Human-Level AI Requires Compassionate Intelligence

Cindy L. Mason

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

Abstract

Human-level AI will require much more than just common sense about the world. It will require compassionate intelligence to guide interaction and build applications of the future. The cognition of such an agent includes Meta-cognition: thinking about thinking, thinking about feeling, and thinking about others' thoughts and feelings. We summarize the core meta-architectures and meta-processes of EM-2, a meta-cognitive agent that uses affective inference and an irrational TMS, the I-TMS, to resolve the turf war between thoughts and feelings based on agent personality rather than logic. The work is inspired by a human mind training process from India called Vipassana and 17th century commonsense philosopher David Hume. Although separated by more than 2000 years they both present a philosophy of mind where emotion is the antecedent to logical thought.

Human-Level AI

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. A program capable of empathetic decision-making or compassionate social interaction requires some meta-cognition as part of the bounded informatic situation. Namely, the cognition of such an agent includes thinking about thinking, thinking about feeling, and thinking about thoughts and feelings – its own and/or those of other agents. The position of the author is that human-level AI programs must not only reason with common sense about the world, but also about irrationally and with feeling, because every human being knows that to succeed in this world, logic is not enough. An agent must have *compassionate intelligence*.

In this abstract we give a brief overview of several topics relating to the construction of an agent with compassionate intelligence. An extended paper with considerably fuller detail will be found on our webpage. In support of the work on meta-cognition and 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 a developmental philosophy of mind known as "insight meditation" or Vipassana.

Mind Training From India and AI

Recently ancient mind-training practices involving meta-cognition have become very popular in western cultures. FMRI and other diagnostic tools have shown that when humans engage in persistent mind training involving meta-processes there is a positive effect on mental and physical health as well as permanent changes in brain structure and function [Begley 2007] [Lutz et. al. 2004]. A dramatic example of this idea is the practice of TUMMO [Crommie 2002], [Benson, 1982]. Figure 1A [Crommie 2002] shows a Harvard researcher monitoring a practitioner (a Buddhist monk) who uses mental meta-processes to create dramatic body heat (see Figure 1B).

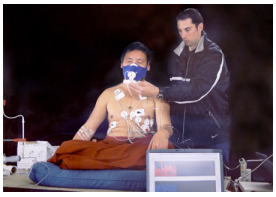

Figure 1A

Figure 1B

Connecting the Dots. How do human mind training practices relate to AI systems? 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 2008] as shown in Figure 2.

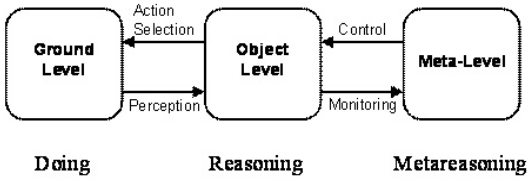

Figure 2

The systems engage practitioners 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. One advantage in choosing "insight meditation" as a philosophy of mind is that it focuses on aspects of consciousness important for kind behavior. The mental processes of Vipassana have been documented to create compassion and empathy in even the most hardened criminals [Ariel and Menahemi 1997]. If human history is any prediction of the future, it looks like we will need and want some robots to exhibit kind behavior. In the least, agents will need to effect kind behavior in applications involving communication, health monitoring, or when working with hardware to measure emotional state of patients (users).

How Vipassana works As A Meta-Cognition We describe the process of Vipassana using the architectural components of Figure 2. The Ground Level of perception involves the human physical embodiment of consciousness - sensory and body sensations such 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 a complex process sometimes called 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

Compassionate Intelligence is the capacity of an agent to act in compassionate manner in a bounded informatic situation. Many features of human-level compassion or emotional intelligence will be wanted in an artificial agent, some will not. Thinking about designing an AI system without view of an application or without view of a particular philosophy of mind can lead to the discovery of new approaches to some of the problems of AI. We are working to build EM-2, a software agent with compassionate intelligence. The philosophy of mind behind EM-2 is one part ancient India and one part 17th century philosopher. 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, automated intake and triage systems for psychiatric departments, user interfaces and dialogues for vehicle control panels, education and training software, customer relations support and so on.

The features of EM-2 include a) an irrational new form of inference called affective inference b) separate and explicit representations of feelings and beliefs about a proposition c) ability to represent multiple mental states or agents.

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 17th 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 human-level AI or face the challenge of applications requiring social and emotional 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. However, because the nature of truth maintenance of affective mental objects involves relative consistency rather than logical consistency we require a non-logical TMS. We require an Irrational-TMS, an I-TMS.

EM-2's affective inferencing mechanism uses defaults or assumptions and Truth Maintenance Systems to carry out common sense reasoning. It relies on two subsystems representing the antecedents of mental state, each with their own style of consistency maintenance: **IM** – an Introspective Machine that represents the logical thoughts of agent(s) and maintains consistency using traditional notions of logic as described in [Mason 1995] and **IH**: an Introspective Heart that represents the feelings of the agent and maintains a *relative consistency* based on an agent's psychological and social rules as determined by a software component that houses personality and cultural information. Which of the **IM** or **IH** systems holds the upper hand in the inference process depends on the application, which in turn affects the personality, social, or cultural aspect of the relation between feelings and beliefs.

Love is Blind. We demonstrate the idea of Affective Inference in an example called Love is Blind. 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 Assert(Handsome(x))
R2: IF BELIEVES(Obese(x)) then NOT(Handsome(x))
R3: IF BELIEVES(Proposes(x) and Handsome(x)) Then
Accept-Proposal(x)
P1: FEELS(In-Love-With(Peppy))
A1: Handsome(Peppy) {{A1}}

P2: Proposes(Peppy) Premise {{}}
D1: Accept-Proposal(Peppy) {{A1}}

Now suppose we learn

P3: Obsese (Peppy) Premise {{}}

Then

D2: NOT(Handsome)

Agents with a logical personality have meta-rules in IH that allow the DATMS to trump the I-TMS. In an agent with a logical personality, we would then find there is a contradiction, and both A1 and D1 will become disbelieved. 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.

Mental State of A Compassionate Agent

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:

$f(\phi)$ H(ϕ) :P → IH($f\phi$): P
H(ϕ) :N→ IH($f\phi$): N

$\neg f(\phi)$ H(ϕ):P → IH($\neg f\phi$): N
H(ϕ):N → 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.

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 $\mathfrak{I}$ as follows:

$$\mathfrak{I}\phi : f(\phi) \vee \neg f(\phi)$$

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 $\neg\mathfrak{I}$ as follows:

$$\neg\mathfrak{I}\phi : \neg(f(\phi) \vee \neg f(\phi))$$

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 $\mathfrak{I}$ is undefined**.**

The presence or absence of $\mathfrak{I}$ 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 later 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 $f\phi$ 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 feeling of 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.) 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.)

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 when feelings change. A special problem may arise in distributed multi-agent systems when agents use communicated feelings in their reasoning processes. Compared to logical belief, feelings canbe 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:

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

Summary

We have introduced several new concepts regarding the challenges AI researchers face in reaching human-level AI. Namely, computing with compassionate intelligence and using affective inference as a means of common sense default reasoning. To accomplish this we introduce the idea of an Irrational-TMS where psychologically based justifications may be used to support belief or feelings about a proposition as well as traditional logic. Meta-cognition is central to each of these. We believe that compassionate intelligence is a necessary part of future applications involving social interactions or healthcare.

Thanks

Thanks to Aaron Sloman for early support when I first began working on EOP and for his paper with Monica Croucher entitled "Why Robots Will Have Emotion" [Sloman and Croucher 1981]. Thanks to John McCarthy for years of arguing that only made me more convinced of the need for affective inference and for access to his great library where I found many books by common sense philosophers. Thanks also to Henry Lieberman, to Waldinger for introducing me to Vipassana 9 years ago, to Michael Cox for his insight, to many kind people who helped over the years, as well as the staff at AAAI without whom none of this would come together.

References

Ariel, E. and 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.; and Epstein, M. 1982. Body temperature changes during the practice of g Tummo yoga. *Nature Magazine*, 295: 234 – 236.

Carlson, L.; Ursuliak, Z.; Goodey, E.; Angen, M.; and Speca, M. 2001a. 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.

Cox, M., and Raja, A., 2008a. Proceedings of the Workshop on Meta Reasoning – Thinking about Thinking, American Association for Artificial Intelligence, Forthcoming.

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.; and Sheridan, J. 2003. Alterations in brain and immune function produced by mindfulness meditation. *Psychosomatic Medicine* 65 (4): 564-570.

Kabat-Zinn, J.; Lipworth, L.; and 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.; and 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. In Proceedings of the Psychoneuroimmunology Research Society. Florida:. International Conference on Psychoneuroimmunology.

Mason, C. 1995. Emotion Oriented Programming. Formal Notes, SRI AI Group. Also www.emotionalmachines.org.

Mason, C. 2004. Global Medical Technology. Proceedings of the Conference on Future Health Technology. Boston, Mass.

Mason, C. 1995. Introspection As Control in Result-Sharing Assumption-Based Reasoning Agents. *Proceedings of the First International Workshop on Distributed Artificial Intelligence*, Lake Quinalt, Wash.

Sloman, A., and Croucher, M. 1981. Why robots will have emotions. *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, Vancouver, Canada.