“Dave... I feel it”

A Nietzschean idea randomizer

Paulo Blikstein
May 2001
We want to make machines that will be proud of us.

Danny Hillis apud Marvin Minsky, 1983
Introduction

Although it seems a common place to start a paper about artificial intelligence with a picture of the famous HAL 9000 computer, it is insightful to remember why it has been discussed since 1968, when Kubrick’s “2001 – A Space Odyssey” was released.

Most of the discussion about HAL malfunction during the Jupiter mission, killing almost the whole crew of the spaceship, is about the liability of the error: was it human mistake or a computer malfunction?

In the sequel novel “2010”, Arthur C. Clarke tries to explain some of the mysteries of “2001”. He declares that the computer was programmed to protect the mission and to protect the crew, and that the decision of killing the crew was just a programming bug, and not an emotional response.

In the 1965 script of the movie, greatly modified until the final version, we find an interesting dialogue:

**BOWMAN**
Hal, I’m in command of this ship. I order you to release the manual hibernation control.

**HAL**
I’m sorry, Dave, but in accordance with sub-routine C1532/4, quote,
When the crew are dead or incapacitated, the computer must assume control, unquote. I must, therefore, override your authority now since you are not in any condition to intelligently exercise it.

Clarke, a science-fiction writer, was clearly optimistic about the future of technology and Artificial Intelligence. Kubrick, who gave the last word in the movie, had clearly other concerns. Whereas Kubrick’s film suggests that HAL is flawed because he has so many human traits, in Clarke’s vision HAL is a technologically perfect victim of human greed.

The result was the everlasting HAL mystery: was it the programmer’s mistake? Did the computer have real feelings? Is that a good thing for a computer to have? Was he really afraid to die?
But why is this mystery so infamous? Why do we feel uncomfortable listening to his monotonous and assertive voice?

**HAL**
I'm not questioning your word, Dave, but it's just not possible. I'm not capable of being wrong.

In the first place, Kubrick emphasizes that HAL was pretentious because of its creators, who believed that he could never fail. The disturbing issue is that HAL is strangely familiar: a mixture of human arrogance and machine perfectionism. But as it is just a computer, we feel free to criticize it as a non-human entity. However, it is no news that it is certainly the most emotional character in the movie.

The public debate over Artificial Intelligence almost always goes towards this same monotonous question: can a computer do everything a human being does? Interestingly enough, there are many different parts of this huge task:

- Can computers be intelligent?
- Can computers feel emotions?
- Can computers learn?
- Can computers perform very complex tasks?

Those questions are often mixed up and confused, resulting in a general disbelief about the real possibilities of Artificial Intelligence, as Marvin Minsky declares:

> It is too easy to say things like, "Computer can't do (xxx), because they have no feelings, or thoughts". But here's a way to turn such sayings into foolishness. Change them to read like this. "Computer can't do (xxx), because all they can do is execute incredibly intricate processes, perhaps millions at a time". (Minsky 1982)

This is a quite revolutionary view on the problem, because it removes the ‘magical’ component often credited to ‘thinking’ and ‘emotion’ and puts it on the scientific battlefield. Nevertheless, in exactly which battlefield are
we? As Franchi states in “Mindless mechanisms, mindful constructions”,
commenting the foreword to “AI in the 1980s and Beyond: An MIT survey”,

More specifically, suppose it is also true, as these authors claim, that AI
theories are, and theories from the humanities are not, sufficiently pre-
cise to be implementable. Once again, no conclusion follows that the
implementable theories are right – or even that they are better. It is
equally possible – and in historical retrospect seems likely – that the
theories embraced by AI were too strictly stated, too narrow to do justice
to the phenomenon of intelligence. Sure enough, there may be light
under the computational lamp (in its present incarnation), but are we
sure the key to intelligence is to be found there? (Franchi 1995)

This sharp division between ‘humanists’ and ‘AI researchers’ created an
artificial boundary between humans and machines, AI and philosophy, sci-
ence and spirituality, as if Artificial Intelligence would only be useful if to-
tally autonomous machines could operate, learn and ‘think’.
This paper will try to discuss some aspects about the difference between
the knowledge of models and understanding of actual mechanisms – a
key point in building more complex machines. Also we will propose an
idea for an application for Artificial Intelligence technology, trying to com-
bine virtues of both worlds: the human brain and electronic computation.
Another goal, in the meta-linguistic level, is to meta-comment and criticize
the paper using a fictitious character (the “critic”). This idea is imitated
from Marvin Minsky’s drafts of “The Emotion Machine”, with one only dif-
ference: we won’t reply the critic.
Critic: What’s the sense of putting me on this article if you are not going to re-
ply? I have other things to do.
Author: Sorry, I can’t reply. I have to keep my word.
Critic: I was making a lot of money working for Marvin. I’ll go back to his book.
Carol: this is not a pipe

Critic: Oh no! There you go again with Magritte!

One important issue when discussing human emotion is the difference between model and mechanism. Finding out the right model doesn’t mean that we know how the mechanism works. Let’s take the example of Carol, from the second draft chapter of “The Emotion Machine” (Minsky 2001). The example given by Marvin Minsky is of a little girl playing with mud:

Let’s imagine a child named Carol. Carol has wandered away from her mother. Equipped with a fork, a spoon, and a cup, she has found a place that has plenty of mud. Her make-believe goal is to make a cake, the way she’s seen her mother do. Here are three versions of what might take place.

(1) **Playing Alone**: She wants to fill her cup with mud, and first tries to do this with her fork, but this fails because the mud slips through. She feels frustrated and disappointed. But when she succeeds by using the spoon, Carol feels satisfied and pleased. (Minsky 2001)

From this simple situation we could derive a model of knowledge building, based on trial and error. Minsky goes on explaining more complex models for learning, but let’s focus on this first idea. An important assumption is that the Carol’s choices are based on functional characteristics of the objects: the spoon may seem a better object to play with mud, or it is simply the only other object available other than the fork.

Here the possible difference between model and mechanism comes into place. Let’s consider the Freudian concept of three distinct entities in the mind (Superego, Ego, Id). He states that our choices and ideas are not totally controlled by the Ego, but rather by complex relationships among the three entities.

As Minsky states in “Why people think computers can’t”:
When a person makes a good decision, we tend to ask what "line of thought" lies behind it. But we don't so often ask what thousand prohibitions might have warded off a thousand bad alternatives. If censors work inside our minds, to keep us from mistakes and absurdities, why can't we feel that happening? Because, I suppose, so many thousands of them work at once that, if you had to think about them, you'd never get much done. They have to ward off bad ideas before you "get" those bad ideas. (Minsky 1982)

This could be an excellent strategy for humans, making it possible to operate in the world and especially in a society. Knowing this model, nevertheless, doesn't mean that we can reproduce the mechanism, as Magritte suggests in his famous picture, “La trahison des images” (The betrayal of images). The dangerous step forward comes exactly here. Carol may have chosen the spoon because it reminds her of the breasts of her mother, or because it could have a phallic shape. In other words, the reason for our choices are very much dependent on mostly unknown and hidden factors – even for the Psychoanalysts. Even knowing them, and having an approximate model, how could you reproduce them in a computer? If we consider it as a kind of hidden memory/processing area, there are no computational methods to handle this very complex concept. How could a computer make decisions based partially on a ‘hidden’ processing system, with its own rules and memory, but without knowing that it exists? What is the mechanism that allows this to happen in our mind? Is the connection between the conscious and the unconscious random? How do the agencies that constitute them exchange information? Are they separate resources?

So, as almost all our emotional life (as well as the ‘rational’, if such a distinction can be made) works through the Ego/Id/Superego model, even if we understand the high level model very well it seems impossible, as for now, to reproduce the machinery.

Another interesting example from “The Emotion Machine” (originally cited in “The Society of Mind” (Minsky 1988)) is the Sleep/Work/Anger model. We may overcome sleep and continue working by invoking anger (of a competitor, for instance). This very ingenious example could also fall in the
same category: maybe we really manage to overcome sleep by feeling challenged by a competitor and working more, but maybe this competitor reminds our father, or a childhood enemy, and that’s the real cause of the rivalry. Without this additional ‘reason’, the ‘anger’ agent wouldn’t be able to overcome sleep. But if we are conscious of the similarity between this competitor and our childhood enemy, we may evaluate the situation again and realize that the competition is not that important, and that we are just reviving a bad memory from childhood. How could the computer reproduce this mechanism? Pretending not to know something that has to be somewhere in its memory? If we arrive at inventing such a software artifact, we may get really close to a better understanding of human and machine emotions, but it doesn’t seem to be the case for now. Maybe better brain scanners could answer many of these questions, but it seems that also reinforcing the transdisciplinary approaches to Artificial Intelligence is another very important step, as Agre apud Franchi concludes:

Philosophy and other traditional disciplines, in sum, provide overall theoretical frameworks. AI, in turn, provides a powerful means of forcing into the open their internal structures and internal tensions. What is needed to make substantial advances in both fields, Agre concludes, is a constructive symbiosis of AI research with humanistic analyses of ideas.

(Franchi 1995)

Critic: OK, this claim of putting philosophy, social sciences and AI together is an old idea, but their methods, goals and culture seem so removed from one another that it is hard to believe that they could really work together.

The forbidden planet

This idea is somehow discussed in the movie “The Forbidden Planet” (1956). Updating Shakespeare’s The Tempest to a science-fiction setting, it is an elegant and intelligent film.

Headed by Commander Adams
(Leslie Nielsen), a patrol ship from Earth is just arriving at Altair after a year in hyperspace. Their mission is to investigate the fate of an exploration vehicle, which landed on the fourth planet 20 years ago and was never heard from again. The ship comes within range of the planet and achieves voice contact with a survivor. Identifying himself as Dr. Morbius (Walter Pidgeon), the scientist explains that all is well and that there's no need to land. Pressed a little further, Morbius reveals how the rest of the crew are long dead and that he can't possibly be responsible for their safety. Intrigued, Adams orders a landing and prepares to investigate the mystery. Shortly after landing, the ship is attacked by a strange invisible monster, which kills one of the group. By the end of the movie, we finally find out the truth. The planet was once inhabited by a very advanced civilization that build a large and powerful machine that would store all the information contained in their ‘brains’. This machine would be a kind of omnipotent computer with powers and talents from every individual in the civilization. After completing it, an invisible and indestructible monster begins to attack and destroy everything, until the whole planet was desert. Then the final revelation comes: the creature was the “monster of the Id”, or the concretization of the Freudian unconscious of all the population. The big machine was meant to be omnipotent, but they haven’t anticipated what would come out of a concretization of the mind: both the Ego and the Id. This allegory invokes the myth that the human brain is a good machine, and reproducing it is a very good idea. Again, depending on the point of view, we can end up with very different views. On one hand, the human brain is capable of love, forgiveness, altruism, Mozart’s 41st symphony and Homer’s “Odyssey”. On the other hand, humans are selfish, egocentric, voracious, greedy and violent.

As Mazlish apud Franchi asserts,

Such a radical question, in its ultimate impossibility, is bound–Mazlish affirms after Freud - to arouse “the same range of ambivalent reactions: the sense of a perfection and infallibility to which we aspire - the angel in us - and the sense of the destructive and degrading in us - the ape in us.” (Franchi 1995)
Marvin Minsky, in “Logical vs. Analogical or Symbolic vs. Connectionist or Neat vs. Scruffy” (Minsky 1990), also points out that we have several bugs:

- Obsessive preoccupation with inappropriate goals.
- Inattention and inability to concentrate.
- Excessively broad or narrow generalizations.
- Excessive accumulation of useless information.
- Superstition; defective credit assignment schema.
- Unrealistic cost/benefit analyses.
- Inability to deal with exceptions to rules.
- Unwillingness to acknowledge loss.
- Depression or maniacal optimism.

There seems to be a belief that scientists can ‘understand’ and ‘isolate’ the good parts of the brain and reproduce them. “The forbidden planet” reminds us that this plan greatly overestimates our capacity to evaluate good and bad things about human behavior. What is more, it reminds us of another issue: is the human brain a good model to depart?

"{it} helps us see the ways in which machines are better than humans: it shows that many of the things that humans do, they do because they are unable to function in the machine-like way that they would prefer. It also shows that there are many activities where the question of machine replacement simply does not arise, or arises as, at best, an asymptotic approximation to human abilities." (Franchi 1995)

Critic: we are not talking about ‘good’ or ‘bad’ parts. This is a very simplistic classification. We are just trying to build models and understand how the brain works. No one is talking about building replicants or omnipotent machines.

I have no words... (mechanism vs. externalization)

The phenomenological explanation of human emotions (criticized by Minsky), based on its externalizations (which is in vogue nowadays) could also be a dangerous and misleading path. One of the reasons is language: because we don’t even know if the words, which we use to classify them, have any relationship at all with the mental processes involved.
What is more, these classifications are very culture-dependent: feelings and emotions are not universal. Maybe our concept of ‘love’ could be considered as ‘pity’ in another culture; or it may not exist at all.

Critic: there you go again with this multi-cultural-awareness discourse.

Trying to use the ‘names’ we give to our emotions as the basis of a model can be a good way to better understand our emotions, but may not lead to a universal computer program that has feelings by its own. As Marvin Minsky writes in “The Society of Mind” [Minsky, 1988 #2], our adult minds have been reconstructed so many times that we hardly remember how it is like to be a baby or a child.

We often consider that emotions are a survival strategy: we could not cope with life without them. As a result, it seems that the study of human emotions could be more important for understanding ourselves than to reproduce them in a machine. After all, are emotions important at all for computation? One example, often cited by Rosalind Picard (Blikstein 2000), from the Affective Computing Group of the MIT Media Lab, is that computers could be more efficient if they could get ‘frustrated’ after an extended period of time trying unsuccessfully to solve a problem. It seems that this category of example considers human emotions as ‘models’ or ‘inspiration’ for computational functions. Having a procedure to measure the amount of time spent in a problem could be a metaphor for frustration, but the opposite ‘direction’ is not applicable: we cannot state that this procedure has anything to do with the way humans get frustrated. Just as an example, we may pursue solving a problem just because we have something else to do that is unconsciously disturbing.

Again, understanding the model of frustration doesn’t mean that we automatically know how to rebuild the same mechanisms in a machine.

What is more, emotion is often taken in a very functionalist approach, as if it was a ‘tool’ manipulated by artists, as Rosalind Picard states in the book “HAL’s Legacy” (Stork 1997):

> Emotion is not simply a luxurious extra in 2001. In film and theater, carefully controlled and expressed emotion has the ability to influence us
and, subsequently, to affect whether we like a production and remember seeing it with pleasure. [...] Emotion is a powerful tool in the hands of artists like Kubrick and Clarke. But what about emotions in computers - for purposes other than entertainment?

In the same book, Marvin Minsky states brilliantly that:

**Minsky:** Well, the person on the street at the time the film was made both loved computers and was scared of them. It's worth mentioning that there are no benevolent super intelligent computers in film – except, I suppose, in The Day the Earth Stood Still.

**Stork:** Do you think that is the nature of high intelligence?

**Minsky:** No, I think that's the nature of Hollywood.

Another important criticism, expressed by Shamanski in “Connectionism, Confusion and Cognitive Science”, about his view on some of the connectionist research in Artificial Intelligence:

Johnson-Laird has noted that “to understand a phenomenon is to have a working model of it”. Interestingly, PDP models appear to prove this statement false, because connectionists can easily replace one unknown (e.g., how the brain mediates some psychological phenomenon) with another – a functioning but unexplained network. The story of Hans [the German horse] shows that it is very easy to be beguiled by interesting behaviour. Researchers must be careful to remember one can generate interesting behaviour in a PDP networks without understanding how the network actually works. (Shamanski 1994)

### A Nietzschean idea randomizer

A very inspiring thought about the nature of intelligence and genius come from Friedrich Nietzsche, specially in his masterpiece “Human, All Too Human: a book for free spirits” [Nietzsche, 1984 #14]. Considering the relevance to this paper, we will reproduce a long passage for the book that inspired the Nietzschean idea randomizer.

Perfection said not to have evolved. When something is perfect, we tend to neglect to ask about its evolution [...] as if it had risen from the ground by magic. The artist knows that his work has its full effect only when it arouses belief in an improvisation [...] and so he encourages this illusion [...] As is self-evident, the science of art must oppose this illusion [...] Believe in inspiration. Artists have an interest in others' believing in sud-
den ideas, so-called inspirations; as if the idea of a work of art, of poetry, the fundamental thought of a philosophy shines down like a merciful light from heaven. In truth, the good artist's or thinker's imagination is continually producing things good, mediocre, and bad, but his power of judgment, highly sharpened and practiced, rejects, selects, joins together [...]. The artist who separates less rigorously, liking to rely on his imitative memory, can in some circumstances become a great improviser [...] All great men were great workers, untiring not only in invention but also in rejecting, sifting, reforming, arranging. [...] Once again inspiration. When productive energy has been dammed up for a while and has been hindered in its outflow by an obstacle, there is finally a sudden outpouring [...] as if a miracle were taking place. This constitutes the well-known illusion which all artists [...] have somewhat too great an interest in preserving. The capital has simply piled up. Worshipping the genius out of vanity. Because we think well of ourselves, but in no way expect that we could ever make the sketch to a painting by Raphael or a scene like [...] Shakespeare, we convince ourselves that the ability to do so is quite excessively wonderful, a quite uncommon accident [...]. Thus our vanity, our self-love, furthers the worship of the genius, for it does not hurt only if we think of it as very remote from ourselves, as a miracle. [...] But [...] the activity of the genius seems in no way fundamentally different from the activity of a mechanical inventor [...]. All these activities are explained when one imagines men whose thinking is active in one particular direction; who do not tire of rearranging their material. The genius, too, does nothing other than first learn to place stones, then to build, always seeking material, always forming and reforming it. Every human activity is amazingly complicated, not only that of the genius: but none is a "miracle." [...] Furthermore, everything that is complete and perfect is admired; everything evolving is underestimated.

This part of the paper is dedicated to describing the proto-idea of the idea randomizer.

One of the most important ideas of the Section Four of “Human, All Too Human” is that “the good artist's or thinker's imagination is continually producing things good, mediocre, and bad, but his power of judgment, highly sharpened and practiced, rejects, selects, joins together”. Nietzsche’s idea is that the quality of the work of a so-called genius is not due to an innate characteristic or divine gift, but to a very sophisticated power of judgment and constant flow of ideas. Minsky has a very similar view:
Do outstanding minds differ from ordinary minds in any special way? I don’t believe that there is anything basically different in a genius, except for having an unusual combination of abilities, none very special by itself. There must be some intense concern with some subject, but that’s common enough. There also must be great proficiency in that subject; this, too, is not so rare; we call it craftsmanship. As I see it, any ordinary person who can understand an ordinary conversation has already in his head most of what our heroes have. So, why can’t “ordinary, common sense” - when better balanced and more fiercely motivated - make anyone a genius? (Minsky 1982)

One direction of Artificial Intelligence research is to explore the boundaries between the natural and the artificial. As Sack states,

Margolin critiques the artificial/natural binary assumed by Simon and argues that it is exactly at the border of these two terms that the work of design (and design criticism) is to be done. [...] Margolin proposes that design work should be aimed at complementing (rather than replacing) the "natural," and, furthermore, that design should be based on a commitment to the spiritual. (Sack 1997)

Considering Nietzsche view, we considered one great advantage of computation: as computers have no Id, no unconscious thoughts that ‘filter’ and ‘censor’ ideas, one can state that they are much better randomizers. Human brains, conditioned by culture, moral values, day-to-day life, are always risking stagnation. New and unexpected connections are less and less likely to happen for most of us, as we go deeper in one field of knowledge or in one professional activity. Having an external randomizer could be a helpful device, as our life seems increasingly conventional, mediated, controlled, monitored, andalienated. It could help us collecting, remembering and recombining ideas, as well as changing our approach to creativity and art: instead of the view of the impossible genius (criticized by Nietzsche), we approach the view that learning to collect, carefully select and combine ideas is the best path to creation.

Critic: do you think that an annotation device can make a person a genius?
As Marvin Minsky declares, the most important part of the meaning of things is its connection to other ideas:

The secret of what something means lies in the ways that it connects to all the other things we know. The more such links, the more a thing will
mean to us. The joke comes when someone looks for the "real" meaning of anything. For, if something had just one meaning, that is, if it were only connected to just one other thing, then it would scarcely "mean" at all! (Minsky 1982)

So, isn’t exactly the lack of emotion, repression and superego an advantage for helping people connecting ideas which connection could have been filtered out or repressed in or mind? In a way, that’s in part the method of Freudian psychoanalysis: trying to put random pieces of discourse to assemble the puzzle of a person’s mind.

Critic: if one combination of ideas was filtered before, it could well happen again, no matter if you see it in your little device.

Maybe the most important part of the creative process of the artist is the artistic intention: even the same exact poem (one written by a poet and the other by a computer) would have very different impacts. Art, invention and authorship are closely linked together. Instead of introducing all of human defects in computers, so that they could start touching the realm of art, it would be better to use computers to feed us with random ideas, combining and recombining them ad infinitum.

How Artificial Intelligence research could help this? Two functionalities would be important:

- A suitable format to store and evaluate ideas.
- A true randomizer.

There are already some systems that try to store ‘common sense’ ideas, as the OpenMind project at the MIT Media Lab. Maybe using a similar database structure, the user would be able to write (or speak) random ideas and comments during the day, which would be then indexed in a proper way. A couple of times a day, the user should evaluate his own ideas, regarding quality and subject. The technologies for all of these tasks are already widely available.

Then comes the second moment – the random walk through the ideas. The system would put together two or three of them, trying to combine very good and very bad ideas together, as well as ideas from different
fields: zoology and mathematics, computation and dance, chemistry and linguistics, sculpture and social sciences, soccer and wavelets.

Some of the important characteristics of the system would be:

- The ideas must not be shared on the Internet. The concept is not to provide a worldwide resource of ideas, but rather a personal companion, capable to randomize when our brains are too stagnated (or tired) to do so. At most, in special occasions, we would be able to send a set of random ideas to a friend.

- The bad ideas should never get out of the system. There should be no way of deleting an idea. Once we stored, it would stay in the system forever. From time to time, the system would present them combined with some good ideas for evaluation.

- There should be no statistical model for selecting better ideas: the evaluation should be entirely based on the user input.

Franchi comments an interesting project by Cohen, “AARON”, a Artificial Intelligence software for painters, with a very sophisticated understanding of color, contrast, hue and brightness. The outcome of the project seems absolutely pertinent:

> What is the relation of AARON's replication of a skill normally attributed to humans only – painting - and intelligence in general? One thing is clear, Cohen emphasizes: AARON's abilities do not constitute human intelligence. But this very fact makes AARON's interactions with human beings even more interesting, by challenging us to rethink our understanding of intelligence, of nature, and, ultimately, of humanity. (Franchi 1995)

After all, we may be condemned for still a long time to our condition: human, all too human, terribly human.

Critic: What’s humanity?
References


