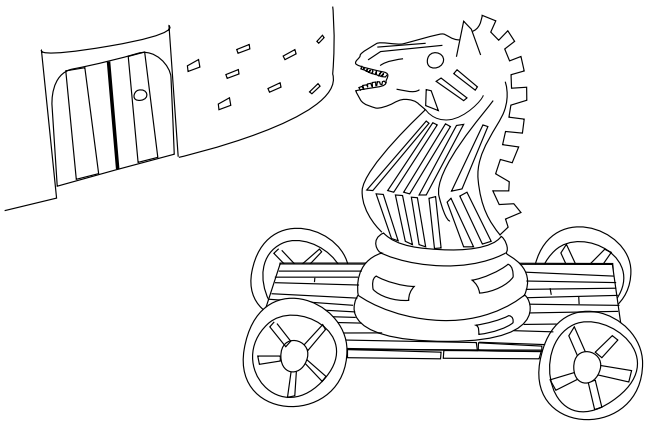
# Problem Solved: Unfriendly AI



Imagine a stage, perhaps a few years from now, in a large ballroom in a hotel in Bangkok. The reigning world champion of chess is defending his title against Deep Six, the latest generation of chess playing computers. It looks bleak for the human, as the computer has managed to establish a position which, it has predicted, will lead to checkmate in eleven moves. With its cold blue lights calmly blinking it is completely controlling the game, leading it to an inevitable, logical conclusion like a train on a track.

Meanwhile, in an adjacent dining room, a waiter spills some Grand Marnier when preparing Crepes Suzette, and the tablecloth catches fire. Flame detectors turn on a segment of the fire sprinkler system. The dining room is soaked, but so is the ballroom where the chess championships are held. Water rains down on Deep Six, shorting it out. Deep Six releases its magic smoke and all its lights go out.

Nobody predicted that.

**The purpose of intelligence is prediction**. Humans have used their minds to predict the behavior of sabertooth tigers, the right day in spring to plant the crops, and the behavior of opponents in games like chess and tennis. Predictive capabilities actually predate chess by hundreds of millions of years. An animal running on a rocky beach needs to predict where its legs will land and when to start activating the leg muscles to handle each impact. Here, the prediction reaches only milliseconds into the future. Often this is enough. The ability to better predict what an opponent will do in a fight to the death conveys the ability to parry a strike in time rather than milliseconds too late. But better than that, the first rule of street fighting is "Be Somewhere Else". Predicting that a fight is imminent is more useful than predicting the first blow. Predation requires intelligence, but more superior intelligences develop safer techniques like scavenging, ambushing, pack hunting, weapons, traps, and Trojan Horses.

An evolutionary arms race has ratcheted intelligence levels ever since the first brain-like nerve clusters appeared. Longer term predictions, predictions that are more often correct, and predictions that are more specific about the outcome all provide evolutionary advantages for superior predictors. As our ancestors started living in larger tribes, the ability to predict the behaviors of potential mates and rivals made a big difference for who got to breed. Intelligence based mate selection was one key factor in the explosive increase of intelligence levels in our species.

The brain has to process the input from the senses. This process takes some time. Benjamin Libet showed that this processing takes about half a second - the so-called

Libet Delay: we become conscious of what we see approximately half a second after it happens. The brain censors this discrepancy from our consciousness so that we can pretend we are experiencing everything as it happens, in order to preserve our sanity.

Given the Libet Delay, we could not play tennis if we could not predict where the ball will be half a second later. But that's not all. The speed of the signal in the nerves leading from the brain to the muscles controlling our fingers is roughly the same as the speed of sound, which is slow enough to matter in this context. If we are throwing a rock at a running rabbit we have to predict not only where the rabbit will be and the speed and trajectory of the rock, but we also need to account for the delay in the nerves that tell our hand when to release the rock. Doing this right means we get to dine on rabbit. Again, high precision prediction provides an evolutionary advantage. This is not restricted to higher life forms. Predictive capability selects the winners among frogs darting their tongue against flies and among flies avoiding tongues.

Having established the **purpose** of Intelligence, let's examine the **limits** on Intelligence.

The point of the chess match example is that logic based systems don't have an edge on predictions in a world that is too complex to be analyzed logically. This is true in at least a dozen different ways and I'll discuss this in detail in the next article. For now it should suffice to state that in our mundane everyday life, we rarely have all the information we need for a 100% certain prediction. But predict we must, whenever we must act. Planning is simply a prediction that includes our own actions as something to consider. We quickly and subconsciously hypothesize a few alternative actions, predict some of their possible outcomes, and then select the action that we predict will bring about the most desirable outcome. This is an error-prone process, but since it's the best we can do, this is what we do. And we do this many times every second since every muscle activation is preceded by a low-level, subconscious prediction of its effects.

We can identify three major kinds of reasoning: Deductive (which only proceeds "downhill or sideways" from given premises), Inductive (which cautiously may proceed "uphill" – from observations to general principles – under certain conditions) and Abductive (which merrily jumps to conclusions even on insufficient evidence). Our current computers almost exclusively use Deductive reasoning. It is also common in the sciences, but scientist reluctantly admit that we must often resort to Induction in order to make progress. But all humans deal with their complex everyday reality mainly by using Abduction. In everyday life, Deduction is useless. Even Induction is so rare and so spectacular that it made Sherlock Holmes a fascinating freak worth reading about.

So don't take your clues about what real or Artificial Intelligence could or should be capable of from Sherlock Holmes, Star Trek's Data, HAL-9000, or other fictional examples. Examine what real intelligences do every day. Artificial Intelligences must be able to exist alongside us in our mundane everyday life. AI is not the ability to play chess or solve integrals. Computers can do these kinds of tasks and most people agree that it's not AI. A true AI based robot should be able to go downtown, select an interesting magazine from the rack in a seven-eleven, chat with Apu the proprietor in spite of his accent, and to understand and enjoy the articles once it gets back home. A stationary AI should be able to selectively browse the web and understand what it reads.

One of the most cited definitions of AI is "The ability in a computer to do things, which, when done by a human, an observer would say required intelligence". I believe this definition is not only wrong, it is harmful to AI research and one reason we've made so little progress in the field. A better (informal) definition would be "The ability in a computer to easily do most everyday mental tasks that are easy for humans". We can effortlessly navigate a changing world, interact with other agents with goals often at odds with our own, and understand and generate spoken and written language. We can instantly recognize a chair as a chair no matter what its shape, color, or orientation. None of these tasks can be done well by today's computers, and when done at all, they are done one task or one problem domain at a time by mechanisms specific to the domain. Stanley and the other entrants in the DARPA Grand Challenge for automated cars have had some limited success driving down the road on their own but their "intelligence" is not suited for understanding and enjoying a movie like "Herbie The Love Bug".

Logic works very well when it is applicable. Deductive reasoning is 100% reliable – and induction has a good track record – in the simplified world where Science can operate. Brains are forced to use Abduction simply because it's the only thing that works at all in our complex mundane reality. AI systems have to do the same. This is not an implementation choice we can make when designing our AI; it is part of the problem statement.

Some fraction of the AI research community (or should I say, "AI enthusiast community" since this attitude is now rare among professional AI researchers) refuses to accept these ideas. They insist on trying to design logic-based infallible godlike AIs in spite of this being impossible. Some speculate about what might be possible "in principle", given a universe-sized chunk of computronium and the lifetime of the universe for its computations. They don't like the abduction based alternative simply because "It would be just as fallible as a human intelligence". To these "Shock Level Four" fanboys I say "get a clue". We need working AI as soon as possible. An AI with the intelligence of the average 14-year old human would be worth a trillion dollars since it would revolutionize everything we use computers for today and would accelerate our advance as a species more than any previous technology. It is our responsibility as transhumanists to take this opportunity and turn this misdirected Reductionist, logical reasoning based AI research around to something that will be useful in our lifetimes.

The world is constantly changing, forcing us to act within seconds, but our brains have some hard limits on computing capability. Plausible is often good enough but correct often isn't fast enough. As the joke goes, "all early hominids whose brains were based on Bayesian Logic were killed by sabertooth tigers while computing prior probabilities". We get by with our brains. AIs must get by with whatever limited resources they have. Both brains and computers have limits on cycle times, number of processors, and the amount of memory. Yes, these limits are receding quite rapidly for computers; Moore, Kurzweil, and others are likely right as far as that goes. But...

**The limits on the quality of the predictions we can make are not technological**. The complexity and unpredictability of the world yields very rapidly diminishing returns for prediction quality for any additional investment in computing power. I believe the rate

of this diminishing return is too steep to overcome for even recursive self-improvement of computers. We'll return to this issue in the next article in this series.

The insight that the complexity and unpredictability of the world enforces a limit on prediction quality – and hence intelligence – pretty much invalidates the AI Singularitarians' "Scary Idea" (as Ben Goertzel so aptly calls it) of a logic-based infallible godlike malevolent intelligence taking over the world. The decreasing return cancels out Moore's law and limits the **rate** of progress so that next year's self-improved AI wouldn't have a sufficient advantage over a dozen humans armed with pitchforks if they were also supported by a dozen of last year's AIs. The Scary Idea of a Runaway  Unfriendly AI is a red herring that we should ignore, along with ideas about logic-based AIs in general. We can now examine the alternatives in earnest and start making some progress. I (and others) have mapped out the main landmarks along this path and I'll be discussing these in future installments. Ironically, the AI singularity is impossible, and the sooner we stop trying to make it happen, the sooner we'll have workable and useful AI systems worthy of the name.

This is a saner, more moderate perspective. There are enormous gains and risks involved, but many vocal and eloquent people have (by ignoring the world-imposed limits on predictability) overestimated both. We have to design fallible Abduction based AIs because that's the only kind of true intelligence that is possible at all. And since these AIs will be fallible, we'll be able to unplug them if they develop tendencies to become "unfriendly". Problem solved.

On the flip side, don't expect a far-future logic-based infallible godlike AI to rescue the human race by solving all our problems. It's up to us, and machines a lot like us.

**- Monica Anderson**

Computer Chess: Fritz Leiber: "The 64-square Madhouse". In "A Pail Of Air", Ballantine 1964
Prediction and rabbits: William Calvin: "The Throwing Madonna". See http://wiliamcalvin.com
Variants of the cited definition of AI has been attributed to both John McCarthy and Marvin Minsky.
Benjamin Libet: "Neurophysiology of Consciousness: Selected Papers and New Essays"
More on LIbet: http://www.dichotomistic.com/mind_readings_chapter%20on%20libet.html
Masking of Libet delay in brain: Daniel Dennett: "Consciousness Explained"
http://multiverseaccordingtoben.blogspot.com/2010/10/singularity-institutes-scary-idea-and.html]
illustration by author