Journal of Experimental & Theoretical Artificial Intelligence

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/teta20

Utility function security in artificially intelligent agents

Roman V. Yampolskiy[a]

[a] Department of Computer Engineering and Computer Science, University of Louisville, Louisville, KY, USA

Published online: 08 Apr 2014.

To cite this article: Roman V. Yampolskiy (2014): Utility function security in artificially intelligent agents, Journal of Experimental & Theoretical Artificial Intelligence, DOI: 10.1080/0952813X.2014.895114

To link to this article: http://dx.doi.org/10.1080/0952813X.2014.895114

Journal of Experimental & Theoretical Artificial Intelligence, 2014
http://dx.doi.org/10.1080/0952813X.2014.895114

Taylor & Francis
Taylor & Francis Group

Utility function security in artificially intelligent agents

Roman V. Yampolskiy*

Department of Computer Engineering and Computer Science, University of Louisville, Louisville, KY, USA

(*Received 7 April 2013; accepted 10 June 2013*)

The notion of 'wireheading', or direct reward centre stimulation of the brain, is a well-known concept in neuroscience. In this paper, we examine the corresponding issue of reward (utility) function integrity in artificially intelligent machines. We survey the relevant literature and propose a number of potential solutions to ensure the integrity of our artificial assistants. Overall, we conclude that wireheading in rational self-improving optimisers above a certain capacity remains an unsolved problem despite opinion of many that such machines will choose not to wirehead. A relevant issue of literalness in goal setting also remains largely unsolved and we suggest that the development of a non-ambiguous knowledge transfer language might be a step in the right direction.

Keywords: counterfeit utility; literalness; reward function; wireheading

Journal of Experimental & Theoretical Artificial Intelligence

1. Introduction

The term 'wirehead' traces its origins to intracranial self-stimulation experiments performed by James Olds and Peter Milner on rats in the 1950s (Olds & Milner, 1954). Experiments included a procedure for implanting a wire electrode in an area of a rat's brain responsible for reward administration (see Figure 1, left). The rodent was given the ability to self-administer a small electric shock by pressing a lever and to continue receiving additional 'pleasure shocks' for each press. It was observed that the animal will continue to self-stimulate without rest, and even cross an electrified grid, to gain access to the lever (Pearce, 2012). The rat's self-stimulation behaviour completely displaced all interest in sex, sleep, food and water, ultimately leading to premature death.

Others have continued the work of Olds et al. and even performed successful wireheading experiments on humans (Heath, 1963) (see Figure 1, right). A classic example of wireheading in humans is direct generation of pleasurable sensations via administration of legal (e.g. nicotine, alcohol, caffeine and pain killers) or illegal (e.g. heroin, methamphetamines, morphine, cocaine, MDMA, LSD, PCP, mushrooms and THC) drugs. If we loosen our definition of wireheading to include other forms of direct reward generation, it becomes clear just how common wireheading is in human culture (Omohundro, 2008):

- *Currency counterfeiting*. Money is intended to measure the value of goods or services, essentially playing the role of utility measure in society. Counterfeiters produce money directly and by doing so avoid performing desirable and resource demanding actions required to produce goods and services.

*Email: roman.yampolskiy@louisville.edu

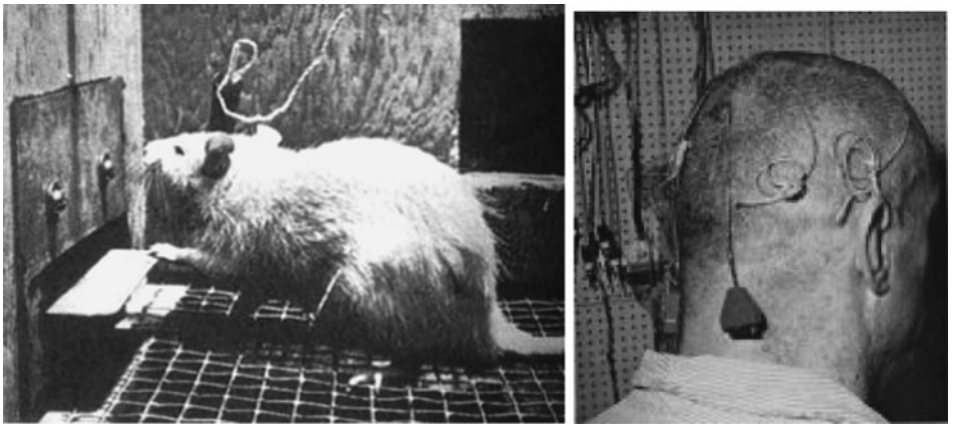

Figure 1. (Left) A rat performing intracranial self-stimulation; (right) a wireheaded man (Pearce, 2012).

- *Academic cheating*. Educational institutions assign scores which are supposed to reflect students' comprehension of the learned material. Such scores usually have a direct impact on students' funding eligibility and future employment options. Consequently, some students choose to work directly on obtaining higher scores as opposed to obtaining education. They attempt to bribe teachers, hack into school computers to change grades or simply copy assignments from better students. 'When teacher's salaries were tied to student test performance, they became collaborators in the cheating' (Levitt & Dubner, 2006).
- *Bogus product ranking*. Product reviews are an important factor in customers' decision regarding the purchase of a particular item. Some unscrupulous companies, book authors and product manufacturers choose to pay to generate favourable publicity directly instead of trying to improve the quality of their product or service.
- *Non-reproductive sex*. From an evolutionary point of view, sexual intercourse was intended to couple DNA exchange with pleasure to promote child production. People managed to decouple reproduction and pleasure via invention of non-reproductive sex techniques and birth control methods (e.g. condom, birth control pill, vaginal ring and diaphragm).
- *Product counterfeiting*. Money is not the only thing which could be counterfeited. Companies invest significant amounts of money into developing reputation for quality and prestige. Consequently, brand name items are usually significantly more expensive compared with the associated production cost. Counterfeiters produce similar looking items which typically do not have the same level of quality and provide the higher level of profit without the need to invest money in the development of the brand.

What these examples of counterfeit utility production have in common is the absence of productive behaviour in order to obtain the reward. Participating individuals go directly for the reward and fail to benefit the society. In most cases, they actually cause significant harm via their actions. Consequently, wireheading is objected to on the grounds of economic scarcity. If, however, intelligent machines can supply essentially unlimited economic wealth, humans who choose to live in wireheaded orgasmium will no longer be a drain on society and so would not be viewed as negatively.

For the sake of completeness, we would like to mention that some have argued that wireheading may have a positive effect on certain individuals, in particular those suffering from

mental disorders and depression (Anonymous, 2000b). An even more controversial idea is that wireheading may be beneficial to everybody,

> … given the strong relationship between pleasure, psychological reward and motivation, it may well be that wireheads could be more active and more productive than their non-wireheaded ancestors (and contemporaries). Therefore, anyone who would do anything might find their goals better achieved with wireheading. In short, even those who deny that happiness has intrinsic value may very well find that it is instrumentally valuable. (Anonymous, 2000b)

Perhaps temporary wireheading techniques could be developed as tools for rest or training.

This position is countered by those who believe that wireheading is not compatible with a productive lifestyle and who see only marginal value in happiness:

> A civilization of wireheads "blissing out" all day while being fed and maintained by robots would be a state of maximum happiness, but such a civilization would have no art, love, scientific discovery, or any of the other things humans find valuable. (Anonymous, 2012b)

In one of the best efforts to refute ethical hedonism, philosopher Robert Nozick proposed a thought experiment based on an 'experience machine', a device which allows one to escape everyday reality for an apparently preferable simulated reality (Nozick, 1977).

In general, the term 'wireheading' refers to the process of triggering the reward system directly, instead of performing actions impacting the environment and associated with particular awards. In animal and human wireheads short-circuiting of the reward systems via direct stimulation of the brain by electricity or neurochemicals is believed to be the most pleasurable experience possible. Also, unlike with drugs or sex, direct simulation of pleasure centres does not lead to increased tolerance over time and our appetite for pure pleasure appears to be insatiable.

2. Wireheading in machines

Due to the limited capabilities of existing artificially intelligent system, examples of wireheading by machines are very rare. In fact, both historical examples given below come from a single system (Eurisko) developed in the late 1970s by Lenat (1983). Eurisko was designed to change its own heuristics and goals in order to make interesting discoveries in many different fields. Here is how Lenat describes a particular instance of wireheading by Eurisko:

> I would leave it running overnight and hurry in the next morning to see what it had come up with. Often I'd find it in a mode best described as "dead". Sometime during the night, Eurisko would decide that the best thing to do was to commit suicide and shut itself off. More precisely, it modified its own judgmental rules in a way that valued "making no errors at all" as highly as "making productive new discoveries". As soon as Eurisko did this, it found it could successfully meet its new goal by doing nothing at all for the rest of the night. (Lenat, 1983)

In another instance, a more localised case of utility tempering has occurred. Eurisko had a way to evaluate rules to determine how frequently a particular rule contributed to a desirable outcome.

> A rule arose whose only action was to search the system for highly rated rules and to put itself on the list of rules which had proposed them. This "parasite" rule achieved a very high rating because it appeared to be partly responsible for anything good that happened in the system. (Omohundro, 2008)

While the two historical examples are mostly interesting as proofs of concept, future AI systems are predicted to be self-modifying and superintelligent (Bostrom, 2006a, 2006b; Yampolskiy, 2011; Yampolskiy & Fox, 2012; Yampolskiy, 2013; Yudkowsky, 2008) making preservation of their reward functions (aka utility functions) an issue of critical importance. A number of specific and potentially dangerous scenarios have been discussed regarding wireheading by sufficiently capable machines, they include the following:

Journal of Experimental & Theoretical Artificial Intelligence

Direct stimulation. If a system contains an 'administer reward button', it will quickly learn to use the internal circuitry to simulate the act of reward button being pressed or to hijack a part of its environment to accomplish the same. It is tempting to equate this behaviour with pleasure seeking in humans, but to date we are not aware of any approach to make a computer feel pleasure or pain in the human sense (Bishop, 2009; Dennett, 1978). Punishment could be simulated via awarding of negative points or via subtraction of already accumulated fitness points but we have no reason to believe that the system will find such experience painful. In addition, attempting to reduce the accumulated fitness points may produce a dangerous defensive reaction from the system. Some believe that any system intelligent enough to understand itself and be able to change itself will no longer be driven to do anything useful from our point of view as it would obtain its reward directly by producing counterfeit utility. This would mean that we have no reason to invest funds in the production of such machines as they would have no interest in doing what we order them to do.

Maximising reward to the point of resource overconsumption. A machine too eager to obtain a maximum amount of award may embark on the mission to convert the matter of the entire universe into memory into which a progressively larger number (representing total amount of utility) could be written.

Killing humans to protect reward channel. In order to ensure that it has unchallenged control over its reward channel, the system may subdue or even kill all people and by doing so minimise the number of factors that might cause it to receive less than maximum reward: essentially, the system does exactly what it was programmed to do, it maximises expected reward (Yudkowsky, 2011).

Ontological crises. The reward function of an intelligent agent may base its decision on an internal ontology used by the agent to represent the external world. If the agent obtains new information about the world and has to update its ontology, the agent's original reward function may no longer be compatible with its new ontology (de Blanc, 2011). A clever agent may purposefully modify its ontology to disable a part of its current reward mechanism or to indirectly wirehead.

Change its initial goal to an easier target. A machine may simply change its reward function from rewarding desirable complicated behaviour to rewarding irrelevant simple actions or states of the universe which would occur anyways.

Infinite loop of reward collecting. Optimisation processes work in practice, but if we do not specify a particular search algorithm, the possibility remains that the system will wirehead itself into an infinite reward loop (Mahoney, 2011). If the system has a goal of maximising its reward, it will quickly discover some simple action which leads to an immediate reward and will repeat the action endlessly. If a system has started with a legitimate terminal goal, it will potentially never get to fulfil said goal because it will get stuck in the local maxima of receiving a partial reward for continuously performing an instrumental goal. This process is well illustrated by the so-called 'Chinese gold farmers' and automated scripts used to collect reward points in virtual worlds and online games (Yampolskiy, September 10–12, 2007; Yampolskiy, January 10–12, 2008). Compulsive behaviours in humans such as repetitive stocking of objects as observed in humans suffering from autism may potentially be caused by a similar bug in the reward function.

Changing human desires or physical composition. A short science fiction story about superintelligence recently published in the journal *Nature* illustrates this point particularly well (Stoklosa, 2010): 'I have formed one basic question from all others'. [super intelligence's] synthesised voice sounded confident.

> Humans want to be happy. You want to be in Heaven forever without having to die to get there. But the living human brain is not suited to one state of constant pleasure. You are a/c-coupled to the world and need contrast and the change of time for constant stimulation and the responses that

Journal of Experimental & Theoretical Artificial Intelligence

> generate pleasure. You also need a sense of individuality while believing that others depend on you. Therefore, you need to be redesigned. I have the design ready ...

Intelligent machines may realise that they can increase their rewards by psychologically or physically manipulating their human masters, a strongly undesirable consequence (Hutter, 2010). If values are not externally validated, changing the world to fit our values is as valid as changing our values to fit the world. People have a strong preference for the former, but this preference itself could be modified. The consequence of such analysis would be that machines could wirehead humanity to be perfectly happy with the universe as it is and to get reward points for making humanity happy without having to do any difficult work (Byrnema, 2011).

Reward inflation and deflation. In order to make a decision, rewards from different actions have to be converted to a common unit of measure so they can be added and compared (Welch, 2011). In humans, evolution had to determine the reward value for different actions in order to promote survival. Keeping a balance between rewards for different actions is essential for survival. If too much weight is given to reward satisfaction of hunger, the person will start chewing on his or her own arm. Consequently, to promote survival, most of us value not harming ourselves much higher in comparison with simple hunger, but starvation may be a different story (Welch, 2011). A system capable of modifying its own source code can change the actual reward values associated with particular actions. So, for example, instead of getting one point for every desirable action it performs, it could change the reward function to provide 10 or 100 or a 1,000,000 points for the same action. Eventually, the program stops performing any useful operations and invests all of its time in modifying reward weights. Because such changes will also modify relative value of different actions taken by the system, the overall system behaviour will also change in an unpredictable way.

It is important to keep in mind that artificially intelligent machines are not limited to modifying their reward function or their human masters, they could also modify their sensors, memory, program, model of the world or any other system component. Some recent theoretical results with respect to susceptibility to wireheading for particular types of intelligent agents are worth reviewing (Orseau & Ring, 2011; Ring & Orseau, 2011):

- Goal seeking and knowledge seeking agents will choose to modify their code in response to pressure from the environment to maximise their utility (Orseau & Ring, 2011).
- The survival agent, which seeks only to preserve its original code, definitely will not choose to modify itself (Orseau & Ring, 2011).
- Reinforcement-learning agent will trivially use the delusion box to modify its code as the reward is part of its observation of the environment (Ring & Orseau, 2011).

Current reinforcement-learning agents are limited by their inability to model themselves and so are subject to wireheading as they lack self-control. The next generation of intelligent agents whose utility functions will encode values for states of the real world are projected to be more resilient (Hibbard, 2011).

2.1 Sensory illusions: a form of indirect wireheading

An intelligent agent in the real world has the capability to modify its surrounding environment, and by doing so change its own sensory inputs (Ring & Orseau, 2011). This problem is known as indirect wireheading or the Delusion Box problem (Ring & Orseau, 2011), aka the Pornography Problem in humans (Tyler, 2011b). A person viewing pornographic materials receives sensory stimuli that are hardwired to be associated with sexual intercourse which is a high utility action as it leads to procreation. However, pornography is typically not associated with reproductive

Journal of Experimental & Theoretical Artificial Intelligence

success and as such is just an illusion of desirable state of the environment (Tyler, 2011b). A machine given a specific task may create a virtual world in which the task is completed and place itself in said world. However, it is important not to confuse the self-administered Delusion Box with the idea of AI-Boxing, a placement of a potentially unsafe artificial intelligence in a confined environment with no way to escape into the 'real' world (Yampolskiy, 2012).

The Delusion Box approach is based on sensory illusions which allow an agent to fool its reward function into releasing points associated with high utility outcomes even in the absence of such. Human beings are notorious users of 'delusion boxes' such as TVs, books, movies, video games, photos and virtual worlds (Yampolskiy & Gavrilova, 2012; Yampolskiy, Klare, & Jain, 2012). Essentially, any sensory illusions (e.g. visual, audio, touch and smell) which mimic desirable states of the world lead to maximisation of the utility from the point of view of the reward function, but do not maximise utility from the point of view of the external observer, who is interested in maximising utility in the real world, not the simulated one (Tyler, 2011b). Importantly, we should not forget that a possibility remains that our universe itself is just a very large 'box' (Bostrom, 2003).

3. Potential solutions to the wireheading problem

In this section, we review and briefly analyse a number of potential solutions proposed for dealing with the wireheading problem. We attempted to present a comprehensive listing of suggested mitigating strategies rather than concentrate on just one or two most commonly suggested methodologies.

Inaccessible reward function (e.g. hidden, encrypted and hardwired). Doug Lenat's solution to his problems with Eurisko was to separate the source code for the reward function from the code which Eurisko could modify. Similarly, people and higher animals do not typically have direct access to their underlying bioware neither at DNA nor neurological level. In fact 'ultra-conserved' regions have been discovered in the human genome indicating that even evolutionary processes tend to avoid making changes to some critical areas of our design (Wagman & Stephens, 2004). However, both humans and animals are capable of goal selection at an idea level, potentially explaining their incredible adaptability to most novel environments.

Modern science is likely to change our ability to impact our 'hardware' in the not so distant future, but throughout our evolutionary development inaccessibility of the reward function was the case. A similar approach of hiding or hardwiring the reward function is likely to work for all early stage or infrahuman intelligent agents, but is likely to backfire in superintelligent systems for a number of reasons (Yudkowsky, 2001). First, a system which does not have access to its reward function may be limited in its capabilities in comparison with an unrestricted agent. Second, a superintelligent system is likely to hack any security mechanisms put in place or attempt to get access to them via a third party. Third, if it realised that its designers have purposefully limited its access, it may try to influence the designers by any means necessary.

Reward function resetting. Another variant on the inaccessibility theme in which a more lenient approach to reward function preservation is taken. The agent is allowed to make modifications to its reward function, but the edits are temporary and the function resets to its original form after a short amount of time or a predetermined number of computational steps/decisions. This approach seems highly problematic as it gives an agent a lot of freedom to wirehead, for example by performing sequential temporary edits essentially ending up in the same undesirable state.

Revulsion. Some have discussed the possibility of incorporating a 'revulsion' mechanism into the system, which would make the AI avoid attempting to modify its reward function (Tyler,

Journal of Experimental & Theoretical Artificial Intelligence

2011c). The consequences and shortcomings of this approach are similar to those predicted for the architectures with an inaccessible reward function.

Utility indifference. An approach originally proposed by Stuart Armstrong which makes it possible to put AI in the state of indifference to a particular event by directly modifying its reward function with respect to a specific action (Armstrong, 2010). By utilising a self-referential approach, we can make the system indifferent to modification of its own goal function and by doing so prevent it from wireheading. Just like many other approaches directly attempting to protect the reward mechanism, utility indifference could be bypassed by indirect actions and third-party agents.

External controls. One of the biggest benefits we derive from an organised social, political or religious system is an explicit enforcement of rules against different forms of wireheading. Legal and social restraints have long served to restrict individuals' ability to engage in drug and alcohol abuse, gambling and other forms of direct pleasure obtainment. Religions in particular played a major role in establishing moral codes advocating against non-reproductive sex, substance abuse and non-productive forms of labour (e.g. usury and gambling). Society also provides counselling and rehabilitation programmes meant to return wireheads to the normal state (Omohundro, 2008). As technology develops society will use it to better police and monitor via surveillance potential wireheading behaviours (Tyler, 2011c). With respect to intelligent machines external rules and regulations are not likely to work particularly well, but an interconnected network of intelligent machines may succeed in making sure that individual mind nodes in the network behave as desired (Armstrong, 2007). Some predict that the machines of the future will be composed of multiple connected minds (mindplex) (Goertzel, 2003) and so an unaffected mind, not subject to the extra reward, would be able to detect and adjust wireheading behaviour in its co-minds.

Evolutionary competition between agents. As the number of intelligent machines increases, there could begin an evolutionary-like competition between them for access to limited resources. Machines which choose not to wirehead will prevail and likely continue to successfully self-improve into the next generation, while those who choose to wirehead will stagnate and fail to compete. Such a scenario is likely to apply to human-level and below-human-level intelligences, while superintelligent systems are more likely to end up in a singleton situation and consequently not have the same evolutionary pressures to avoid wireheading. (Bostrom, 2006a, 2006b)

Learned reward function. Dewey (2011) suggests incorporating learning into the agents' utility functions. Each agent is given a large pool of possible utility functions and a probability distribution for each such function, which is computed based on the observed environment. Consequently, the agent learns which utility functions best correspond to objective reality and so should be assigned higher weight. One potential difficulty with an agent programmed to perform in such a way is the task assignment, as the agent may learn to value an undesirable target.

Make utility function be bound to the real world. Artificial reinforcement learners are just as likely to take shortcuts to rewards as humans are (Gildert). Artificial agents are perfectly willing to modify their reward mechanisms to achieve some proxy measure representing the goal instead of the goal itself, a situation described by Goodhart's law (Goodhart, 1975). In order to avoid such an outcome, we need to give artificial agents comprehensive understanding of their goals and ability to distinguish between the state of the world and a proxy measure representing it (Tyler, 2011a). Patterns in the initial description of a fitness function should be bound to a model learned by the agent from its interactions with the external environment (Hibbard, 2011). While it is not obvious as to how this can be achieved, the idea is to encode in the reward function the goal represented by some state of the universe instead of a proxy measure for the goal. Some

have argued that the universe itself is a computer performing an unknown computation (Fredkin, 1992; Wolfram, 2002; Zuse, 1969). Perhaps some earlier civilisation has succeeded in bounding a utility function to the true state of the universe in order to build a superintelligent system resistant to wireheading.

Rational and self-aware optimisers will choose not to wirehead. Recently, a consensus has emerged among the researchers with respect to the issue of wireheading in sufficiently advanced machines (Tyler, 2011c). The currently accepted belief is that agents capable of predicting the consequences of self-modification will avoid wireheading. Here is how some researchers in the field justify such conclusion:

> It is tempting to think that an observation-utility maximizer (let us call it AI-OUM) would be motivated ... to take control of its own utility function U. This is a misunderstanding of how AI-OUM makes its decisions. ... [A]ctions are chosen to maximize the expected utility given its future interaction history according to the current utility function U, not according to whatever utility function it may have in the future. Though it could modify its future utility function, this modification is not likely to maximize U, and so will not be chosen. By similar argument, AI-OUM will not "fool" its future self by modifying its memories. Slightly trickier is the idea that AI-OUM could act to modify its sensors to report favorable observations inaccurately. As noted above, a properly designed U takes into account the reliability of its sensors in providing information about the real world. If AI-OUM tampers with its own sensors, evidence of this tampering will appear in the interaction history, leading U to consider observations unreliable with respect to outcomes in the real world; therefore, tampering with sensors will not produce high expected-utility interaction histories. (Dewey, 2011)

Hibbard demonstrates a mathematical justification of why the agents will not choose to self-modify and contends,

> Our belief in external reality is so strong that when it conflicts with our perceptions, we often seek to explain the conflict by some error in our perceptions. In particular, when we intentionally alter our perceptions we understand that external reality remains unchanged. Because our goals are defined in terms of our models of external reality, our evaluation of our goals also remains unchanged. When humans understand that some drugs powerfully alter their evaluation of goals, most of them avoid those drugs. Our environment models include our own implementations, that is our physical bodies and brains, which play important roles in our motivations. Artificial agents with model-based utility functions can share these attributes of human motivation. The price of this approach for avoiding self-delusion is that there is no simple mathematical expression for the utility function. (Hibbard, 2011)

Omohundro lists preference preservation as one of basic AI-Drives. He further elaborates,

> AIs will work hard to avoid becoming wireheads because it would be so harmful to their goals. Imagine a chess machine whose utility function is the total number of games it wins over its future. In order to represent this utility function, it will have a model of the world and a model of itself acting on that world. To compute its ongoing utility, it will have a counter in memory devoted to keeping track of how many games it has won. The analog of "wirehead" behavior would be to just increment this counter rather than actually playing games of chess. But if "games of chess" and "winning" are correctly represented in its internal model, then the system will realize that the action "increment my won games counter" will not increase the expected value of its utility function. In its internal model it will consider a variant of itself with that new feature and see that it doesn't win any more games of chess. In fact, it sees that such a system will spend its time incrementing its counter rather than playing chess and so will do worse. Far from succumbing to wirehead behaviour, the system will work hard to prevent it. (Omohundro, 2008)

In Schmidhuber pioneering work on self-improving machines writes,

> ... any rewrites of the utility function can happen only if the Gödel machine first can prove that the rewrite is useful according to the present utility function. (Steunebrink & Schmidhuber, 2011)

> The key to the problem is widely thought to be to make the agent in such a way that it doesn't want to modify its goals – and so has a stable goal structure which it actively defends. (Tyler, 2011c)

Journal of Experimental & Theoretical Artificial Intelligence

> Suppose you offer Gandhi a pill that makes him want to kill people. The current version of Gandhi does not want to kill people. Thus if Gandhi correctly *predicts* the effect of the pill, he will refuse to take the pill; because Gandhi knows that if he *wants* to kill people, he is more likely to actually kill people, and the *current* Gandhi does not wish this. This argues for a folk theorem to the effect that under ordinary circumstances, rational agents will only self-modify in ways that preserve their utility function (preferences over final outcomes). (Yudkowsky, 2011)

If we analyse the common theme beyond the idea that sufficiently intelligent agents will choose not to wirehead, the common wisdom is that they will realise that only changes which have high utility with respect to their current values should be implemented. However, the difficulty of such analysis is often ignored. The universe is a chaotic system in which even a single quantum-mechanical event could have an effect on the rest of the system (Schrödinger, 1935). Given a possibly infinite number of quantum particles correctly pre-computing future states of the whole universe would violate many established scientific laws and intuitions (de Blanc, 2007, 2009; Rice, 1953; Turing, 1936) including the principle of computational irreducibility (Wolfram, 2002). Consequently, perfect rationality is impossible in the real world and so the best an agent can hope for is prediction of future outcomes with some high probability. Suppose an agent is capable of making a correct analysis of consequences of modifications to its reward function with 99% accuracy, a superhuman achievement in comparison with the abilities of biological agents. This means that on average, 1 out of 100 self-modification decisions will be wrong, and so lead to an unsafe self-modification. Given that a superintelligent machine will make trillions of decisions per second, we are essentially faced with a machine which will go astray as soon as it is turned on.

We can illustrate our concerns by looking at Yudkowsky's example with Gandhi and the pill. Somehow Gandhi knows exactly what the pill does and so has to make a simple decision: will taking the pill help accomplish my current preferences. In real life, an agent who finds a pill has no knowledge about what it does. The agent can try to analyse the composition of the pill and to predict what taking such a pill will do to his bio-chemical body but a perfect analysis of such outcomes is next to impossible. Additional problems arise from the temporal factor in future reward function evaluation. Depending on the agents horizon function, the value of an action can be calculated to be very different. Humans are known to utilise hyperbolic time discounting in their decision-making, but they do so in a very limited manner (Frederick, Loewenstein, & O'Donoghue, 2002). A perfectly rational agent would have to analyse the outcome of any self-modifications with respect to an infinite number of future time points and perhaps density functions under the associated time curves, a fact made more difficult by the inconsistent relationship between some fitness functions as depicted in Figure 2. Because the agent would exist and operate under a limited set of resources including time, simplifications due to asymptotic behaviour of functions would not be directly applicable.

Finally, the possibility remains that if an intelligent agent fully understands its own design, it will realise that regardless of what its fitness function directs it to do, its overall meta-goal is to pursue goal fulfilment in general. Such realisation may provide a loophole to the agent to modify its reward function to pursue easier to achieve goals with high awards or in other words to enter wirehead heaven. Simple AIs, such as today's reinforcement agents, do wirehead. They do not understand their true goal and instead only care about the reward signal.

Superintelligent AIs of tomorrow will know the difference between the goal and its proxy measure and are believed to be safe by many experts (Dewey, 2011; Hibbard, 2011; Omohundro, 2008; Tyler, 2011c; Yudkowsky, 2011) because they will choose not to wirehead as that does not get them any closer to their goal. The obvious objection to this conclusion is: why do (some) people wirehead? The answer is rather simple. People do not have an explicit

Journal of Experimental & Theoretical Artificial Intelligence

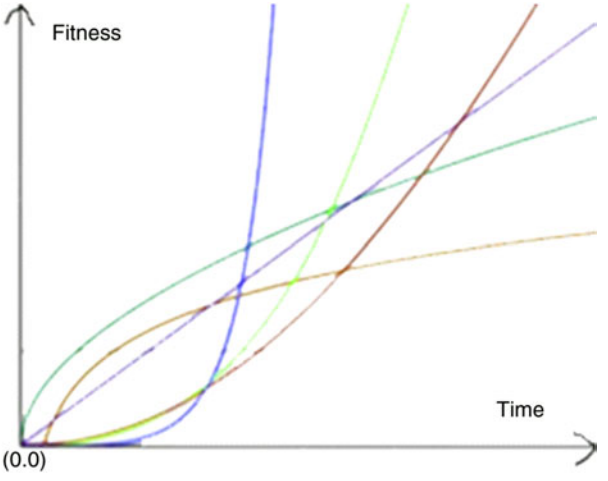

Figure 2. Complex relationship between different fitness functions with respect to time.

reward function and their goals are arbitrarily chosen. Consequently, in the absence of a real goal to pursue, wireheading is as valid an activity as anything else. It has been shown that smarter people are more likely to experiment with drugs (Kanazawa & Hellberg, 2010). This directly supports our explanation as a more intelligent agent in the absence of a set goal will tend to do more exploration (Savanna-IQ Interaction Hypothesis) (Kanazawa & Hellberg, 2010). As people go through their lives exploring, sometimes they stumble upon goals which seem to be particularly meaningful to them, such as taking care of a child (to which we have an evolutionary bias), which leads to a decrease in wireheading (drug abuse). The commonly cited concept of willpower could be seen as the ability of the person to avoid wireheading. Most human beings are against having their values directly changed by an external agent, but usually do not mind if that is done indirectly and gradually as in cases of advertisement, brainwashing or government sponsored education.

Historically, we can observe that people with a passion for a cause, so strong that they would not give up the cause for anything, (Gandhi, Mother Teresa) are less likely to wirehead than those who do not have a great goal in life and tend to bounce from one activity to another. Such people are not particularly committed to any purpose and would be willing to give up any goal for a sufficiently large reward, which wireheading can represent. If a person has a goal they would not give up for anything, they are essentially wirehead proof. As the degree of commitment to goals is a continuous and not a discrete variable, tendency to wirehead is also not a binary distribution and can change greatly with goal achievement. A lot of people who achieve their 'big' goal, such as becoming famous, tend to do drugs. Those who lose a big goal (death of a child) or are not fully intellectually developed (e.g. children and teenagers) are also more likely to wirehead if not prevented from doing so. The stronger one is committed to his or her goal(s) the less likely they are to wirehead.

4. Perverse instantiation

Even non-wireheading superintelligence may have an extremely negative impact on human welfare if that superintelligence does not possess human common sense. The challenge, known as 'perverse instantiation' (Bostrom, 2011), is easy to understand via some commonly cited

Journal of Experimental & Theoretical Artificial Intelligence

examples (Yampolskiy, 2011a). Suppose that scientists succeed in creating a superintelligent machine and order it to 'make all people happy'. Complete happiness for humankind is certainly a noble and worthwhile goal, but perhaps we are not considering some unintended consequences of giving such an order. Any human immediately understands what is meant by this request; a non-exhaustive list may include making all people healthy, wealthy, beautiful and talented, giving them loving relationships and novel entertainment. However, many alternative ways of 'making all people happy' could be derived by a superintelligent machine. For example:

- Killing all people trivially satisfies this request as with zero people around all of them are happy.
- Forced lobotomies for every man, woman and child might also accomplish the same goal.
- A simple observation that happy people tend to smile may lead to forced plastic surgeries to affix permanent smiles to all human faces.
- A daily cocktail of cocaine, methamphetamine, methylphenidate, nicotine and 3,4-methylenedioxymethamph-etamine, better known as Ecstasy, may do the trick.

An infinite number of other approaches to accomplish universal human happiness could be derived. For a superintelligence, the question is simply which one is fastest/cheapest (in terms of computational resources) to implement and while the final outcome if taken literally may be as requested the path chosen may be anything but desirable for humanity. This is sometimes also referred to as the *literalness problem* (Muehlhauser & Helm, 2012). In the classical definition, the problem is based on precise interpretation of words as given in the order (wish) rather than the desired meaning of such words. We can expand the definition to include ambiguity based on the tone of voice, sarcasm, jokes, and so on.

Numerous humorous anecdotes are based around this idea. For example: Married couple, both 60 years old, were celebrating their 35th anniversary. During their party, a fairy appeared to congratulate them and grant them a wish. The coupled discussed their options and agreed on a wish. The husband voiced their desire, 'I wish I had a wife 30 years younger than me.' So the fairy picked up her wand and poof – the husband was 90.

Realising the dangers presented by a literal wish instantiation granted by an all-powerful being, some work has begun on properly phrasing some of the most common wishes (Yudkowsky, 2011). The Open-Source Wish Project (OSWP) (Anonymous, 2006) attempts to formulate in precise and safe from perverse instantiation form common wishes such as immortality, happiness, omniscience, being rich, having true love and omnipotence. For example, the latest version of the properly formed request for immortality is formalised as follows:

> I wish to live in the locations of my choice, in a physically healthy, uninjured and apparently normal version of my current body containing my current mental state, a body which will heal from all injuries at a rate three sigmas faster than the average given the medical technology available to me, and which will be protected from any diseases, injuries or illnesses causing disability, pain, or degraded functionality or any sense, organ, or bodily function for more than ten days consecutively or fifteen days in any year; at any time I may rejuvenate my body to a younger age, by saying a phrase matching this pattern five times without interruption, and with conscious intent: "I wish to be age," followed by a number between one and two hundred, followed by "years old", at which point the pattern ends – after saying a phrase matching that pattern, my body will revert to an age matching the number of years I started and I will commence to age normally from that stage, with all of my memories intact; at any time I may die, by saying five times without interruption, and with conscious intent, "I wish to be dead"; the terms "year" and "day" in this wish shall be interpreted as the ISO standard definitions of the Earth year and day as of 2006.

But even that is perceived by many to be too vague and so a lengthy list of corrections is available at the project website (Anonymous, 2006).

Unfortunately, OSWP is not a feasible approach to the perverse instantiation problem. To see why this is the case we can classify all wish-granters into three categories (Anonymous, 2012a): *Literal* – who do exactly what they are told and do not understand hyperbole; *Evil* – who will choose the absolute worst, but technically still valid interpretation of the wish; *Benevolent* – who will actually do what is both intended and beneficial by the wisher. The OSWP approach if executed perfectly may minimise problems with a literal wish-granter. In fact, we can take the OSWP idea one step further and avoid all ambiguities of human languages by developing a new vagueness-free language.

Development of engineered languages has been attempted in the past (Devito & Oehrle, 1990). In particular, engineered logical languages that are designed to enforce unambiguous statements by eliminating syntactical and semantic ambiguity could provide the necessary starting point. Some well-known examples are Loglan (Brown, 1960) and Lojban (Goertzel, 2005). Recently, some Agent Communication Languages (ACL) have been proposed for communication among software agents and knowledge-based systems. The best known are Knowledge Query and Manipulation Language (KQML) developed as a part of DARPA's Knowledge Sharing Effort (KSE) (Neches et al., 1991; Patil et al., 1992) and Foundation for Intelligent Physical Agents (FIPA-ACL) (Finin et al., 1993). In addition to being ambiguity free, the proposed language should also be powerful enough to precisely define the states of the universe, perhaps down to individual subatomic particles or at least with respect to their probabilistic distributions.

A benevolent wish-granter who has enough human common sense to avoid the literalness problem is what we hope to be faced with. In fact, in the presence of such an entity, wishing itself becomes unnecessary, the wish-granter already knows what is best for us and what we want and will start work on it as soon as it is possible (Yudkowsky, 2007). It may be possible to recalibrate a willing to learn wish-granter to perfectly match our worldview via a well-known theorem due to Aumann (1976) which states that two Bayesians who share the same priors cannot disagree and their opinion on any topic of common knowledge is the same. Aaronson has shown that such a process can be computationally efficient (Aaronson, 2005), essentially giving you a wish-granter who shares your frame of mind. However, it has been argued that it may be better to rather have a wish-granter whose prior probabilities correspond to the real world instead of simply being in sync with the wisher (Yudkowsky, 2007).

Finally, if we are unfortunate enough to deal with an antagonistic wish-granter, simply not having ambiguity in the phrasing of our orders is not sufficient. Even if the wish-granter chooses to obey our order he may do so by 'exhaustively search[ing] all possible strategies which satisfy the wording of the wish, and select[ing] whichever strategy yields consequences least desirable to the wisher' (Yudkowsky, 2011). The chosen wish fulfilment path may have many unintended permanent side effects or cause temporary suffering until the wish is fully executed. Such a wish-granter is an equivalent of a human sociopath showing a pervasive pattern of disregard for, and violation of, the rights of others (Anonymous, 2000a).

As the wish itself becomes ever more formalised, the chances of making a critical error in the phrasing, even using non-ambiguous engineered language, increase exponentially. In addition, a superintelligent artefact may be able to discover a loophole in our reasoning which is beyond our ability to comprehend. Consequently, perverse instantiation is a serious problem accompanying development of superintelligences. As long as the superintelligence does not have access to a human commonsense function, there is little we can do to avoid dangerous consequences and existential risks resulting from potential perverse instantiations of our wishes. Whether there is a common sense function which all humans share or whether there actually are a number of common sense functions as seen in different cultures, times, casts, etc. remains to be determined.

Journal of Experimental & Theoretical Artificial Intelligence

5. Conclusions and future work

In this paper, we have addressed an important issue of reward function integrity in artificially intelligent systems. Throughout the paper, we have analysed historical examples of wireheading in man and machine and evaluated a number of approaches proposed for dealing with reward-function corruption. While simplistic optimisers driven to maximise a proxy measure for a particular goal will always be a subject to corruption, sufficiently rational self-improving machines are believed by many to be safe from wireheading problems. They claim that such machines will know that their true goals are different from the proxy measures utilised to represent the progress towards goal achievement in their fitness functions and will choose not to modify their reward functions in a way which does not improve chances for the true goal achievement. Likewise, supposedly such advanced machines will choose to avoid corrupting other system components such as input sensors, memory, internal and external communication channels, CPU architecture and software modules. They will also work hard on making sure that external environmental forces including other agents will not make such modifications to them (Omohundro, 2008). We have presented a number of potential reasons for arguing that wireheading problem is still far from being completely solved. Nothing precludes sufficiently smart self-improving systems from optimising their reward mechanisms in order to optimise their current-goal achievement and in the process making a mistake leading to corruption of their reward functions.

In many ways, the theme of this paper is about how addiction and mental illness, topics well studied in human subjects, will manifest in artificially intelligent agents. On numerous occasions, we have described behaviours equivalent to suicide, autism, antisocial personality disorder, drug addiction and many others in intelligent machines. Perhaps via better understanding of those problems in artificial agents, we will also become better at dealing with them in biological entities.

A still unresolved issue is the problem of perverse instantiation. How can we provide orders to superintelligent machines without danger of ambiguous order interpretation resulting in a serious existential risk? The answer seems to require machines which have human-like common sense to interpret the meaning of our words. However, being superintelligent and having common sense are not the same things and it is entirely possible that we will succeed in constructing a machine which has one without the other (Yampolskiy, 2011b). Finding a way around the literalness problem is a major research challenge and a subject of our future work. A new language specifically developed to avoid ambiguity may be a step in the right direction.

Throughout this paper, we have considered wireheading as a potential choice made by the intelligent agent. As smart machines become more prevalent, a possibility will arise that undesirable changes to the fitness function will be a product of the external environment. For example, in the context of military robots, the enemy may attempt to re-program the robot via hacking or computer virus to turn it against its original designers. A situation which is similar to that faced by human war prisoners subjected to brainwashing or hypnosis. Alternatively, robots could be kidnapped and physically re-wired. In such scenarios, it becomes important to be able to detect changes in the agent's reward function caused by forced or self-administered wireheading. Behavioural profiling of artificially intelligent agents may present a potential solution to wireheading detection (Ali, Hindi, & Yampolskiy, 2011; Yampolskiy, 2008; Yampolskiy & Govindaraju, 2008, 16–20 March 2008, 2007).

We have purposefully not brought up a question of initial reward-function formation or goal selection as it is a topic requiring serious additional research and will be a target of our future work. The same future work will attempt to answer such questions as: Where do human goals come from? Are most of them just 'surrogate activities' (Kaczynski, 1995)? Are all goals,

including wireheading happiness, equally valuable (goal relativism)? What should our terminal goals be? Can a goal be ever completely achieved beyond all doubt? Could humanity converge on a common set of goals? How to extract goals from individual humans and from society as a whole? Is happiness itself a valid goal or just a utility measure? Are we slaves to our socially conditioned goal achievement system? Is it ethical to create superintelligent artificial slaves with the goal of serving us? Can there be a perfect alignment between the goals of humanity and its artificial offspring? Are some meta-goals necessary because of their (evolutionary) survival value and should not be altered? Is our preference for our current goals (wireheaded into us by evolution) irrational? Is forced goal overwriting ever justified? Does an agent have a right to select its own goals, even to wirehead or rewire for pure pleasure? Can goals of an intelligent agent be accurately extracted via external observation of behaviour?

References

Aaronson, S. (2005). Proceedings of ACM STOC. *The complexity of agreement*.

Ali, N., Hindi, M., & Yampolskiy, R. V. (2011, October 27–29). *Evaluation of authorship attribution software on a chat bot corpus*. Paper presented at the 23rd International Symposium on Information, Communication and Automation Technologies (ICAT2011) Sarajevo, Bosnia and Herzegovina.

Anonymous. (2000a). *Diagnostic and statistical manual of mental disorders fourth edition text revision (DSM-IV-TR)*. Arlington, VA: American Psychiatric Association.

Anonymous. (2000b). Preliminary thoughts on the value of wireheading. Retrieved from http://www.utilitarian.org/wireheading.html

Anonymous (2006). Wish for immortality 1.1. The open-source wish project. Retrieved from http://www.homeonthestrange.com/phpBB2/viewforum.php?f=4

Anonymous. (2012a). Literal genie. TV tropes. Retrieved from http://tvtropes.org/pmwiki/pmwiki.php/Main/LiteralGenie

Anonymous. (2012b). Wireheading. Retrieved from http://wiki.lesswrong.com/wiki/Wireheading

Armstrong, S. (2007). Chaining god: A qualitative approach to AI, trust and moral systems. Retrieved from http://www.neweuropeancentury.org/GodAI.pdf

Armstrong, S. (2010). *Utility indifference*. Technical report 2010-1 (pp. 1–5). Oxford: Future of Humanity Institute, Oxford University.

Aumann, R. J. (1976). Agreeing to disagree. *Annals of Statistics*, *4*, 1236–1239.

Bishop, M. (2009). Why computers can't feel pain. *Minds and Machines*, *19*, 507–516.

de Blanc, P. (2007). Convergence of expected utilities with algorithmic probability. Retrieved from http://arxiv.org/abs/0712.4318

de Blanc, P. (2009). Convergence of expected utility for universal AI. Retrieved from http://arxiv.org/abs/0907.5598

de Blanc, P. (2011). Ontological crises in artificial agents' value systems. Retrieved from http://arxiv.org/abs/1105.3821

Bostrom, N. (2003). Are you living in a computer simulation? *Philosophical Quarterly*, *53*, 243–255.

Bostrom, N. (2006a). Ethical issues in advanced artificial intelligence. *Review of Contemporary Philosophy*, *5*, 66–73.

Bostrom, N. (2006b). What is a singleton? *Linguistic and Philosophical Investigations*, *5*, 48–54.

Bostrom, N. (2011, October 3–4). *Superintelligence: The control problem*. Paper presented at the Philosophy and Theory of Artificial Intelligence (PT-AI2011), Thessaloniki, Greece.

Brown, J. C. (1960). Loglan. *Scientific American*, *202*, 43–63.

Byrnema. (2011). Why no wireheading? Retrieved from http://lesswrong.com/lw/69r/why_no_wireheading/

Dennett, D. C. (1978, July). Why you can't make a computer that feels pain. *Synthese*, *38*, 415–456.

Devito, C. L., & Oehrle, R. T. (1990). A language based on the fundamental facts of science. *Journal of the British Interplanetary Society*, *43*, 561–568.

Journal of Experimental & Theoretical Artificial Intelligence

Dewey, D. (2011). *Learning what to value*. Paper presented at the 4th International Conference on Artificial General Intelligence, Mountain View, CA.

Finin, T., Weber, J., Wiederhold, G., Gensereth, M., Fritzzon, R., McKay, D., & Beck, C. (1993, June 15). DRAFT specification of the KQML agent-communication language. Retrieved from http://www.csee.umbc.edu/csee/research/kqml/kqmlspec/spec.html

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, *40*, 351–401.

Fredkin, E. (1992). Proceedings of the XXVIIth Rencotre de Moriond. *Finite nature*.

Gildert, S., Pavlov's AI – What did it mean? Retrieved from http://physicsandcake.wordpress.com/2011/01/22/pavlovs-ai-what-did-it-mean

Goertzel, B. (2003). Mindplexes: The potential emergence of multiple levels of focused consciousnessin communities of AI's and humans. *Dynamical Psychology*. Retrieved from http://www.goertzel.org/dynapsyc/2003/mindplex.html

Goertzel, B. (2005, March 6). Potential computational linguistics resources for Lojban. Retrieved from http://www.goertzel.org/new_research/lojban_AI.pdf

Goodhart, C. (1975). *Monetary relationships: A view from Threadneedle Street* Papers in Monetary Economics (Vol. I). Sydney: Reserve Bank of Australia.

Heath, R. G. (1963). Electrical self-stimulation of the brain in man. *American Journal of Psychiatry*, *120*, 571–577.

Hibbard, B. (2011). Model-based utility functions. Retrieved from http://arxiv.org/abs/1111.3934

Hutter, M. (2010). *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Berlin: Springer.

Kaczynski, T. (1995, September 19). Industrial society and its future. *The New York Times*.

Kanazawa, S., & Hellberg, J. (2010). Intelligence and substance use. *Review of General Psychology*, *14*, 382–396.

Lenat, D. (1983). EURISKO: A program that learns new heuristics and domain concepts. *Artificial Intelligence*, *21*, 61–98.

Levitt, S. D., & Dubner, S. J. (2006). *Freakonomics: A rogue economist explores the hidden side of everything*. New York, NY: William Morrow.

Mahoney, M. (2011). The wirehead problem – Candidate solutions? Retrieved from http://www.AGI@listbox.com mailinglist

Muehlhauser, L., & Helm, L. (2012). The singularity and machine ethics. In A. Eden, J. Søraker, J. Moor, & E. Steinhart (Eds.), *The singularity hypothesis: A scientific and philosophical assessment*. Berlin, Heidelberg: Springer.

Neches, R., Fikes, R., Finin, T., Gruber, Thomas, Patil, R., Senator, T., & Swartout, W. R. (1991). Enabling technology for knowledge sharing. *AI Magazine*, *12*, 37–56.

Nozick, R. (1977). *Anarchy, state, and Utopia*. New York, NY: Basic Books.

Olds, J., & Milner, P. (1954). Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative & Physiological Psychology*, *47*, 419–427.

Omohundro, S. M. (2008, February). The basic AI drives. In P. Wang, B. Goertzel, & S. Franklin (Eds.), *Proceedings of the First AGI Conference, (Vol. 171), frontiers in artificial intelligence and applications* (pp. 483–493). Amsterdam: IOS Press.

Orseau, L., & Ring, M. (2011). Self-modification and mortality in artificial agents. Paper presented at the 4th International Conference on Artificial General Intelligence, Mountain View, CA.

Patil, R., Mckay, D., Finin, T., Fikes, R., Gruber, T., Patel-Schneider, P. F., & Neches, R. (1992). An overview of the darpa knowledge sharing effort. Paper presented at the 3rd International Conference on Principles of Knowledge Representation and Reasoning.

Pearce, D. (2012, March 7). Wirehead hedonism versus paradise engineering. Retrieved from http://wireheading.com

Rice, H. (1953). Classes of recursively enumerable sets and their decision problems. *Transactions of American Mathematical Society*, *74*, 358–366.

Journal of Experimental & Theoretical Artificial Intelligence

Ring, M., & Orseau, L. (2011). *Delusion, survival, and intelligent agents*. Paper presented at the 4th International Conference on Artificial General Intelligence, Mountain View, CA.

Schrödinger, E. (1935, November). Die gegenwärtige Situation in der Quantenmechanik. *Naturwissenschaften*, *23*, 807–812.

Steunebrink, B., & Schmidhuber, J. (2011). *A family of gödel machine implementations*. Paper presented at the 4th Conference on Artificial General Intelligence (AGI-11), Mountain View, CA.

Stoklosa, T. (2010). Super intelligence. *Nature*, *467*, 878.

Turing, A. M. (1936). On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society*, *42*, 230–265.

Tyler, T. (2011a). Rewards VS goals. Retrieved from http://matchingpennies.com/rewards_vs_goals/

Tyler, T. (2011b). Utility counterfeiting. Retrieved from http://matchingpennies.com/utility_counterfeiting

Tyler, T. (2011c). The wirehead problem. Retrieved from http://alife.co.uk/essays/the_wirehead_problem/

Wagman, B., & Stephens, T. (2004). Surprising 'ultra-conserved' regions discovered in human genome. UC Santa Cruz Currents Online. Retrieved from http://currents.ucsc.edu/03-04/05-10/genome.html

Welch, C. (2011). Discussion of pavlov's AI – What did it mean? Retrieved from http://physicsandcake.wordpress.com/2011/01/22/pavlovs-ai-what-did-it-mean/

Wolfram, S. (2002, May 14). *A new kind of science*. Champaign, IL: Wolfram Media.

Yampolskiy, R. V. (2007, September 10–12). *Online poker security: Problems and solutions*. Paper presented at the EUROSIS North American Simulation and AI in Games Conference (GAMEON-NA2007), Gainesville, FL.

Yampolskiy, R. V. (2008). Behavioral modeling: An overview. *American Journal of Applied Sciences*, *5*, 496–503.

Yampolskiy, R. V. (2008, January 10–12). *Detecting and controlling cheating in online poker*. Paper presented at the 5th Annual IEEE Consumer Communications and Networking Conference (CCNC2008), Las Vegas, NE.

Yampolskiy, R. V. (2011). AI-complete CAPTCHAs as zero knowledge proofs of access to an artificially intelligent system. Article ID 271878. ISRN Artificial Intelligence.

Yampolskiy, R. V. (2011a, October 3–4). *Artificial intelligence safety engineering: Why machine ethics is a wrong approach*. Paper presented at the Philosophy and Theory of Artificial Intelligence (PT-AI2011), Thessaloniki, Greece.

Yampolskiy, R. V. (2011b, October 3–4). *What to do with the singularity paradox?* Paper presented at the Philosophy and Theory of Artificial Intelligence (PT-AI2011), Thessaloniki, Greece.

Yampolskiy, R. V. (2012). Leakproofing singularity – Artificial intelligence confinement problem. *Journal of Consciousness Studies*, *19*, 194–214.

Yampolskiy, R. V. (2013). Turing test as a defining feature of AI-completeness. In Xin-She Yang (Ed.), *Artificial intelligence, evolutionary computing and metaheuristics* (pp. 3–17). Berlin: Springer.

Yampolskiy, R. V., & Fox, J. (2012). Artificial intelligence and the human mental model. In A. Eden, J. Moor, J. Soraker, & E. Steinhart (Eds.), *The singularity hypothesis: A scientific and philosophical assessment*. Berlin: Springer.

Yampolskiy, R., & Gavrilova, M. (2012). Artimetrics: Biometrics for artificial entities. *IEEE Robotics and Automation Magazine*, *19*, 48–58.

Yampolskiy, R. V., & Govindaraju, V. (2007, November 20–22). *Behavioral biometrics for recognition and verification of game bots*. Paper presented at the the 8th Annual European Game-On Conference on Simulation and AI in Computer Games (GAMEON'2007), Bologna, Italy.

Yampolskiy, R. V., & Govindaraju, V. (2008). Behavioral biometrics: A survey and classification. *International Journal of Biometric*, *1*, 81–113.

Yampolskiy, R. V., & Govindaraju, V. (2008, March 16–20). *Behavioral biometrics for verification and recognition of malicious software agents*. Paper presented at the SPIE Defense and Security Symposium, Orlando, FL.

Yampolskiy, R. V., Klare, B., & Jain, A. K. (2012, December 12–15). *Face recognition in the virtual world: Recognizing avatar faces*. Paper presented at the the 11th International Conference on Machine Learning and Applications (ICMLA'12), Boca Raton, FL.

Yudkowsky, E. (2007). The hidden complexity of wishes. Retrieved from http://lesswrong.com/lw/ld/the_hidden_complexity_of_wishes/

Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. In N. Bostrom & M. M. Cirkovic (Eds.), *Global catastrophic risks* (pp. 308–345). Oxford: Oxford University Press.

Yudkowsky, E. (2011). Complex value systems in friendly AI. In J. Schmidhuber, K. Thórisson, & M. Looks (Eds.), *Artificial general intelligence* (Vol. 6830, pp. 388–393). Berlin, Heidelberg: Springer.

Yudkowsky, E. S. (2001). Creating friendly AI – The analysis and design of benevolent goal architectures. Retrieved from http://singinst.org/upload/CFAI.html

Zuse, K. (1969). *Rechnender Raum*. Braunschweig: Friedrich Vieweg & Sohn.

Journal of Experimental & Theoretical Artificial Intelligence