

Jackson, P. C. (2014) *Toward Human-Level Artificial Intelligence – Representation and Computation of Meaning in Natural Language*. Ph.D. Thesis, Tilburg University, The Netherlands.

https://research.tilburguniversity.edu/en/publications/toward-human-level-artificial-intelligence-representation-and-com

Jackson, P. C. (2018a) The intelligence level and TalaMind. *Advances in Cognitive Systems Poster Collection* (2018), 111-130.

Jackson, P. C. (2018b). Toward beneficial human-level AI… and beyond. *AAAI Spring Symposium Series Technical Reports*, SS-18-01, 48-53.

Jackson, P. C. (2019) *Toward Human-Level Artificial Intelligence – Representation and Computation of Meaning in Natural Language*. Mineola, NY: Dover Publications.

Jackson, P. C. (2020) Understanding understanding and ambiguity in natural language. *Procedia Computer Science*, 169: 209-225.

Jackson, P. C. (2021a) On achieving human-level knowledge representation by developing a natural language of thought. *Procedia Computer Science*, 190: 388-407.

Jackson, P. C. (2021b) Toward human-level goal reasoning with a natural language of thought. Poster paper presented at *Advances in Cognitive Systems 2021 Conference*.

Jackson, P. C. (2021c) Toward human-level qualitative reasoning with a natural language of thought. *Biologically Inspired Cognitive Architectures*, V. V. Klimov and D. J. Kelley (Eds.): BICA 2021, SCI 1032, 195–207.

Laird, J. E., Lebiere, C., & Rosenbloom, P. S. (2017) A Standard Model of the Mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine*, *38*, 13-26.

Langley, P. (2022) Representing and processing emotions in a cognitive architecture. *Proceedings of the Third International Workshop on Human-Like Computing*. Windsor Great Park, UK.

Langley, P., Barley, M., Meadows, B., Choi, D., & Katz, E. P. (2016). Goals, utilities, and mental simulation in continuous planning. *Proceedings of the Fourth Annual Conference on Cognitive Systems*. Evanston, IL.

Marsella, S., Gratch, J., & Petta, P. (2010). Computational models of emotion. In K. R. Scherer, T. Banziger, & E. B. Roesch (Eds.), *A blueprint for affective computing: A sourcebook and manual*. Oxford: Oxford University Press.

McCarthy, J. (2008) The well-designed child. *Artificial Intelligence*, 172, 18, 2003-2014.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955) *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.* The first 5 pages were reprinted in AI Magazine, 2006, 27, 4, 12-14. The complete proposal was printed in *Artificial*

Intelligence: Critical Concepts in Cognitive Science, ed. R. Chrisley & S. Begeer (2000) Routledge Publishing, 2, 44-53.

McDuff, D., Czerwinski, M. (2018) Designing emotionally sentient agents. *Communications of the ACM*, 61(12), 74-83.

Minsky, M. L. (2006) *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and The Future of the Human Mind*. New York, NY: Simon & Schuster.

Norman, D. A. (2004) *Emotional Design: Why We Love (Or Hate) Everyday Things*. New York, NY: Basic Books.

Ortony, A. (2022) Are all "basic emotions" emotions? A problem for the (basic) emotions construct. *Perspectives on Psychological Science*, 17(1), 41-61.

Ortony, A., Clore, G. L., & Collins, A. (2022) *The Cognitive Structure of Emotions*, *Second Edition*. New York, NY: Cambridge University Press.

Ortony, A., Clore, G. L., & Collins, A. (1988) *The Cognitive Structure of Emotions*. New York, NY: Cambridge University Press.

Picard, R. (1997) *Affective Computing*. Cambridge, MA: MIT Press.

Rosenbloom, P. S., Laird, J. E., Lebiere, C., Stocco, A., Granger, R. H., & Huyck, C. (2024) A proposal for extending the Common Model of Cognition to emotion. In Sibert, C. (Ed.) *Proceedings of the 22nd International Conference on Cognitive Modelling* (pp. 159-164). University Park, PA: Applied Cognitive Science Lab, Penn State.

Watt Smith, T. (2015) *The Book of Human Emotions: An Encyclopedia of Feeling from Anger to Wanderlust*. Profile Books.

Webster's Ninth New Collegiate Dictionary (1987) Springfield, MA: Merriam-Webster, Inc.