Engineering Kindness: Building A Machine With Compassionate Intelligence

Cindy Mason, CMT, Ph.D.

Stanford University

cmason@steam.stanford.edu, cindymason@media.mit.edu

We present the first step towards building robots with compassionate intelligence - EM-2 a software agent with the capacity for compassionate, namely it has the ability to take into account the feelings of self and others in problem solving. Because emotions and feelings have a special relationship to time, to build EM-2 we are developing a representation and reasoning apparatus that maintains logical and emotional consistency in the face of knowledge about facts and feelings that can change over time. The approach is based on using experience building a multi-agent system for default reasoning and common sense knowledge garnered from human sciences. The architecture for EM-2 is inspired by a human mind training process from India called Vipassana centered on kindness and self awareness and 18th century commonsense philosopher David Hume's view that thought originates in the heart rather than the brain. Although separated by more than 2000 years they both present a philosophy of mind where emotion is an antecedent in logical thought, essential in compassion. The cognition of such an agent includes a new kind of inference called affective inference and meta-cognition involving representation and reasoning about its own beliefs, feelings, and the beliefs and feelings of others. We summarize the core meta-architectures and meta-processes of EM-2. We describe the process of Vipassana, a mental process for humans that cultivates kindness, from the viewpoint of a machine. We define affective inference and summarize the I-TMS, an Irrational TMS that resolves the turf war between thoughts and feelings based on agent personality rather than logic.

Keywords: Emotion, Time, EIQ, TMS, meditation, compassionate intelligence, human-level AI

Engineering Kindness: Building A Machine With Compassionate Intelligence

C. Mason
Stanford University
cmason@steam.stanford.edu

INTRODUCTION

Emotions have a special relationship to time. Happiness in one mental state can create actions we later regret. But regret is not possible without the passage of time. Some emotions are fleeting, like surprise, yet others like grief and love, can last a very long time. Giving satisfaction to a craving removes the feeling of craving but the craving then returns again and again. These emotional states have great influence and power in our minds and bodies yet we have barely scratched the surface of how to integrate our understanding into western medicine or computational systems that now permeate all of society.

Artificial Intelligence programmers deal with information in fairly logical terms - how many pixels in an image, how many keywords or concepts in a page, or how many links in a network. In order to process information with a computer it is the case that information needs to be formally, if not logically, organized. However, to create programming languages, data structures and algorithms that are as proficient at pattern recognition and learning as humans, we may need to consider an emotional approach.

Often, people deal with information, lots of information, on an emotional plane. The canonical example is that of overhearing our name at a cocktail party. Of all the sensory input we receive in any moment - all the sights, smells, tastes and sounds we receive second by second, the sound of our name being spoken has the instantaneous effect of filtering out the competing inputs and allowing us to hone in on anything and everything having to do with that most important input signal, our name. Simply stated, programming with emotions may lead AI researchers to develop human level AI, with solutions that are elegant and human, rather than brute force or complex.

The work presented here represents a new paradigm for AI computing – developing first principles of computational emotional intelligence by 1) organizing models of memory and mental state not only according to the meaning of objects and concepts but also according to the emotional meaning of those objects and concepts which change over time 2) incorporating emotion into an automatic inferencing mechanism that supports the revision of inferences over time 3) revising beliefs not according to logic but according to personality and psychological make-up and finally 4) incorporating multi-agent programming methods for the purpose of taking into account another agent's feelings during decision making and being responsive to changes of other agents over time.

It is the author's perspective that a first principles approach is a powerful foundation from which all of AI can be redeveloped to encompass an aspect of intelligence that was neglected in AI from the beginning of the field – namely, emotional intelligence. The viewpoint of the author is that with a first principles approach to including emotional intelligence in AI a range of new AI algorithms and data structures can then be developed and applied to all of the categories of AI - including pattern analysis, machine learning, robotics, vision, natural language, search, and more. In the least, it is certainly the next big step in the creation of social agents, realistic characters, dialogues and movement in games and virtual environments. Developing advanced models of mind that harness emotional and social intelligence is essential to the future of data analysis tasks involving the mind/brain and in tasks involving very large or continuous data collections.

BACKGROUND

The drive towards emotion oriented programming and affective computing is also fueled by the realization that the separation between man and machine grows closer each day. Robotic companions, robotic nurses or doctors, software tutors, and other systems that need to interact closely with humans must be able to analyze or model emotional content for user interaction, advanced information processing, and mind-man-machine devices. At this junction there are a variety of architectures for emotional computing, such as the system by Camurri and Coglio (Camurri & Coglio, 1998). The CogAff project provides an important schema from which to contrast and compare various affective agent architectures (Sloman, 2010). Some of the affective software architectures are inspired by practical applications such as how to display emotions in a robotic face, and others are driven to create architectures for a specific style of agent.

The work presented here represents a long term project begun in 1998 at a time when very few AI researchers, if any, were working on affect. The paper by Sloman and Croucher in 1981 was largely ignored (Sloman & Croucher, 1981). EM-2 was inspired by the author's work with patients in critical care at Stanford Hospital who were able to shave weeks off expected hospital stays by daily sessions in psychophysiophilosophy (Mason, 2003). The ideas behind psychophysiophilosophy are rooted in studies of philosophies and medicine from Asia and India where the architecture of the mind is explicitly described in its role in medicine and healing, including visualization and meta-cognition. The present work is greatly influenced by the meta-cognition practice of Vipassana (Gunaratana, 2002; Hanh, 1976).

Analyzing the issues that arise a result of combining affect and reason, or any type of cognition and affect, we are immediately faced with the problem of how to manage beliefs that can change over time and the need for a mechanism to monitor or introspect about those beliefs and feelings. Emotions are by their very nature highly volatile and do not follow a standard chain of reasoning. In the work of David Hume (1711 – 1776), one of the great common sense philosophers, he suggests that thinking arises from the heart (Hume, 1997.) The interested reader may wish to consider Hume's treatment on the issue of the origination of ideas in An Enquiry Concerning Human Understanding (Hume, 1997.) In computational terms, we can say that deduction or induction may be a result of feelings. That is, logical facts (whether learned or deduced) can be dependent upon emotions (and vice versa). Inherent in this style of reasoning is the need to revise what we know and feel because things, people and feelings change over time. This is common sense.

For example, I have a feeling in my heart that I love Don, and so when I learn that his mother died, I will re-schedule my commitments or travel plans around the schedule of his mother's funeral. Suppose I later discover his ex-wife will be at the funeral and because she once tried to run me over with her car, I notice my heart is racing, and so reasonably, fear becomes involved in the inferencing calculus. I re-schedule again, around the ex-wife's departure schedule. Logical facts that are later subject to revision, often due to incomplete information, are often referred to as a "belief". A collection of interdependent beliefs (and feelings) is usually referred to as a "possible world." The revision of beliefs and the context switch that occurs among various possible worlds can be accomplished by a computationally intensive architectural component known as a truth maintenance system (TMS). The subject of truth maintenance is often quite technical and lengthy so we refer the interested reader to some of the original papers in this area for single agent (deKleer, 1986; Doyle, 1979) and multi agent systems (Mason & Johnson, 1989; Bridgeland & Huhns, 1990). Multi-agent TMS is of particular value to the area of affective software agents because they allow direct representation and reasoning of another agent's beliefs, providing an agent with a glimpse of another agent's mental state. This feature is an important starting point when building an agent that can take the other agent's mental state and feelings into account when making individual decisions.

COMPASSIONATE AI

A program capable of compassionate decision-making, learning or compassionate social interaction requires several things. First, the intentional stance or philosophy of mind must concern itself with a positive intention toward others. The immediate consequence of this philosophy is the second requirement: the construction of representation and reasoning apparatus that support empathy as part of the bounded informatic situation. The third requirement follows from the second one: the cognition of such an agent must include an inferencing calculus that supports the representation of affect in general and a representation that includes reference not only to its own affect but to that of other agents' affect. Humans incapable of empathy or compassion are often categorized with mental illness and are generally considered anti-social if not dangerous by others (sociopaths and psychopaths). Historically, AI machines have no intentional capacity for empathy.

The EM-2 agent architecture is inspired and influenced by the philosophy of Vipassana or "insight meditation," an ancient mind training (meditation) system from India that concentrates on the development of kindness and compassion (Salzberg, 1995) (Rhys-Davids, 2003). One advantage in choosing "insight meditation" as a philosophy of mind is that it focuses on aspects of consciousness important for kind behavior. It is described as generating "loving-kindness, friendliness, benevolence, amity, friendship, good will, kindness, love, sympathy and active interest in others," see http://en.wikipedia.org/wiki/Mettā. The mental processes of Vipassana have been documented to create compassion and empathy in even the most hardened criminals (Ariel and Menahemi, 1997), and the architecture of mind is described in detail in a variety of places, most readily accessed through the audio Series Talks (Fronsdale, 2003) at Insight Meditation Center in Redwood City, California (http://www.insightmeditationcenter.org/), a sister center to the Cambridge Insight Meditation

Center organized by Jon Kabot-Zinn (http://www.cimc.info/), or see (Gunarana, 2002; Salzberg, 1995; Hanh, 1976; Rhys-Davids, 2003).

We address the second requirement, namely the construction of representation and reasoning apparatus that supports empathy, by reusing previously developed program components from a multi-agent system (ROO) and from an emotion oriented programming language called EOP used to build EM-1(Mason 1998). Essential components of mental state for an agent in a multi-agent system include representation of both internally and externally generated beliefs, the ability to distinguish by name the agents' beliefs and those generated by other agents and a mechanism to keep track of them in working memory as they become co-mingled during decision making, problem solving, or learning. Emotion Oriented Programming (EOP) is a rule-based language that supports explicit representation of emotional concepts such as mood and feeling as part of agent mental state.

We address the third requirement of compassionate AI with help from an 18th century philosopher named David Hume. Hume, like many other common sense philosophers, was obsessed with understanding the origination of thought. He proposed that thought is a consequence of something felt in the heart. In creating EM-2 we developed a new kind of inference called affective inference where emotion can be the antecedent to logical consequences in a forward chaining rule-based system.

Tying it all together is meta-cognition. EM-2 has rule-based meta-cognition involving representation and reasoning about its own beliefs, feelings, and the beliefs and feelings of others. Namely, the cognition of EM-2 includes a meta-representation of mental state objects that supports thinking about thinking, thinking about feeling, and thinking about thoughts and feelings – its own and/or those of other agents.

In this paper we give a brief overview of several topics relating to the construction of EM-2, an agent with compassionate intelligence. We give fragments of code from the language EOP (Emotion Oriented Programming). EOP was used to build the first pass at a software agent that could represent and reason with both emotional and mental state, EM-1 (Mason, 1998). We extended EM-1 to incorporate multiple agents and a TMS that uses a psychologically realistic, but irrational approach to consistency. The agent architecture of EM-2 and meta-level predicates and processes of EOP are based in part on previous work on multi-agent systems (Mason, 1995) and on the architecture of mind described in insight meditation or Vipassana.

HUMAN LEVEL AI

In contrast to the majority of current research in AI, the position of the author is that human-level AI programs must not only reason with common sense about the world, but also about feeling and sometimes with irrational reasoning, because every human being knows that to succeed in this world, logic is not enough. An agent must have compassionate intelligence. The heart of what it means to be both human and intelligent includes compassion and empathy, social and emotional common sense, as well as more traditional methods of AI suitable to tasks.

> "Whenever we change our emotional states, we find ourselves thinking in different ways. Our minds get directed toward different concerns, with modified goals and priorities and with different ways to describe what we see. Thus Anger can alter the way we perceive, so that innocent gestures get turned into threats. Love, too, can change our point of view turning blemishes into embellishments. Love also alters how we behave; The heart of what it means to be both human and intelligent includes compassion and empathy, social and emotional common sense, as well as more traditional methods of AI suitable to tasks."
>
> Marvin Minsky
> The Emotional Machine

EM-2'S PHILOSOPHY OF MIND

Recently ancient philosophies have become important in western culture. Mind training practices based on eastern philosophies (e.g. meditation) have come under the scrutiny of FMRI and other diagnostic tools and have shown that when humans engage in persistent mind training there is permanent changes in brain structure and function as well as a positive effect on mental and physical health (Begley, 2007; Lutz et. al., 2004). A dramatic example of this idea is the practice of TUMMO (Crommie, 2002; Benson, 1982). Crommie (Crommie, 2002) describes and illustrates Harvard researchers monitoring a TUMMO practitioner (a Buddhist monk) who uses mental meta-processes involving compassion to create dramatic body heat. The effects of the TUMMO practice were physically demonstrated by the semi-nude practitioner who sat for prologued period on ice without harm.

Vipassana, or Insight Meditation, has been practiced for 2500 years. Vipassana meditation practitioners engage in an active mental process of observation or visualization of mental objects, often with a representation of self, along with meta-processes that effect transformation in behavior, state of mind, affect, and or body function. At the heart of meditation is the engagement of a cognitive process known as mindfulness. Mindfulness is simply the act of paying attention to what we pay attention. The following example by Fronsdale illustrates this point (Fronsdale, 2003).

Consider that while you drive, you are concerned with the "doing" of driving – noticing signs, other cars, staying in your lane, and also with "reasoning" processes in navigation, planning and scheduling. While we look out for signs, cars, and children, we are not concerned with and not noticing our dirty windshield. Stopping for gas, we clean it and get back in the car. We do this as part of routine driving. As we start driving again, we become aware of the extra effort that was used to see objects and cars, and the road, due to the dirt.

Mindfulness is accomplished with a meta-level process sometimes called the "observer". The "observer" process is cultivated over a period of time to be attentive to thoughts, body sensations, and feelings, without judgment, resistance, or clinging. Regardless of the moral or ethical semantics of the objects, regardless of the difficulty of the emotion or the lack of apparent rationality of the object, the intentional stance in Vipassana towards these mental objects by the practitioner is gentle reflection.

This intentional stance, namely, the encouragement to notice when we cling or resist and encouragement to allow rather than judge, creates a safe mental space where anything that arises in our consciousness can be met with positive acceptance almost as a mother to a child. In this context thoughts, feelings, and body sensations that would otherwise never be noticed come into awareness, creating "insight" into oneself and others, hence the name "insight" meditation. The mindfulness practice is said to gradually dissolve the barriers to the full development of our human wisdom and compassion.

VIPASSANA AS A COMPUTATION: META-COGNITION

Natural systems of cognition have always been inspirational to AI researchers (e.g. vision, memory, locomotion.) Cultures where human mind training has evolved for hundreds and thousands of years present an untapped resource of ideas for researchers working towards human-level AI or compassionate intelligence. Many of these special mind training methods use an architecture of mind that is based on meta-cognition and meta-processes similar in structure and function to the diagram developed by Cox and others (Cox and Raja, 2011) as shown in Figure 1.

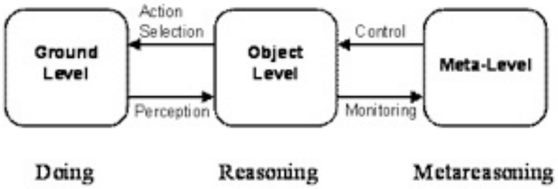

Figure 1. The architecture for meta-mind used for computation and in describing Vipassana.(Copyright 2011 Cox and Raja). Used with permission.

In Vipassana, the idea of non-judgment towards an object is well suited for a computational model and the "observer" could readily be described as a meta-level mental process for being attentive to agent mental state. We describe the process of Vipassana using the architectural components of Figure 1.

The Ground Level of perception involves the human physical embodiment of consciousness - sensory and body sensations such as the perception of inhaling or exhaling, the sense of contact with a supporting structure such as a chair or cushion, smell, sound, pain, pleasure, heat, coolness, itching, etc. At the Object Level is our "stream of consciousness". It includes the thoughts, feelings, or other mental representations of our ground level perceptions as well as daily chatter, problem solving, self-talk, mental imagery, and so forth. The Meta-level consists of the observer.

The "observer" is involved in the creation of Meta-Level mental objects concerning the stream of consciousness at the Object-Level. Examples of Meta-Level objects include labels for individual feelings or thoughts, labels for patterns and groups of thoughts and feelings, questions about thoughts, questions about feelings, images, self or other mental objects.) These meta-level objects regarding the stream of consciousness are used in various ways to direct attention to either Object Level or Ground Level for the purpose of Noticing and Observing or for the purpose of answering a question such as "What happens if I pay attention to the pain in my back?" The result of Noticing, Observing, or Asking Questions is the creation of more objects at both the Object and Ground level. Thoughts or realizations about those mental objects sometimes give rise to what is called "insight." Hence the name, "insight meditation." "Insight" about one's thoughts, feelings, or breath (body) can and do profoundly change the state of mind and the state of heart of the practitioner.

COMPASSIONATE INTELLIGENCE

Many features of human-level compassion will be wanted in an artificial agent, some will not. To motivate and ground our discussion here we ask the reader to consider applications requiring social or compassionate intelligence – examples include robotic nurses and nursing assistants, robotic nannies, automated intake and triage systems for psychiatric departments, user interfaces and dialogues for vehicle control panels, gaming characters and avatars, education and training software, customer relations support and so on.

Compassionate Intelligence is the capacity of an agent to act in a compassionate manner in a bounded informatic situation. Essential to the act of compassion is empathy – the ability to consider the possible (emotional) world of another. The computational components of EM-2 relating to empathy include a) the separate and explicit representations of feelings and beliefs about a proposition b) the ability to represent propositions and mental states of other agents. These features are discussed in the following sections.

Agent architecture

The architecture for an EM-2 agent is shown in Figure 2. As figure shows we use the name A_i to represent an arbitrary EM-2 agent. We subscript the agent name since it will be important later to distinguish among the agent A_i (self) and other agents. The architecture for A_i supports the Vipassana as Meta-Computation as follows. The "Observer" is effectively the execution of Inferencing and Learning (IL) and Truth Maintenance (I-TMS) processes on Meta-Object Memory (M) which contains Feelings and Beliefs about Objects in Object Memory (O) and on Objects. Objects are created through this process as well as by ground level actions that include both sensing (S) and communication (C). To be clear, Meta-Objects refer to Objects, and can be expressed as a predicate applied to an object. For example "Feels(Object)" or "Believes(Object)" where Object can be any physical or mental construct. Objects may be reasoned about and manipulated independently of the beliefs and feelings about those objects that reside at the meta-level (M). Reasoning and learning processes can be applied as both object-level and meta-level computation. In theory, we may create an arbitrary number of levels, especially in a multi-agent scenario, e.g. Feels (Feels (Object)), Feels (Feels (Feels (Object))) and so on, with endlessly rising levels. For simplicity, our discussion focuses on two levels.

In addition to sensing and communication, new objects may come from the common sense repository and as a result of learning. Because our agent resides in an environment with other agents capable of meta-cognition, the communications can be at both object and meta-level. In this way, an agent can "learn" what another agent "Feels" or "Believes" about an object. The context based truth maintenance system (I-TMS) works to maintain consistency among feelings and beliefs using common sense knowledge and knowledge based models of personality. In addition to communication, learning takes place as a result of conflict discovery and resolution (I-TMS) and also as a result of the common sense collective (CS). The focus in our chapter is on the mechanism for compassionate computation.

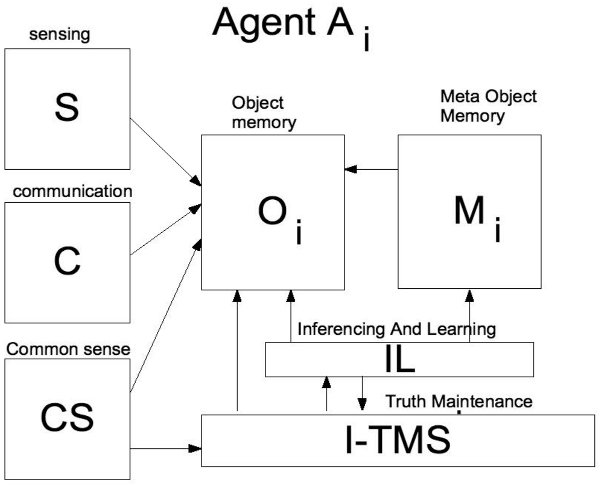

Figure 2. The component architecture for an EM-2 Agent.

AFFECTIVE INFERENCE

There is no question that in the natural course of thinking sometimes an emotion can give rise to a thought, or that thoughts may also give rise to an emotional state, which in turn gives rise to more thoughts or further emotional states, and so on. The meta-cognition process of insight meditation highlights the interdependency of feelings and thoughts in human cognition. Many 18th century "common sense" philosophers such as Hume, Locke, and others spent considerable time on the issue of affect and rational thought. As we approach the goal of creating compassionate intelligence, it is essential to have some form of affective inference. By this we mean

Definition: *Affective Inference is a method of inferencing whereby emotion can be the antecedent to a consequent thought, and vice versa.*

This style of inferencing presupposes a programming language where an agent's mental state contains explicitly named objects of emotional concept such as mood, emotional state, disposition, attitude, and so on in addition to traditional non-affective concepts and objects, and that these emotional objects require a separate but interdependent computational apparatus.

Affective inference differs from logical inference in some important ways. First, by its very nature affective inference is volatile – moods and emotions change over time and so will the inferences that depend on them. An essential component of an affective inference machine or agent is a truth maintenance mechanism. Typically, consistency refers to the idea of preventing logical theories involving P & ~P. However, because the nature of affective mental objects involves relative consistency rather than logical consistency we require a non-logical TMS. We require an Illogical-TMS, an I-TMS, where consistency is based on what "makes sense" relative to a number of factors, including common sense knowledge, personality, and logic. It is through this method of consistency maintenance that we are able to create agents with behaviors and decisions that reflect the experience of the world rather than a formal model of the world. We agree that this is perhaps unusual approach to consistency maintenance, as compared to traditional algorithms for both single (deKleer, 1986; Doyle, 1979) or multi-agent systems (Mason and Johnson, 1989; Bridgeland and Huhns, 1990), however, our intention is the creation of agents with an emotional stance.

EM-2’s affective inferencing mechanism is based on reasoning with defaults or assumptions. The style of reasoning is based on knowledge about what is normal or commonly expected. For example, “men are tall” or “birds fly.” A significant amount of common sense knowledge has been gathered about emotions through an open collective project on the common sense of happiness begun in 2008 called Open Heart Common Sense (http://www.openhearttreasures.com) and about objects in the world in an open collective at MIT called Open Mind Common Sense (http://openmind.media.mit.edu/). These knowledge collectives are processed into machine readable form. We are actively working on integrating this knowledge into EM-2. In both Open Heart and Open Mind, the common sense knowledge is filtered and reformatted from user input to semantic networks where objects or concepts are represented internally as triplets or assertions for ease of processing by other algorithms such as truth maintenance.

Truth Maintenance Systems look for conflicts between the assertions that actually reside in memory (mental state or reality) and the expected or anticipated assertions of common sense knowledge. As shown in Figure 2, the I-TMS relies on two subsystems, each representing the antecedents or support for each object of mental state with their own style of consistency maintenance. The first is IM – an Introspective Machine that represents the logical thoughts of agent(s) at the object and meta-object level and maintains consistency using traditional notions of logic and truth maintenance as described in (Mason, 1995). The second subsystem is IH: an Introspective Heart that represents the feelings of the agent at the object and meta-object level. Unlike the IM, it maintains a relative consistency based on an agent’s psychological and social rules as determined by knowledge about personality and cultural information.

Conflicts between logic and feelings can lead to indefinite loops, resulting in thrashing during the I-TMS algorithms. While this is a rich issue with philosophical and computational implications, we chose to solve this problem with knowledge and by allowing the possibility that an agent programmer may need more than one possible solution. That is, which of the IM or IH systems holds the upper hand in the I-TMS and therefore in the inference process depends on the application, personality, social, and cultural aspects of the agent. For example, agents with a logical personality and a romantic personality will have different contradiction resolutions and hence differences in objects and the feelings about those objects represented in mental state. Before we engage in the detail of IM and IH, we illustrate this concept by example. In the following section, called “Love is Blind”, we demonstrate the point described by Minsky, that “Love can change our point of view, turning blemishes into embellishments,” (Minsky, 2006).

LOVE IS BLIND

We demonstrate the idea of Affective Inference using EM-2 language constructs. The example illustrates the use of explicit representation of how an agent “feels” about a proposition. The example is interesting because it gives a different outcome depending on the personality of the agent.

```
R1: IF (Feels (In-Love-With(x)) ) then
                        (Assume(Handsome(x)))

R2: IF (Believe (Obese(x))) THEN NOT(Handsome(x))

R3: IF (Believe (Proposes(x)) ) and
       (Believe (Handsome(x)) ) THEN
    Accept-Proposal(x)

P1: Feeling(In-Love-With(Peppy))      {{P1}}   B

P2: Proposes(Peppy)                    {{ }}    B

A1: Handsome(Peppy)                   {{A1}}   B

D1: Accept-Proposal(Peppy)            {{A1}}   B
```

On start-up Agent EM-2's mental state contains the rules R1, R2, and R3 along with the premises, P1 and P2. Mental Object A1 is created as an assumption using the premise P1 and the rule R1. The object of the agent's love feeling, Peppy, is declared to be handsome precisely because the agent loves Peppy. Next, mental object D1 is derived by the application of rule R3 to A1 and P2. That is, if someone (Peppy) is handsome and that someone (Peppy) proposes then we create the mental object "Accept-Proposal(Peppy)." So far, all the objects in memory are currently believed, as indicated by status "B".

Now suppose we learn

P3: Obese (Peppy)	{{}}	B

Then later derive

D2: NOT (Handsome)	{{P3}}	B

as a result of the rule R2 and P3. There is now a conflict between D2 and A1. Unlike a traditional TMS, the manner in which this conflict will resolve, and the resulting set of mental objects that are believed or disbelieved do not depend solely on logic but on common sense about the personality and social/cultural make-up of the agent as well. The heart (IH) and mind (IM) both engage in the process of relabeling.

Agents with a logically inclined personality have meta-rules in IH that allow it to trump the IM. In an agent with a logical personality, we would then find there is a contradiction, and both A1 and D1 will become disbelieved. We also make a record of the contradiction involving A1 and D1, since this, too could change in the future.

A1: Handsome (Peppy)	{{A1}}	D
D1: Accept-Proposal (Peppy)	{{A1}}	D

Agents with personalities interested in attachments, loyalty and a tendency towards socializing will give preference to feelings in a conflict. A romantic agent's IH subsystem would not contradict this.

A1: Handsome (Peppy)	{{A1}}	B
D1: Accept-Proposal (Peppy)	{{A1}}	B
D2: Not(Handsome (Peppy))	{{P3}}	D

DISCUSSION

In the Love is Blind example, there are 3 rules, R1, R2 and R3. Premises, denoted by P, are observed, felt, sensed or created at agent start-up. In the example, Love is Blind, there are two kinds of premises. Those based on feeling and those based on logic. Notice that P1 is a feeling object as distinct from logical objects P2 or P3 and has the label "{{P1}}". A feelings premise allows the agent to begin a chain of reasoning based on an initial feeling, as consistent with common sense philosopher David Hume's standpoint on human reasoning. The labels for P2 and P3 are both "{{}}" and contain only the empty set, which is a typical premise label for an agent using logic oriented TMS.

The reason for label distinction among premises here is because feelings and logical facts are treated differently during the truth maintenance process and by the meta-operators. These ideas are discussed further in the next section. Assumptions are denoted with the letter "A," and derivations, denoted by the letter "D." Each premise, assumption and derivation has a label indicated by brackets "{}" which contain the mental contexts in which the object has been derived. It is used by the Truth Maintenance System to determine the contexts in which it may be believed/disbelieved and in explaining how the object was created. Next to the label is a meta-tag indicated the agent's belief status for the mental object. The I-TMS provides the ability to introspect about a mental object from the left hand side of a rule through the query operators "Believe," "Disbelieve," "Known" and "Unknown." "Believe" indicates there is a context in which all of its support is Believed. "Disbelieve" indicates there is no support for the mental object. "Known" indicates the object exists in mental state in either "Believe" or "Disbelieve" status, while "Unknown" indicates an object is not "Known". There is also an operator called "FEELS" which allows the agent to introspect on its feelings about an object. In our example we concern ourselves simply with the Believe, Disbelieve and Feels operator.

In our example, when the meta-operator Believe is applied in the evaluation of the left hands side of a rule, when those objects are marked B the clause will evaluate as true, and become executable. If at any time the belief status changes from B to D, it is no longer executable.

MENTAL STATE OF A COMPASSIONATE AGENT

The Subsystem IM

The response of the introspective belief machine IM_i is based on matching ϕ against the network of concepts and processes held in mental state. Queries of the form $\Box\ \phi$ presented to IM_i by other consciousness functions can be answered by presenting ϕ to the machine M_i. This formulation of computational introspection is intuitively appealing as it appears to mimic human introspection. While multiple levels of introspection are possible, because meta-objects can be considered as objects, for our purposes a single level of introspection suffices.

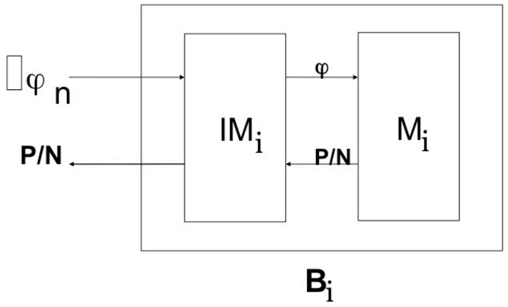

Figure 3. The logic subsystem IM of the I-TMS maintains what is believed.

As shown in Figure 3 the logical beliefs B of agent A_i, can be described as a two-level introspective machine, with an introspective component IM, and the belief machinery, M, that maintains consistency and determines whether a concept is a member of mental state. It may also detect the contexts or relative links to other concepts as needed. In order to access the contents of mental state other consciousness functions may submit queries to the belief subsystem posed in computational language L whose exact form is inconsequential except that there must be an explicit reference to an agent's mental state. In a healthy agent mind, a mental concept ϕ will be provided to other consciousness functions if and only if there is support for ϕ, which means agent A_i believes ϕ. We represent these expressions by the form $\Box\ \phi$. In general, ϕ may represent concepts created by consciousness functions of A_i or another agent, A_i, as when $\boldsymbol{\phi}$ has been communicated. We use the notation ϕ_n to represent a concept originating from agent A_n when the origin of ϕ bears relevance to the discussion.

The Subsystem IH

The following is a machine-theoretic description of the mental state subsystem IH when the agent evaluates a rule containing an query $f(\phi)$ regarding the agent's feelings about antecedent ϕ :

$$f(\phi) \qquad H(\phi) :P \rightarrow IH(f\phi): P$$

$$H(\phi) :N \rightarrow IH(f\phi): N$$

$$\neg f(\phi) \qquad H(\phi) :P \rightarrow IH(\neg f\phi): N$$

$$H(\phi):N \rightarrow IH(\neg f\phi): P$$

When faced with the query $f(\phi)$, IH poses the query ϕ to H, and simply returns Positive if H says Positive, and Negative if H says Negative. From the cognitive perspective of the agent, "Positive" means that the agent has considered its set of feelings and has concluded that it has favorable feelings about ϕ and therefore that it feels ϕ. In other words, the agent double checked with its heart component of mental state and is positive it feels ϕ. When presented with $\neg f(\phi)$ IH will respond Negative if H says Positive, indicating that that agent does feel ϕ. "Positive" reply from IH means that the agent does not feel ϕ. The agent does not feel that ϕ so $\neg f(\phi)$ is part of mental state.

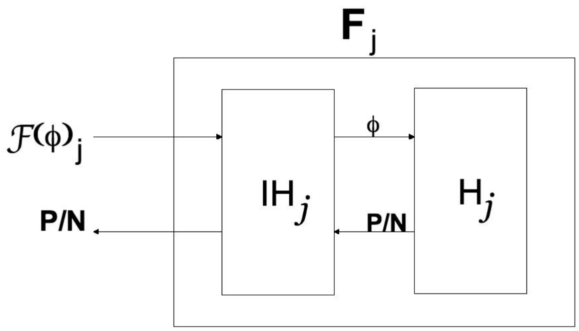

Figure 4. The Feelings Subsystem, F_i of the I-TMS maintains relative consistency among interdependent feelings.

We have now defined an agent's cognitive outlook in terms of its state of positive or negative feelings on the proposition ϕ. Together the set of positive feelings and the set of negative feelings constitute what is felt by the agent. We may define the set of known feelings as the set of ϕ that satisfy a query in L of the form $f(\phi) \vee \neg f(\phi)$. We define the "known feelings" modal operator $\Im$ as follows:

$$\Im\phi : f(\phi) \vee \neg f(\phi)$$

that is, the set of all ϕ that are in the agent's feelings list regardless of state of its feeling toward ϕ. It follows that the set of unknown feelings is the set of ϕ that satisfy a query in L of the form $\neg(f(\phi) \vee \neg f(\phi))$. We define the "unknown feelings" modal operator with $\neg\Im$ as follows:

$$\neg\Im\phi{:}\neg(f(\phi) \vee \neg f(\phi))$$

that is, the set of all propositions ϕ for which the IH "shrugs its shoulders" – it answers negative to both $f(\phi)$ and $\neg f(\phi)$ (alternatively, you may chose to implement the concept of "unknown" feelings with both positive, or by creating a "don't know" state, etc.) The idea of the unknown feelings operator is to describe an agent that has feelings but is neither positive nor negative towards ϕ (humans might refer to this as indifference, being neutral, or undecided.) It is important to distinguish between an agent that has feelings but simply does not know what they are and an agent with no feelings. In the latter case, the operator $\Im$ is undefined.

The presence or absence of $\Im$ in agent mental state is a means by which agents may be divided into two camps: those that have the capacity to reason with feelings and those that do not. Presently most agents fall into the latter category.

Communicated Feelings In an agent with compassionate intelligence, propositions in the feeling system may occur not only as a result of affective inference and knowledge as discussed in the previous section, but also as a result of communication. That is A_I believes $f\phi_J$ as a result of a previous communication of $f\phi$ to A_I by A_J. The set of propositions an agent has feelings about includes not only locally generated propositions but also propositions shared by other agents. It follows that agents may reason about another agent's feelings about a proposition as well as its beliefs. This is the heart of compassion and of indifference – to consider and take into account or not the feelings of another agent when engaged in reasoning, planning and scheduling, decision making, and so on.

It is possible that A_I also feels $f\phi_J$, that is, $ff\phi_J$ - in this case, the agent might be called empathetic. We could describe this $f\phi_J\ f\phi_I$ where semantics(ϕ_J) ~ semantics(ϕ_I), if agent A_I believes $f\phi_J$ where $J \neq I$ (it believes something about the feelings of another agent) then agent A_I believes that agent A_J feels ϕ. It is possible that A_I also feels $f\phi_J$, that is, $f\phi_I$ - in this case, the agents feel the same.)

Consistency Maintenance

When agents reason with feelings, as when reasoning with defaults or assumptions, the inferences that an agent holds depending on those feelings may be retracted over time when feelings change. A special problem worth noting may arise in distributed multi-agent systems when agents use communicated feelings in their reasoning processes. Compared to logical belief, feelings can be relatively more volatile. The distributed state gives rise to a situation where an agent's feelings change after they have been communicated. In this case collaborative agents may be "out of synch" (in humans we refer to this as "out of touch.") Using our model the situation may be described as:

$$\text{For } A_I\ IM_I\ \ (f(\phi_J))\text{: P and For } A_J\ IH_J\ \ (f(\phi_J))\text{: N}$$

where A_I believes that A_J feels ϕ, but A_J does not feel ϕ. The situation is remedied once A_I receives notification of the change by $A_{J,}$ but not without some cost.

Agents reasoning with affective inference may require increased demands for computation and communication depending on the degree of persistence of agent state (agent personality) or the dynamics of the domain (e.g. frequent sampling of hardware devices measuring affective state of drivers or pilots.) In general, TMSes can be computational intensive, and like traditional or non-affective inferencing, alternative solutions and improvements will be needed as a result. Due to space constraints we do not address many issues nor can we discuss topics completely or in depth. We continue work on EOP and EM-2 as well as the I-TMS.

COMPASSION AND OUR MACHINES

The mechanisms for reasoning with regards to another's feelings only makes sense if there is wisdom to go along with it. This is a very important point. For a machine to engage in our world with a compassionate stance, we are faced with the task of articulating the common sense of compassion. Not all engineers and scientists are born with the gift for empathy, sympathy or compassion. We require collaboration with educators, psychologists, mothers, priests, our pets and even the kindness of strangers, to achieve the level of interaction that would enable the compassionate stance in a computational machine. The idea of programming our interfaces and embodied agents with a compassionate stance has great potential for positive influence in our cultures.

> "If you go into a music shop and pluck a string of a violin, each of the other instruments in the store will resonate with that sound. Similarly human beings can resonate with each other to such an extent that they can exchange understanding at a subtle level."
>
> Rollo May
> Healers On Healing

Such is the influence of the objects and machines in our environment as well. It will be important to know what kind of training or what kind of morals we can rely on for a machine in our environment as more and more robotic companions appear in society. To this extent, we began collecting self-reported common sense knowledge from the general public regarding happiness, kindness and self awareness (Mason & Smith, 2008). We are also moving forward with a collection of memes and bemes as they relate to the creation of a positive cyberpersonality and cyber consciousness. Please feel free to visit this website and contribute your knowledge: www.openhearttreasures.org.

DISCUSSION AND SUMMARY

Agents that can represent and reason about the feelings of self and of other agents in decision making and inferencing are a necessary step towards human-level AI. Computing with compassionate intelligence requires the development of a reasoning apparatus that can adapt mental state over time according to changes in feelings of self and others. We approach the problem by developing affective inference as a means of common sense default reasoning. Unlike traditional logical inference, affective inference involves the justification of facts based on emotion and vice versa. Emotions can bootstrap the creation of objects but because emotions are particularly sensitive to the passage of time and the "consistency" of emotions often has more to do with social and emotional psychology and integrity, we have developed a mechanism for revising beliefs and feelings over time that maintains consistency not necessarily in a traditionally logical fashion but in accordance with social and emotional semantics. This opens a very large can of worms theoretically in terms of understanding the properties of such a system and we will continue to work on developing deeper understanding of these real issues. However, the application of this idea to engineering and engendering of kindness in cyberspace cannot be started soon enough. In support of the US White House project on anti-bullying in cyberspace, which involves a number of departments (Education, Health and Human Sciences) and universities, we proposed one way this can be supported is by creating memes and bemes. Memes and bemes reflect states of mind of users and evolve over time.

The work presented here does not address a number of controversial issues involving emotions – namely the origination of emotion, psychological theories of emotion and so on. Our belief in the engineering of tools despite a lack of consensus on practical and that building systems helps to shed light on issues. Human potential is greatly expanded when emotional support is provided, and we are inspired to engineer systems with emotional intelligence by our experience with remarkable patients using emotional therapies while recovering from bone marrow stem cell transplant procedures at Stanford Hospital [Mason 2003].

The agent architecture and processes of EM-2 support the philosophy of mind that emotion can give rise to thought and should be explicitly represented in the inferencing process. Like reasoning with common sense knowledge, emotions are subject to change over time and any inferences or concepts created that are based on those feelings or common sense knowledge must be revised. To accomplish this we introduce the idea of an Ilogical-TMS (I-TMS) where psychologically based justifications may be used to support belief or other feelings about a proposition as well as traditional logic. We chose a knowledge based semantics or psychologically based approach to conflict resolution in the I-TMS in part because many of the domains we work in, such as very big scale data analysis, present such knowledge as a way of resolving conflict. In general however, an agent programmer may need more than one way to approach the conflict resolution problem. A traditional logic based approach to resolving conflicts between logic and feelings in the I-TMS can lead to indefinite loops, resulting in thrashing during the I-TMS algorithms. The algorithmic approach to conflict resolution is a rich issue with philosophical and computational implications and deserves further work but is not an easy problem. Meta-cognition is central to both the semantics and the syntax/logic based approach to conflict resolution and these issues easily relate to the ideas of Goedel, Turing and others. We hope to continue this work and welcome collaborations.

Meta-cognition is also central in representing and reasoning about another agent's beliefs and feelings. The ability to explicitly represent the beliefs and feelings of the self as distinct from another agent in mental state while consider several possible worlds is central to the ability to reason with compassion – that is, to taking into account another agent's feelings and beliefs during reasoning, planning, decision making and problem solving. It is very important to notice that it is not enough simply to have these mechanisms. The intention of what to do with the feelings and beliefs of another agent (human or machine) is the cornerstone of creating compassionate agents. While EM-2's architecture reflects the influence of the meta architecture of mind in the Vipassana meditation practice, there is a connection between what we observe at the meta level and what we do with it. These naturally take place over the course of time. As humans, our observations and awareness at the meta-level or observer level are in hindsight but lead to an opening of the heart. No matter what the agent architecture or knowledge structure, this will not happen in a machine unless we make it so. Meta-computation is thus an essential component in reviewing past behavior.

We believe that compassionate intelligence is a necessary but not sufficient means to create artificial general intelligence. It is essential for the next generation of games, film, telecommunication and many applications involving social interactions. If we do not provide affective mechanisms for computer systems with close human interaction, we will miss a great opportunity for improving the human condition as well as build systems that are blind to many forms of intelligence. As mentioned earlier, endowing robots with the traits of friendliness, benevolence, amity, friendship, good will, kindness, love, sympathy and active interest in others is desirable. However, it must be said that the decision to give an agent affective computing depends entirely on the context and the application, as illustrated in John McCarthy's science fiction story, The Robot and The Baby, (http://www.formal.stanford.edu/jmc/robotandbaby/robotandbaby.html). It is entirely possible that if we create synthetic mind that we will also need synthetic mental health practitioners, e.g. robot psychologists.

It is worth noting that in our paper, we have often used the term human level AI and compassionate intelligence as dual terms. Although one might object to this duality it is the opinion of the author after having digested much about brain anatomy and become certified in a medical system where emotions and physical health are inseparable, as well as having continued to monitor the evolution of neuroscience perspectives on the importance of emotion in all aspects of life, that one cannot reach human level intelligence in an AI system without compassionate intelligence. That said, although compassionate intelligence is necessary for human level AI, they are perhaps distinct concepts. Along the same lines, while our three-leveled architecture of mind may appear to have the power to differentiate 'feeling' and 'emotion', a philosophical discussion of their distinction, in terms of knowledge and perception or consciousness, will be left for another time. We believe the distinction does not affect the usefulness of the mechanism. It is possible the perspective that EM-2 agents are designed to act similar to human level reasoning by taking into account beliefs and desires, that they can be evaluated as an evolutionary approach to goal-driven agents, is useful.

We are expecting to find that psychologically valid computational models of mind and emotion are useful in very large scale data analysis. We are currently applying working on a case study to emulate our emotional reactions to features in very large data sets with the hope that we will avoid many computations. The applications for this kind of model are quite open ended. Indeed, Infosys is well on its way to building a research lab around the idea of emotion oriented programming and affective computing. We continue to develop EM-2 with more features that take into account multiple agents and communication and its implications on the I-TMS. In general, the problem of I-TMS and multiple agents is computationally intense and provides a number of philosophical and practical problems, much like trying to keep track of what someone else feels and thinks in order to take that into account. Not an easy feat, even for humans. Perhaps, if we can create machines that can keep track of others' feelings and beliefs, it will be a helpful social tool, much like our calendar or contacts list.

REFERENCES

Ariel, E. & Menahemi, A. (1997). *Doing Time, Doing Vipassana*, Karuna Films.

Begley, S., (2007). *Train Your Mind, Change Your Brain: How a New Science Reveals Our Ability to Transform Ourselves*. New York: Ballantine.

Benson, H., Lehmann, J., Malhotra, M., Goldman, R., Hopkins, J., & Epstein, M. (1982). Body temperature changes during the practice of g Tummo yoga. *Nature Magazine, 295,* 234 – 236.

Bridgeland, D. M. & Huhns, M. N. (1990). *Distributed Truth Maintenance*. Proceedings of AAAI–90: Eighth National Conference on Artificial Intelligence. AAAI Press.

Carlson, L.; Ursuliak, Z.; Goodey, E.; Angen, M.; & Speca, M. (2001). The effects of a mindfulness meditation-based stress reduction program on mood and symptoms of stress in cancer outpatients: 6-month follow-up. *Support Care Cancer, 9*(2), 112-23.

Camurri, A. & Coglio, C. (1998). An Architecture for Emotional Agents. *IEEE Multi-Media, 5*(4) 24-33.

Cox, M. & Raja, A. (2011). Meta-Reasoning – An Introduction. In (Cox & Raja, Ed.)*, Thinking about Thinking* (pp. 3- 14). Cambridge, MA: MIT Press.

Cromie, W. (2002). Research: Meditation changes temperatures: Mind controls body in extreme experiments. Cambridge, Massachusetts, *Harvard University Gazette:4.*

Davidson, R., Kabat-Zinn, J., Schumacher, J., Rosenkranz, M., Muller, D., Santorelli, S., Urbanowski, F., Harrington, A., Bonus, K., & Sheridan, J. (2003). Alterations in brain and immune function produced by mindfulness meditation. *Psychosomatic Medicine, 65*(4), 564-570.

de Kleer, J. (1986). Problem solving with the ATMS. *Artificial Intelligence, 28,* 197-224.

Doyle, J. (1979). A Truth Maintenance System, *Artificial Intelligence, 12,* 231-272.

Fronsdale, G. (2003). *Introduction to Meditation*, www.audiodharma.org/talks-gil.html.

Gunaratana, V.H. (2002). *Mindfulness in Plain English.* Boston, MA, Wisdom Publications.

Hanh, T. N. (1976). *The miracle of mindfulness! : A manual of meditation*. Boston, MA: Beacon Press.

Hume, D. (1997). *An enquiry concerning human understanding : A letter from a gentleman to his friend in Edinburgh / David Hume*. (Ed.) Eric Steinberg, Indianapolis, IN: Hackett Pub. Co.

Kabat-Zinn, J., Lipworth, L. & Burney, R. (1985). The clinical use of mindfulness meditation for the self-regulation of chronic pain. *Journal of Behavioral Medicine 8*(2): 163-190.

Lutz, A., Greischar, L., Rawlings, N., Ricard, M. & Davidson, R. (2004). Long-Term Meditators Self-Induce high-Amplitude Gamma Synchrony During Mental Practice. *Neuroscience 101*(46): 16369–16373.

Mason, C. (2003). Reduction in Recovery Time and Side Effects of Stem Cell Transplant

Patients Using Physiophilosophy. *Late Breaking News Abstract at the Proceedings of the International Conference on Psychoneuroimmunology*, FL:PNIRS.

Mason, C. & Smith, C., (2008). Open Heart Treasures – A Public Collection of the Common Sense of Happiness. Retrieved November, 2008 from www.openhearttreasures.org.

Mason, C. (1998). Emotion Oriented Programming. Formal Notes, SRI AI Group. See also www.emotionalmachines.org.

Mason, C. (2004). Global Medical Technology in Bushko (Ed.) *Future of Health Technology* Boston, MA: IOS Press.

Mason, C. (1995). Introspection As Control in Result-Sharing Assumption-Based Reasoning Agents. *Proceedings of the 13th International Workshop on Distributed Artificial Intelligence,* Lake Quinalt, WA: AAAI Press.

Mason, C. & Johnson, R. (1989). DATMS: A Framework for Assumption Based Reasoning. In M. Huhns (Ed.) *Distributed Artificial Intelligence Vol.:2*, (pp. 293- 317). London: Pitman.

May, R. (1989). The Empathic Relationship: A Foundation of Healing. In Carlson & Shield (Eds.) *Healers on Healing* (pp 108-110). New York, NY: Penguin Putnam.

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

Rhys-Davids, C.A.F. (2003). *Buddhist Manual of Psychological Ethics, of the Fourth Century B.C., Being a Translation, now made for the First Time, from the Original Pāli, of the First Book of the Abhidhamma-Piṭaka, entitled Dhamma-Sangani (Compendium of States or Phenomena).* Whitefish, Montana: Kessinger Publishing.

Salzberg, S. (1995). *Lovingkindness: The Revolutionary Art of Happiness.* Boston, MA: Shambhala Publications. ISBN 1-57062-176-4.

Sloman, A., & Croucher, M. (1981). Why Robots Will Have Emotions. *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, Vancouver, Canada.

Sloman, A. (2010). http://www.cs.bham.ac.uk/research/projects/cogaff/cogaff.html.