

**The Android's Wardrobe:  
Emotions, Preferences, and Pre-Rational Decisions in Machine Consciousness**

**Paul Fahn, Ph.D. ([p.fahn@samsung.com](mailto:p.fahn@samsung.com))**

**Samsung Electronics**

**March 15, 2009**

## **Abstract**

Will emotions have any proper role in machine intelligence? Emotional systems are present in humans and other animals and seem to serve as mechanisms for primitive decision-making. As a result of evolution, however, humans are capable of logical thinking, which produces superior decisions, thus obviating the need for emotions. There would therefore appear to be no reason for AI researchers to attempt to endow androids and other “conscious machines” with emotions, although the ability for such machines to detect or predict emotions in humans may be useful. There remains one class of everyday decisions, however, where emotions may never be completely replaced by rationality: simple personal preferences, such as which shirt to buy in a store, or what food to order from a restaurant menu. We speculate that these “pre-rational” decisions, as we call them, may be implemented in machine intelligence by randomized algorithms or similar stochastic processes.

## **Introduction and Preliminaries**

What is the role of emotions in machine consciousness? Perhaps we should first ask: Do intelligent machines need emotions at all? Given all the terrible distortions that emotions bring to human judgment, behavior, and decision-making, the answer to this last question, one would hope, is No.

Today's most intelligent machines – domain-specific software such as chess-playing computers or travel reservation systems – do not use emotions at all, and there is no reason to suggest that adding emotions to such software would improve their problem-solving skill or level. Thus, at first glance, there is no need for emotions in machine intelligence. But let us examine the subject more closely.

To refine the question, we must turn our gaze from today's machines, which are really just domain-specific reasoning systems, to tomorrow's “conscious” machines, which will by definition be Artificial General Intelligent (AGI) systems operating in unstructured, real-world environments, and be capable of solving a wide range of problems, including problems on which such systems were never specifically trained.

As an exemplar of such a system, let's choose an android, i.e., a humanoid AGI acting in human environments, interacting with humans, and faced with the more or less the same tasks faced by humans. However, we do not assume that such an android must "pass" as humans, or imitate humans in every aspect of behavior; for example, an android would not drink a cup of coffee in the morning in order to raise its energy or intelligence level, the way that humans do. Our question now becomes: Does an android need to have emotions?

I also want to clarify a couple of points. First, we are not asking whether conscious machines will need to perceive or understand animal (including human) emotions; depending on the level of interaction it has with people or animals, an android may need some model of animal emotions and how to perceive or react to them. Second, emotions are not the same as states or goals. Obviously an android or any other AGI will have states and goals, and outside observers *could* call the states "emotions" by analogy with human emotional states such as anger, fear or jealousy, and *could* call the goals "emotional drives", by analogy with humans' drive towards happiness or avoidance of pain. However, these are simple anthropomorphisms, and should be avoided.

## The Role of Emotions in Humans

The neurophysiology and evolutionary purpose of emotions in mammals is only coarsely understood at this moment. Nevertheless, there are some basic points that have been generally accepted and will be useful to our investigation:

- The emotions and its supporting systems are based in the limbic system, which is a collection of neural organs.
- The limbic system is one of the oldest parts of the brain, evolutionarily speaking. In fact, it was present before mammals became a separate order; the limbic system is also present in reptiles.
- The emotions seem to have been the original decision-making mechanisms in mammals, as they still are for most mammals. For example, the emotion "fear" influences the decision whether to run away from a possible danger; this is called a "fight or flight" decision.
- Rational thinking is an alternative decision-making mechanism available to humans.
- In humans, emotion-driven decisions are often made impulsively and beyond the view of our conscious minds; for example, the decision to run when confronted by an angry tiger. By contrast, rational decision making is often done in a time-consuming manner, in full conscious awareness; for example, the decision of what move to make next in a chess game.

The development of rationality has given humans a superior decision-making tool that is not available to other species, and this capacity is largely responsible for the technological and cultural sophistication exhibited by human civilizations, at least relative to animal societies.

However, emotions still distort our decision-making progress to a considerable degree, as has been clearly demonstrated by copious research in behavioral finance. For example, in a 2005 experiment by Shiv et al. [4], patients with brain damage that impaired their emotional systems out-performed a group of healthy individuals in a task of repeated investment decisions. In another example, Ariely et al. [1] showed that merely looking at a random two-digit number (the last two digits of their social security number) influenced the amount that participants were willing to pay for a bottle of wine; this is the so-called “anchoring effect”, where exposure to arbitrary numbers or other items can influence people away from purely rational decisions.

One of the most promising aspects of machine consciousness is thus the prospect of purely rational intelligences. Without the pernicious influence of emotion, androids and other AGI’s should have a great advantage over humans, whether in playing chess, investing, formulating public policy, or running corporations. Furthermore, the speed advantage of emotional decision-making in crisis situations should cease to apply, in the vast majority of circumstances, as the speed of semiconductors continue their improvement; for example, an android will decide to run away from the tiger, using only logical reasoning, as quickly as his human (or canine) counterpart.

## Pre-Rational Decisions

So far, we have concluded that many decisions which are influenced by emotions in humans can be replaced by purely rational processes in machines. However, now consider the following examples of decisions:

- On an airplane, a stewardess asks you what you would like to drink
- In the morning, you choose which shirt to wear
- In a store, you choose which pair of shoes to buy
- You are the first to arrive for a meeting, and must decide which chair to sit in
- You choose what restaurant to go to
- At a restaurant, you choose what food to order

In each of these cases, and countless more everyday choices, we may first use rationality to narrow down the set of acceptable choices, and then use our emotional system to make a final selection. In the first example, choosing a beverage on an airplane, we may use rationality to rule out hot drinks (if we are already feeling hot), or rule out high-calorie drinks if we are on a diet, but afterwards we use our emotional systems to check which of the remaining beverages seem to please us most at that moment; if asked afterwards to justify our choice, we could not do so except by statements such as “I wanted orange juice”, “I felt like having tea”, or similar words that make reference to subjective feelings.

Common features of such decisions include

- They are under-constrained: whatever rational reasoning we may apply to limit the selection still leaves many acceptable choices. This is most easily seen in the examples of choosing food from a large restaurant menu, or choosing an article of clothing from a large store.
- We solve the problem using both rational and emotional mental processes; the final choice is made by the emotional system.
- The decisions are not completely determined by the situation or set of choices: we may choose differently when faced with an identical choice on the next day.
- The choices we make are both influenced by our “personality” and can be aspects of our “personality”; some such choices may even be said to help define our “personality”.
- The ability to make such choices is shared with other animals, who are also capable of choosing what to eat, where to sit, and so on.

We call this class of decisions “pre-rational”, because they are made using biological systems that pre-date rationality, evolutionarily speaking. These are decisions that cannot be replaced by purely logical reasoning. The role of “emotions” in machine consciousness is to make pre-rational decisions.

We note that the mental systems used to make pre-rational decisions may be identified with what Kahneman and Tversky called “System 1 Processes” [3]. Kahneman and Tversky noted that such processes are generally fast, intuitive, effortless, and “automatic”, whereas rational decisions (which they call “System 2 processes”) are slow, controlled and effortful. Their justly famous investigations revealed many cases where System 1 processes are used in situations much more suited for rationality, such as making numerical estimates, or deciding whether to accept gambles with certain odds and payoffs. Since androids will use strict rationality to make such decisions, we can confine our concerns to preference-based selections, where pure rationality is not so fruitful.

## Pre-Rational Decisions in Machines

How will androids choose which clothes to wear, or where to sit in a meeting room? These are the pre-rational decisions that will be handled by “machine emotional systems”, although this terminology is mainly indicative of the analogous way that such decisions are handled in humans and other biological creatures.

A human, when confronted with a pre-rational decision, can let his or her conscious mind “consult” with his emotional system to compare feelings based on prospective alternatives: Do I feel like drinking black tea or green tea? Do I want pasta or risotto? Should I wear my brown sweater or my blue sweater? The conscious mind is not capable of verbalizing the process in more detail than this, but the result is a sort of emotional level-check that indicates a preference for one option. Consulting one’s emotional system is thus like consulting an oracle, submitting a set of options as input and receiving levels of emotional desire or attraction as outputs.

To a machine consciousness, a pre-rational decision will approximately take the form of the following steps:

- the machine's deliberative, rational thought systems (analogous to humans' conscious train of thought) formulates the set of available choices, filtering out those that don't match explicit constraints,
- the deliberative system then calls an internal "preference oracle" with the remaining choices and relevant contextual information,
- the preference oracle returns a set of numerical "preferences" in return for each possible choice, and
- the deliberative system selects the option with the highest preference score.

The preference oracles, in our sketch of this architecture, appear as mysterious "black boxes" to the high-level deliberative systems, just as our own intuitions or emotional preferences often appear mysterious to our own conscious awareness.

How are these preference oracles to be implemented in intelligent machines? This is not clear, and the internal details will likely be part of each android manufacturer's "secret sauce", since these oracles may in a sense be the keystones to the android's personality, creating a sense of realistic individualism to the android, and influencing whether humans find the android friendly, annoying, spontaneous, fun-loving, or boring. In short, it is the architecture of the preference oracles that, when implemented successfully, will imbue the android with the ineffable sense of a living being.

We speculate that the actual internal computations of the preference oracles will be randomized algorithms or other stochastic processes, since individuals make different choices at different times when faced with the same situation: a person doesn't always choose the same dish from a menu, or wear the same color shirt every day. What sort of randomized algorithm? Randomized using what distribution? Will the distribution evolve over time (akin to personality development)? And although the randomized algorithms may be pre-programmed by the android manufacturer for commonly expected situations, how will the android behave when faced with a completely unexpected choice in an unexpected situation?

It is also possible that the preference oracle will may be deterministic, but use such opaque pattern-finding methods, incorporating an extremely wide set of contextual cues, that the actual outcome may appear random to an outside observer or even to the machine's own "conscious" thought, even if the actual hardware does not incorporate any true physical randomness.

How these preference oracles are built, whether they learn over time, whether they are deterministic or random, how they interface with the deliberative systems – these are all questions for future research, and for those planning to develop v0.9 androids, which will be the prototypes of machine consciousness.

## Acknowledgements

The author would like to thank Sheena Iyengar for useful discussions on human decision-making, and Ben Goertzel for discussions about the architecture of choice in intelligent machines.

## References

1. D. Ariely, G. Loewenstein, and D. Prelec, "Coherent Arbitrariness: stable demand curves without stable preferences", *Quarterly Journal of Economics*, 118(1), 73-105, 2003.
2. Sheena Iyengar, *How We Choose*, Twelve Publishing, 2009 (to be published).
3. Daniel Kahneman, "Maps of Bounded Rationality" (Nobel Lecture), 2002, [http://nobelprize.org/nobel\\_prizes/economics/laureates/2002/kahnemann-lecture.pdf](http://nobelprize.org/nobel_prizes/economics/laureates/2002/kahnemann-lecture.pdf).
4. B. Shiv, G. Loewenstein, A. Bechara, H. Damasio, and A. Damasio, "Investment behavior and the negative side of emotions", *Psychological Science*, 16(6): 435-439, 2005.