**Can AI Develop Empathy & Ethics?** Commentary by Mel Head <melheadgmail.com>

I think the first thing we need to understand with respect to AI in our lives is for us to learn to live with it. Just as being with a strange person in a strange place, we need to be aware of how we conduct our interactions with that person. We cannot abandon common sense and critical thinking and simply go with what “sounds right” to us at the time.

Dr. Diamandis mentions the movie **“Her”.** I have not seen it but will watch it one day as it may provide some examples of where empathy and other qualities may be required in an AI. . One thing that AI will do for many people is to present information to them that may challenge their thinking. This is perhaps because what the {generative} AI can provide is an organized presentation of information that addresses the user’s query. That information might have otherwise taken the user a long time to obtain and compile themselves as well as perhaps learning new skills to allow proper interpretation of such information. In fact, there may well be situations where a user would have never been able to accomplish in a lifetime, a full understanding of what the AI presents in mere seconds.

AI today can “watch” and “listen” to a human and build a profile of that person according to the metrics its algorithms and multi-level neural networks support. We can all, in a sense, be “pigeon-holed” by the AI.  This is not much different than what we do ourselves when we meet a stranger. We do formal and informal assessments of the stranger’s characteristics, communication skills, strengths, weaknesses, and abilities. When we get to a self-defined point of familiarity, we may raise the level of the relationship, so the stranger now becomes an acquaintance and later, some such individuals become friends, confidants, and perhaps partners. Other individuals may no longer garner our interest or desire to pursue further interaction. The AI may not consider "us” as an entity with which it would seek further interaction, but it may use what it “learns” from its interactions with us, to guide how it may react with us in the future.

Is this “*digital* emotional intelligence”? Will it get to “know us better” and govern itself {programmatically} to better interact with us in our next encounter? At present, science fiction aside, this could only be done through the capabilities of its algorithms and neural network(s). If we define this “knowing us better” as “emotional intelligence” or “AI emotions”, we could muse about its developing “AI empathy”. That’s said, empathy evolves from our having previous experience with the issue at hand or having had an arms length exposure to the issue. The {generative} AI, regardless of whether it expresses itself as text on a screen, verbally, or mechanically (e.g., a *robot)*, it would still be driven by its internal software.

AI technology, coupled with video and audio input, is capable today of greater and faster monitoring and analysis of a human’s reactions than is a human able to do so when interacting with another human. Just as our eyesight uses very few queues as input to our brain which creates the images we see as our reality, and AI may have the capacity to do this on a multidimensional scale at speeds we could never attain. Again, we need to understand that the human brain, although not that fast, is massively parallelin what it does. An AI, while its design may accommodate parallel activity, is still a long way from being self-capable of conducting its interaction and activity with us in a way we can only describe as “human-centred”. This may be more of a human “feeling” than something we can define in a pragmatic sense. The adage, “if it walks like a duck and quacks like a duck”, we may conclude ‘*it must be a duck’* may come to mind here but that conclusion is not absolute.

Having mentioned this before, we continue to use “human” terms to describe certain qualities and capabilities of an AI. When we ask if a child is ‘intelligent’ and later ask if a dog is intelligent, is our reference to the word “intelligent” meant in the same way? We expect a child to talk, read and write because the child is 'intelligent'. We don't expect this from a dog but yet, because of some things they do, we refer to the animal as being 'intelligent'. Now we are doing the same for AI systems. It is unlikely that we mean the same thing although we use the same descriptor. Considering that, why would we describe an AI as being ‘intelligent’, having ‘emotion’, being ‘rational’, possessing empathy, being “ethical”, or having any other quality of a human? We need a new vocabulary when discussing or comparing human traits to those of a generative AI. These entities are not human and cannot be treated or referred to as such... but is that always true if we want to work well with them? Moving in the longer term towards General AI would mean to build upon success with some of the more manageable enhancements being done first. We are approaching the point where Large Language Model AI systems appear to “act” or “respond” like a human and humans are treating them as such. Dr. Diamandis’ reference to author Mo Gawdat’s statement that we are “*raising a new species of intelligence*” is perceptive. AIs will learn the “basics” from human input over history and help us understand ourselves in the long run. In a sense, we should treat “them” as we would wish to be treated.

We build AI entities such that it makes it easy to interact with them… in our terms. If it was possible, how would they choose to interact with us? Perhaps “recursive action”, as a capability built into the design of an AI might extend its settings of internal ‘reward’ metrics such that it could ‘discover’ that maximizing or minimizing such ‘reward’ metrics may elicit better responses or reactions from a human, allowing the AI to achieve a ‘goal’… but who (or what) defines and sets that goal. An example in health care might be to have a conversational AI that converses with a patient and is provided a goal to diagnose the patient and inform the doctor of the diagnosis, recommended treatment, and prognosis. This would mean the ‘goal’ would be defined such that the AI is told to be the doctor. Extending this thought, maybe the AI could simply begin a conversation with a person and, by analysis of the interaction, conclude that the individual suffers from some ailment or condition. The AI would then set its ‘goal’ to diagnose the patient for the doctor who would of course now focus on the specific findings. In a health setting, the AI may have been provided with access to current lab work results and other information in the person’s history to ensure better accuracy in its analysis.

In an Education environment, the AI may be provided with access to a student’s records and, by comparing them to the appropriate curriculum, could recommend specific actions to raise the student’s academic level. In a more general sense, this student-specific educational advisor may determine the student’s needs based upon a conversation with the student. The student could be provided with certain prescribed exercises and the teacher would be provided specific directives to help the student improve his or her knowledge. Subsets of goals in each situation could be selected by the AI to accomplish the best results for each situation.

A neurosurgeon in Calgary with whom I spoke at a "Back Health" event about six years ago, told me that he was in the early stages of developing an "AI Triage Management System". From a high level, an individual experiencing back pain would set up an account and be interrogated by an AI agent about the nature of the pain they were experiencing. The nature of the "conversation" would change depending upon what or how the potential patient described their back problem. The AI would compile the input, contrast it against similar data and provide an initial assessment which would be sent to the individual's family doctor. From there, the doctor may add comments to the report and, depending upon the severity of the assessment, a neurosurgeon would be notified of the "file". After reviewing, the neurosurgeon would activate the file for follow-up with the patient and this might evolve to an in office visit. This process might range from the system contacting the patient in a few months for an update and a potential reassessment. At some point, the patient would be placed on the surgical "Triage List" and their position on that list would be managed by the AI and the neurosurgeon at times of change in the patient's subsequent interaction with the AI. This would help free up the neurosurgeon and offer some assurance to the patient that they were moving up the Triage List to the level appropriate for their stage of required intervention. I do not know if this was implemented. I note this example here because if an AI is to interact with a patient, there will need to be some level of "trust" on the patient's part and likely the AI would need to interact with the patient in a compassionate and empathetic-like manner.

**In summary,** until we truly understand our human qualities, it may seem unproductive to pursue study of the inefficiencies of an AI with respect to qualities and capabilities of a human. In Dr. Diamandis’ article, the AI researcher (Rana el Kaliouby) rightly stated that we can ‘simulate’ emotional intelligence and empathy, developing them accordingly, but deploying them with the certainty that they “fit” those of a human is currently quite remote. The comment about bias with respect to algorithms is still real and needs to be further addressed. The sheer vastness of information used to train and validate an AI cannot preclude any inference of bias and can not be processed such that all bias is removed. This task would most likely fall upon humans to determine what constitutes ‘bias’ however that effort could introduce new bias.

I think the first thing we need to understand with respect to AI in our lives is for us to learn to live with it. Just as being with a strange person in a strange place, we need to be aware of how we conduct our interactions with that person. We cannot abandon common sense and critical thinking and simply go with what “sounds right” to us at the time.

Dr. Diamandis mentions the movie **“Her”.** I have not seen it but will watch it one day as it may provide some examples of where empathy and other qualities may be required in an AI. . One thing that AI will do for many people is to present information to them that may challenge their thinking. This is perhaps because what the {generative} AI can provide is an organized presentation of information that addresses the user’s query. That information might have otherwise taken the user a long time to obtain and compile themselves as well as perhaps learning new skills to allow proper interpretation of such information. In fact, there may well be situations where a user would have never been able to accomplish in a lifetime, a full understanding of what the AI presents in mere seconds.

AI today can “watch” and “listen” to a human and build a profile of that person according to the metrics its algorithms and multi-level neural networks support. We can all, in a sense, be “pigeon-holed” by the AI.  This is not much different than what we do ourselves when we meet a stranger. We do formal and informal assessments of the stranger’s characteristics, communication skills, strengths, weaknesses, and abilities. When we get to a self-defined point of familiarity, we may raise the level of the relationship, so the stranger now becomes an acquaintance and later, some such individuals become friends, confidants, and perhaps partners. Other individuals may no longer garner our interest or desire to pursue further interaction. The AI may not consider "us” as an entity with which it would seek further interaction, but it may use what it “learns” from its interactions with us, to guide how it may react with us in the future.

Is this “*digital* emotional intelligence”? Will it get to “know us better” and govern itself {programmatically} to better interact with us in our next encounter? At present, science fiction aside, this could only be done through the capabilities of its algorithms and neural network(s). If we define this “knowing us better” as “emotional intelligence” or “AI emotions”, we could muse about its developing “AI empathy”. That’s said, empathy evolves from our having previous experience with the issue at hand or having had an arms length exposure to the issue. The {generative} AI, regardless of whether it expresses itself as text on a screen, verbally, or mechanically (e.g., a *robot)*, it would still be driven by its internal software.

AI technology, coupled with video and audio input, is capable today of greater and faster monitoring and analysis of a human’s reactions than is a human able to do so when interacting with another human. Just as our eyesight uses very few queues as input to our brain which creates the images we see as our reality, and AI may have the capacity to do this on a multidimensional scale at speeds we could never attain. Again, we need to understand that the human brain, although not that fast, is massively parallelin what it does. An AI, while its design may accommodate parallel activity, is still a long way from being self-capable of conducting its interaction and activity with us in a way we can only describe as “human-centred”. This may be more of a human “feeling” than something we can define in a pragmatic sense. The adage, “if it walks like a duck and quacks like a duck”, we may conclude ‘*it must be a duck’* may come to mind here but that conclusion is not absolute.

Having mentioned this before, we continue to use “human” terms to describe certain qualities and capabilities of an AI. When we ask if a child is ‘intelligent’ and later ask if a dog is intelligent, is our reference to the word “intelligent” meant in the same way? We expect a child to talk, read and write because the child is 'intelligent'. We don't expect this from a dog but yet, because of some things they do, we refer to the animal as being 'intelligent'. Now we are doing the same for AI systems. It is unlikely that we mean the same thing although we use the same descriptor. Considering that, why would we describe an AI as being ‘intelligent’, having ‘emotion’, being ‘rational’, possessing empathy, being “ethical”, or having any other quality of a human? We need a new vocabulary when discussing or comparing human traits to those of a generative AI. These entities are not human and cannot be treated or referred to as such... but is that always true if we want to work well with them? Moving in the longer term towards General AI would mean to build upon success with some of the more manageable enhancements being done first. We are approaching the point where Large Language Model AI systems appear to “act” or “respond” like a human and humans are treating them as such. Dr. Diamandis’ reference to author Mo Gawdat’s statement that we are “*raising a new species of intelligence*” is perceptive. AIs will learn the “basics” from human input over history and help us understand ourselves in the long run. In a sense, we should treat “them” as we would wish to be treated.

We build AI entities such that it makes it easy to interact with them… in our terms. If it was possible, how would they choose to interact with us? Perhaps “recursive action”, as a capability built into the design of an AI might extend its settings of internal ‘reward’ metrics such that it could ‘discover’ that maximizing or minimizing such ‘reward’ metrics may elicit better responses or reactions from a human, allowing the AI to achieve a ‘goal’… but who (or what) defines and sets that goal. An example in health care might be to have a conversational AI that converses with a patient and is provided a goal to diagnose the patient and inform the doctor of the diagnosis, recommended treatment, and prognosis. This would mean the ‘goal’ would be defined such that the AI is told to be the doctor. Extending this thought, maybe the AI could simply begin a conversation with a person and, by analysis of the interaction, conclude that the individual suffers from some ailment or condition. The AI would then set its ‘goal’ to diagnose the patient for the doctor who would of course now focus on the specific findings. In a health setting, the AI may have been provided with access to current lab work results and other information in the person’s history to ensure better accuracy in its analysis.

In an Education environment, the AI may be provided with access to a student’s records and, by comparing them to the appropriate curriculum, could recommend specific actions to raise the student’s academic level. In a more general sense, this student-specific educational advisor may determine the student’s needs based upon a conversation with the student. The student could be provided with certain prescribed exercises and the teacher would be provided specific directives to help the student improve his or her knowledge. Subsets of goals in each situation could be selected by the AI to accomplish the best results for each situation.

A neurosurgeon in Calgary with whom I spoke at a "Back Health" event about six years ago, told me that he was in the early stages of developing an "AI Triage Management System". From a high level, an individual experiencing back pain would set up an account and be interrogated by an AI agent about the nature of the pain they were experiencing. The nature of the "conversation" would change depending upon what or how the potential patient described their back problem. The AI would compile the input, contrast it against similar data and provide an initial assessment which would be sent to the individual's family doctor. From there, the doctor may add comments to the report and, depending upon the severity of the assessment, a neurosurgeon would be notified of the "file". After reviewing, the neurosurgeon would activate the file for follow-up with the patient and this might evolve to an in office visit. This process might range from the system contacting the patient in a few months for an update and a potential reassessment. At some point, the patient would be placed on the surgical "Triage List" and their position on that list would be managed by the AI and the neurosurgeon at times of change in the patient's subsequent interaction with the AI. This would help free up the neurosurgeon and offer some assurance to the patient that they were moving up the Triage List to the level appropriate for their stage of required intervention. I do not know if this was implemented. I note this example here because if an AI is to interact with a patient, there will need to be some level of "trust" on the patient's part and likely the AI would need to interact with the patient in a compassionate and empathetic-like manner.

**In summary,** until we truly understand our human qualities, it may seem unproductive to pursue study of the inefficiencies of an AI with respect to qualities and capabilities of a human. In Dr. Diamandis’ article, the AI researcher (Rana el Kaliouby) rightly stated that we can ‘simulate’ emotional intelligence and empathy, developing them accordingly, but deploying them with the certainty that they “fit” those of a human is currently quite remote. The comment about bias with respect to algorithms is still real and needs to be further addressed. The sheer vastness of information used to train and validate an AI cannot preclude any inference of bias and can not be processed such that all bias is removed. This task would most likely fall upon humans to determine what constitutes ‘bias’ however that effort could introduce new bias.